HANDBOOKS IN ECONOMICS

HANDBOOK OF ECONOMIC GROWTH

Editors:
Philippe Aghion
Steven N. Durlauf

NORTH-HOLLAND

VOLUME 2 2014

HANDBOOK OF
ECONOMIC GROWTH

Edited by

PHILIPPE AGHION
Harvard University

and

STEVEN N. DURLAUF
University of Wisconsin at Madison

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo
North Holland is an imprint of Elsevier

N·H

CONTRIBUTORS

Philippe Aghion
Harvard University, NBER, and CIFAR, USA.

Ufuk Akcigit
University of Pennsylvania and NBER, USA.

Alberto Alesina
Harvard University, USA.
IGIER Bocconi, Italy.

Yann Algan
Sciences Po, France.

Holger Breinlich
University of Essex, CEP and CEPR.

Pierre Cahuc
ENSAE-CREST, Ecole Polytechnique, France.

A.W. Carus
Faculty of Economics, University of Cambridge, United Kingdom.

Gregory Clark
University of California, Davis, CA 95616, USA.

Diego Comin
Harvard University, NBER and CEPR, USA.

Nicholas Crafts
Warwick University.

Matthias Doepke
Department of Economics, Northwestern University and NBER, 2001 Sheridan Road, Evanston, IL 60208, USA.

Gilles Duranton
Wharton School, University of Pennsylvania, 3620 Locust Walk, Philadelphia, PA 19104, USA.
CEPR.

Paola Giuliano
UCLA Anderson School of Management, NBER and CEPR.

Berthold Herrendorf
Department of Economics, Arizona State University, Tempe, AZ 85287, USA.

Peter Howitt
Brown University and NBER, USA.

Christopher M. Meissner
Department of Economics, University of California, Davis and NBER, Davis, CA 95616, USA.

Martí Mestieri
Toulouse School of Economics, France.

Nathan Nunn
Department of Economics, Harvard University, NBER, and BREAD, 1805 Cambridge Street, Room M25, Cambridge, MA 02138, USA.

Sheilagh Ogilvie
Faculty of Economics, University of Cambridge, United Kingdom.

Kevin Hjortshøj O'Rourke
All Souls College, Oxford, United Kingdom.

Gianmarco I.P. Ottaviano
London School of Economics, CEP and CEPR, United Kingdom

Diego Puga
Centro de Estudios Monetarios y Financieros (CEMFI), Casado del Alisal 5, 28014 Madrid, Spain.

Richard Rogerson
Princeton University & NBER, Princeton, United States.

Enrico Spolaore
Department of Economics, Tufts University, NBER, CESIfo and CAGE, Medford, MA 02155-6722, USA.

Jonathan R.W. Temple
University of Bristol and CEPR.

Ákos Valentinyi
Cardiff Business School, IE-CERSHAS & CEPR, United Kingdom.

Romain Wacziarg
UCLA Anderson School of Management, NBER and CEPR, 110 Westwood Plaza, Los Angeles, CA 90095, USA.

David N. Weil
Brown University and NBER, USA.

Yang Yao
China Center for Economic Research, National School of Development, Peking University, China.

Fabrizio Zilibotti
Department of Economics, University of Zurich, Muehlebachstrasse 86, Zurich, CH 8008, Switzerland.

CHAPTER ONE

Culture, Entrepreneurship, and Growth

Matthias Doepke[*] and **Fabrizio Zilibotti**[†]

[*]Department of Economics, Northwestern University and NBER, 2001 Sheridan Road, Evanston, IL 60208, USA
[†]Department of Economics, University of Zurich, Muehlebachstrasse 86, Zurich, CH 8008, Switzerland

Abstract

We discuss the two-way link between culture and economic growth. We present a model of endogenous technical change where growth is driven by the innovative activity of entrepreneurs. Entrepreneurship is risky and requires investments that affect the steepness of the lifetime consumption profile. As a consequence, the occupational choice of entrepreneurship hinges on risk tolerance and patience. Parents expecting their children to become entrepreneurs have an incentive to instill these two values in their children. Cultural transmission is Beckerian, i.e. parents are driven by the desire to maximize their children's happiness. We also consider, in an extension, a paternalistic motive for preference transmission. The growth rate of the economy depends on the fraction of the population choosing an entrepreneurial career. How many entrepreneurs there are in a society hinges, in turn, on parental investments in children's patience and risk tolerance. There can be multiple balanced growth paths, where in faster-growing countries more people exhibit an "entrepreneurial spirit." We discuss applications of models of endogenous preferences to the analysis of socio-economic transformations, such as the British Industrial Revolution. We also discuss empirical studies documenting the importance of culture and preference heterogeneity for economic growth.

Keywords

Culture, Entrepreneurship, Innovation, Economic growth, Endogenous preferences, Intergenerational preference transmission

JEL Classification Codes

J24, L26, N30, O10, O32, O33, O43, Z10

1.1. INTRODUCTION

The relationship between economic development and culture—broadly defined as the set of preferences, values, and beliefs that are at least partially learned—has attracted increasing attention in the economic literature over the last decade.

The notion that accounting for cultural heterogeneity is important for explaining individual behavior and economic success was a familiar one to classical economists. For instance, Smith (1776) described members of different social classes of his time as distinct types of human beings driven by different motives: *"A merchant is accustomed to employ his money chiefly in profitable projects; whereas a mere country gentleman is accustomed to employ it*

Handbook of Economic Growth, Volume 2A
ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53538-2.00001-0

chiefly in expense. The one often sees his money go from him and return to him again with a profit: the other, when once he parts with it, very seldom expects to see any more of it" (p. 432).

A century later, Karl Marx postulated that culture is the effect, rather than the cause, of the structure of production relations. In his view, culture, religion, and ideology (the "superstructure") are mere reflections of the material interests of the class that controls the means of production. Marx' materialism was disputed by Max Weber, who argued, that cultural and spiritual factors are independent drivers of socio-economic transformations. For Weber, the emergence of a "spirit of capitalism" with the ensuing emphasis on the virtue of entrepreneurial success was a major engine of the industrial revolution, not just a mere reflection of it. Weber did not fully reverse Marx' perspective, but rather acknowledged that the causation can run both ways.[1] For instance, he held the view that Protestant Asceticism had been an engine of economic transformation, but "was in turn influenced in its development and its character by the totality of social conditions, especially economic" (Weber, 1905, p. 183).

In contrast to the thinking of Smith, Marx, and Weber, the marginalist revolution in economics in the late 19th century sidelined cultural factors. According to the neoclassical paradigm, economics should focus on optimal individual choice and efficient resource allocation, while treating preferences and technology as exogenous primitives. Consistent with this paradigm, until recently economists have regarded preference formation, and culture more broadly, as issues lying outside the realm of economics. Over time, however, as economic imperialism has broken into new territories, exogenous preferences and technology have become straitjackets. The erosion of the neoclassical tenets began from technology. It is by now widely recognized, following the intuition of Schumpeter (1942), that technology cannot be viewed as exogenous if one wants to understand the mechanics of the growth process of industrial as well as developing economies. Rather, the efforts and risk-taking behavior of a particular group of individuals that aims to change the set of technological constraints, namely inventors and entrepreneurs, are the engines of economic growth. This observation motivated the development of the neo-Schumpeterian endogenous technical change paradigm throughout the 1990s (see, e.g. Aghion and Howitt, 1992).

Recently, the paradigm shift has extended to the realm of preferences. The availability of large data sets such as the World Value Survey has revealed that there is a great deal of heterogeneity in values and preferences across both individuals (see, e.g. Guiso and Paiella, 2008; Beauchamp et al. 2011), and world regions (see, e.g. Inglehart et al. 2000). Preference heterogeneity has also become a salient issue in mainstream macroeconomics. For instance, Krusell and Smith (1998), Coen-Pirani (2004), De Nardi (2004), Guvenen (2006), Hendricks (2007), and Cozzi (2011) have argued that individual variation in

[1] "It is, of course, not my aim to substitute for a one-sided materialistic an equally one-sided spiritualistic causal interpretation of culture and of history" (Weber, 1905, p. 183).

preferences is necessary for calibrated macroeconomic models with incomplete markets to reproduce the large wealth inequality observed in the data.

Preference heterogeneity as such is not in conflict with the neoclassical paradigm. Traditionally, extra-economic factors have served as the motivations for error terms in regressions and individual or regional fixed effects. However, treating preferences and culture as exogenous factors in growth and development theory is problematic if, on the one hand, cultural factors respond to changes in the economic and institutional environment (see Alesina and Glaeser, 2004; Alesina and Giuliano, 2009), and, on the other hand, culture and preferences have an important feedback on institutions and economic performance (see Greif, 1994; Grosjean, 2013; Guiso et al. 2006; Gorodnichenko and Gerard, 2010; Tabellini, 2010).

Motivated by these observations, a growing number of studies incorporate endogenous cultural change into economic models.[2] A particularly important link is the one connecting preferences, culture, and innovation (see Mokyr, 2011). In many recent models of endogenous technical change, innovation and economic growth ultimately are determined by policy and preference parameters, such as the time discount rate and risk aversion. Yet, there is a lack of studies of the joint determination of preferences and technology. A key issue is the extent to which different societies differ in terms of the average propensity of their citizens to carry out entrepreneurial or innovative activities. This is the focus of the investigation of this chapter.

To this aim, we present a model of endogenous technical change where growth is driven by the innovative activity of entrepreneurs. The focal point of the analysis is the occupational choice between being a worker and being an entrepreneur in an economy with capital market imperfections. Entrepreneurs face more risk and make investments that force them to defer consumption. As a consequence, the occupational choice hinges on patience and risk tolerance. These preference traits are distributed heterogeneously in the population and subject to the influence of family upbringing. Cultural transmission is driven by the desire of parents to maximize their children's happiness, conditional on the expectations they hold about the children's future occupation. Parents expecting their children to become entrepreneurs have stronger incentives to raise them to be patient and risk tolerant.

At the aggregate level, the growth rate of the economy depends on the fraction of entrepreneurs in the population, since this determines the rate of technological innovation. The theory identifies a self-reinforcing mechanism linking preferences and growth. In a highly entrepreneurial society, a large proportion of the population is patient and risk tolerant. These preferences sustain high human capital investment and risky innovation, leading to a high growth rate and incentives for entrepreneurial preferences to develop

[2] The recent literature in behavioral economics has proposed a psychological foundation for endogenous preferences. Fehr and Hoff (2011) argue that individual preferences are susceptible to institutional, familiar, and social influences due to their intrinsic psychological properties.

in the next generation, too. Societies with identical primitives may end up in different balanced growth paths characterized by different degrees of entrepreneurial culture, innovativeness, and growth. In addition, changes in institutions or policies can feed back into the evolution of culture and preferences, giving rise to potentially long-lasting effects on economic growth and development.

This chapter is organized as follows. Section 1.2 presents a model of endogenous technical change with an occupational choice, where entrepreneurship is the driver of innovation. Sections 1.3 and 1.4 endogenize culture and preference transmission analyzing, respectively, the endogenous accumulation of patience and risk tolerance. While in Sections 1.3 and 1.4 the cultural transmission of preferences hinges on an altruistic Beckerian motive, Section 1.5 considers an alternative model incorporating parental paternalism. Section 1.6 reviews the existing theoretical and empirical literature. Section 1.7 concludes. Proofs of propositions and lemmas are deferred to the mathematical appendix.

1.2. A FRAMEWORK FOR ANALYZING THE INTERACTION OF CULTURAL PREFERENCES, ENTREPRENEURSHIP, AND GROWTH

In this section, we develop a dynamic model where culture and economic growth are jointly determined in equilibrium. The underlying process of technical change is related to the model of Romer (1990), where growth takes the form of an expanding variety of inputs. However, unlike Romer we assume that innovation is driven by a specific group of people, namely entrepreneurs, whose economic lives (for example, in terms of risk and lifetime consumption profiles) are distinct from those of ordinary workers. Cultural preferences determine people's propensity to entrepreneurship, and conversely the return to entrepreneurship affects parents' incentives for forming their children's preferences. In other words, there is a two-way interaction between culture and growth. In this section, we develop the general setup, turning to specific dimensions of endogenous preferences further below.

1.2.1 A Model of Endogenous Innovation

Consider an endogenous growth model where innovation takes the form of an increasing variety of intermediate inputs. New inputs are created by people in a specific occupation, namely entrepreneurs (as in Klasing, 2012). Innovative activity has two key features: it involves investments and deferred rewards (as in Doepke and Zilibotti, 2008), and it may also involve risk (as in Doepke and Zilibotti, 2012 and Klasing, 2012). In addition, financial markets are incomplete: agents can neither borrow to smooth consumption over the life cycle, nor hedge the entrepreneurial risk.[3] Since entrepreneurs and regular workers face

[3] While these assumptions are stark, models with moral hazard typically imply imperfect consumption smoothing or risk sharing. Empirically, we observe that entrepreneurs can neither borrow without

different consumption profiles (across both time and states of nature), the choice between these two occupations hinges on heterogeneous cultural preferences.

The measure of the intermediate input varieties invented before the start of period t is denoted by N_t. Time is discrete. Final output at time t is produced using the production function:

$$Y_t = \frac{1}{\alpha}\left(\int_0^{N_t} \bar{x}_t\left(i\right)^{\alpha} di + \int_{N_t}^{N_{t+1}} x_t\left(i\right)^{\alpha} di\right) Q^{1-\alpha},$$

where Q is a fixed factor (e.g. land or unskilled labor) that will be normalized to unity; $\bar{x}_t\left(i\right)$ is the supply of intermediates i that were invented up until time t; and $x_t\left(i\right)$ is the supply of new varieties i invented during period t. Following Matsuyama (1999), we assume that old varieties with $i \in [0, N_t]$ are sold in competitive markets, whereas new varieties $i \in (N_t, N_{t+1}]$ are supplied monopolistically by their inventors. Put differently, inventors enjoy patent protection for only one period.

Innovation (i.e. the introduction of $N_{t+1} - N_t$ new varieties) is carried out by entrepreneurs. The return to entrepreneurial effort is assumed to be stochastic. In particular, entrepreneurs do not know in advance how successful they will be at inventing new varieties. With probability $\kappa > 0$ an entrepreneur will be able to run $\left(1 + \nu\right) N_t$ projects, whereas with probability $1 - \kappa$ he or she will manage only $\left(1 - \nu\frac{\kappa}{1-\kappa}\right) N_t$ projects, where $\nu \geq 0$. In the aggregate, κ is the fraction of successful entrepreneurs. Intermediate-good production is instead carried out by workers using a linear technology that is not subject to uncertainty.

In order for the equilibrium to feature balanced growth, we assume that a knowledge spillover increases the productivity of both workers and entrepreneurs as knowledge accumulates. More precisely, productivity is indexed by N_t, and thus grows at the equilibrium rate of innovation. Given these assumptions, the labor market-clearing condition at time t is given by:

$$N_t X_t^W = N_t \bar{x}_t + \left(N_{t+1} - N_t\right) x_t,$$

where the left-hand side is the labor supply by workers in efficiency units, and the right-hand side is the labor demand given the production of intermediates $\bar{x}_t$ and x_t.[4] The corresponding market-clearing condition for entrepreneurs is:

$$N_t X_t^E = \left(\frac{N_{t+1} - N_t}{\xi}\right),$$

where X_t^E is the number of entrepreneurs, and the parameter ξ captures the average productivity per efficiency unit of entrepreneurial input in innovation. Hence, an efficiency unit of the entrepreneurial input produces measure ξ of new varieties. Denoting

constraints to finance their investments, nor separate their personal economic success from the fate of their enterprises. Thus, our stylized model captures some important features of the real world that are well-understood outcomes of models of imperfect information.

[4] Note that the market-clearing expression is written under the assumption that all old varieties $i \in [0, N_t]$ are supplied at the same level, $\bar{x}_t$, and that all new varieties $i \in (N_t, N_{t+1}]$ are supplied at the same level, x_t. We show later that this the case in equilibrium.

the growth rate of technology by $g_t \equiv (N_{t+1} - N_t)/N_t$ allows us to simplify the two market-clearing conditions as follows:

$$X_t^W = \bar{x}_t + g_t x_t, \tag{1.1}$$

$$X_t^E = \frac{g_t}{\xi}. \tag{1.2}$$

We now turn to the goods-market equilibrium. The representative competitive final-good producer maximizes profits by solving:

$$\max_{\bar{x}(i),x(i)} \left\{ \frac{1}{\alpha} \left(\int_0^{N_t} [\bar{x}_t(i)^\alpha - \alpha \bar{p}_t(i) \bar{x}_t(i)] di + \int_{N_t}^{N_{t+1}} [x_t(i)^\alpha - \alpha p_t(i) x_t(i)] di \right) \right\},$$

where $\bar{p}_t(i)$ and $p_t(i)$ are the prices of old and new intermediates, respectively.[5] The first-order conditions for the maximization problem imply:

$$\bar{x}_t(i) = \bar{p}_t(i)^{\frac{1}{\alpha-1}} \quad \text{and} \quad x_t(i) = p_t(i)^{\frac{1}{\alpha-1}}. \tag{1.3}$$

Next, we consider the intermediate-goods producers. Let w_t^W denote the market wage of workers, and let $\omega_t^W = w_t^W / N_t$ denote the wage per efficiency unit of labor. The maximization problem for the competitive producers of old intermediates with $i \in [0, N_t]$ can then be written as:

$$\max_{\bar{x}_t(i)} \left\{ \left(\bar{p}_t(i) - \omega_t^W \right) \bar{x}_t(i) \right\},$$

so that we have $\bar{p}_t(i) = \omega_t^W$ and, hence:

$$\bar{x}_t(i) = \left(\omega_t^W \right)^{\frac{1}{\alpha-1}}. \tag{1.4}$$

The producers of new goods (i.e. the firms run by entrepreneurs) are monopolists that maximize profits subject to the demand function (1.3). More formally, they solve:

$$\max_{x_t(i),p_t(i)} \left\{ \left(p_t(i) - \omega_t^W \right) x_t(i) \right\}$$

subject to (1.3). The solution to this problem yields:

$$p_t(i) = \frac{\omega_t^W}{\alpha} \equiv p_t, \tag{1.5}$$

$$x_t(i) = \left(\frac{\omega_t^W}{\alpha} \right)^{\frac{1}{\alpha-1}} \equiv x_t, \tag{1.6}$$

[5] The fixed factor $Q = 1$ is owned by firms, so that profits correspond to the return to the fixed factor. For simplicity, we assume that firms are held by "capitalist" dynasties that are distinct from the workers and entrepreneurs, although allowing for trade in firm shares would not change our results.

and the realized profit per variety is:

$$\Pi_t = \left(p_t - \omega_t^W\right) x_t = (1-\alpha)\left(\frac{\alpha}{\omega_t^W}\right)^{\frac{\alpha}{1-\alpha}}.$$

We can now solve for the equilibrium return to labor and entrepreneurship as functions of the aggregate supply of regular and entrepreneurial labor. First, combining (1.1), (1.4), and (1.6) yields:

$$X_t^W = \left(\omega_t^W\right)^{\frac{1}{\alpha-1}} + g_t\left(\frac{\omega_t^W}{\alpha}\right)^{\frac{1}{\alpha-1}}.$$

Using (1.2) to eliminate g_t, and rearranging terms, yields the following expression for the workers' normalized wage:

$$\omega_t^W = \left(\frac{1+g_t\alpha^{\frac{1}{1-\alpha}}}{X_t^W}\right)^{1-\alpha} = \left(\frac{1+\alpha^{\frac{1}{1-\alpha}}\xi X_t^E}{X_t^W}\right)^{1-\alpha}.$$

Next, denote by w_t^E the expected profit of entrepreneurs, and let $\omega_t^E = w_t^E/N_t$.[6] Then, the following expression for the return to entrepreneurship obtains:

$$\omega_t^E = \xi\Pi_t = \xi^{1-\alpha}(1-\alpha)\left(\frac{\alpha^{\frac{1}{1-\alpha}}\xi X_t^W}{1+\alpha^{\frac{1}{1-\alpha}}\xi X_t^E}\right)^{\alpha}.$$

Finally, let $\eta_t \equiv w_t^E/w_t^W$ denote the expected entrepreneurial premium. Taking the ratio between the expressions of the two returns obtained above yields:

$$\eta_t = \frac{(1-\alpha)\,\alpha^{\frac{\alpha}{1-\alpha}}\xi X_t^W}{1+\alpha^{\frac{1}{1-\alpha}}\xi X_t^E}. \tag{1.7}$$

Innovation and growth are ultimately pinned down by the share of the population choosing entrepreneurship. The occupational choice, in turn, hinges on both technological variables and the endogenous distribution of individual preferences. We therefore turn, next, to the structure of preferences in the economy.

1.2.2 Demographics and Structure of Preferences

The model economy is populated by overlapping generations of altruistic people who live for two periods. Every person has one child, and a measure one of people is born each period. The lifetime utility V_t of a person born at time t is given by:

$$V_t = \chi U(c_{1,t}) + \beta U(c_{2,t}) + zV_{t+1}, \tag{1.8}$$

[6] Recall that the entrepreneurial return is stochastic. Each entrepreneur earns $(1+\nu)\,w_t^E$ with probability κ and $\left(1-\nu\frac{\kappa}{1-\kappa}\right)w_t^E$ with probability $1-\kappa$.

where $c_{1,t}$ is consumption when young, $c_{2,t}$ is consumption when old, and V_{t+1} is the lifetime utility of the person's child. Preferences are pinned down by the shape of the period utility function $U(\cdot)$ and by the weights χ, β, and z attached to young-age consumption, old-age consumption, and the utility of the child, respectively. Below, we endogenize the determination (via intergenerational transmission) of specific preference parameters. More specifically, we assume that people can shape certain aspects of their children's preferences, but cannot change their own preferences. Economic decisions within a generation are taken therefore for fixed preference parameters. This feature allows us to discuss economic choices and preference transmission separately.

People have one unit of time in each period. When young, they make a career choice between being workers or being entrepreneurs. Workers supply one unit of labor to the labor market in each period. Entrepreneurs supply a fraction ψ of their time to the labor market when young, and use the remainder $1 - \psi$ for human capital investment.[7] When old, entrepreneurs use all their time for innovating, with a return to innovation as described in Section 1.2.1.

As generations overlap, at time t labor is supplied by the people born in periods $t-1$ and t. Let λ_t denote the fraction of entrepreneurs in the generation born at time t. Then, aggregate labor supply at time t is given by:

$$X_t^W = 1 - \lambda_t + \lambda_t \psi + 1 - \lambda_{t-1}, \tag{1.9}$$

namely, it is the sum of labor supply by young workers, young entrepreneurs, and old workers. The supply of entrepreneurial input is given by the labor supply of old entrepreneurs:

$$X_t^E = \lambda_{t-1}. \tag{1.10}$$

Equations (1.2) and (1.10) imply that the growth rate of the economy is given by $g_t = \lambda_{t-1}\xi$.

1.2.3 Balanced Growth Path for Fixed Preferences

To establish a benchmark, we first analyze balanced growth paths for the case of fixed preferences. That is, parents do not affect their children's preferences, and the preference parameters χ, β, and z, as well as the $U(\cdot)$ function are fixed. For simplicity, we focus initially on the case where entrepreneurship is not risky, $\nu = 0$. In a balanced growth path, the growth rates of output and consumption are constant, as is the fraction of the population comprised of entrepreneurs. This balanced growth path requires that preferences feature a constant intertemporal elasticity of substitution, so that period utility

[7] Other ways of modeling the cost of becoming an entrepreneur would yield similar results as long as the cost results in lower utility at young age, and therefore has the characteristic of an investment.

is given by:

$$U(c) = \frac{c^{1-\sigma}}{1-\sigma}.$$

We restrict attention to the case $0 \leq \sigma < 1$, because the analysis of the economy with endogenous preferences will require utility to be positive (although this can be generalized, see Doepke and Zilibotti, 2008). We also impose the following restriction:

$$(1+\xi)^{1-\sigma} z < 1,$$

which guarantees that discounted utility is well defined.

Given that with fixed preferences everyone's preferences are the same, the key condition for a balanced growth path with a positive growth rate is that the entrepreneurial premium, η, makes people just indifferent between being workers and being entrepreneurs.[8] The indifference condition for people born at time t can be written as:

$$\chi u\left(w_t^W\right) + \beta u\left(w_{t+1}^W\right) + zV_{t+1} = \chi u\left(\psi w_t^W\right) + \beta u\left(w_{t+1}^E\right) + zV_{t+1},$$

where the left-hand side is the utility of workers and the right-hand side is the utility of entrepreneurs. Note that the utility derived from children is identical for both occupations, and therefore does not feature in the indifference condition. In a balanced growth path, wages and entrepreneurial returns are given by $w_t^W = N_t \omega^W$ and $w_t^E = N_t \omega^E$, respectively, where ω^W and ω^E are constants and N_t grows at the constant rate g. Canceling common terms allows us to rewrite the indifference condition in this form involving only variables that are constant in the balanced growth path:

$$\chi \frac{\left(\omega^W\right)^{1-\sigma}}{1-\sigma} + \beta \frac{((1+g)\omega^W)^{1-\sigma}}{1-\sigma} = \chi \frac{(\psi \omega^W)^{1-\sigma}}{1-\sigma} + \beta \frac{((1+g)\omega^E)^{1-\sigma}}{1-\sigma}. \tag{1.11}$$

Condition (1.11) can be further simplified by dividing both sides of the equality by $(\omega^W)^{1-\sigma}$, and rewriting it in terms of the entrepreneurial premium $\eta = \omega^E/\omega^W$:

$$\chi + \beta(1+g)^{1-\sigma} = \chi(\psi)^{1-\sigma} + \beta((1+g)\eta)^{1-\sigma}. \tag{1.12}$$

Next, consider the expression for the entrepreneurial premium, (1.7). Plugging in the balanced growth levels of X^W and X^E from (1.9) and (1.10), we can express the premium as a function of the fraction of entrepreneurs, λ:

$$\eta = (1-\alpha)\,\alpha^{\frac{\alpha}{1-\alpha}} \xi \frac{2-(2-\psi)\lambda}{1+\alpha^{\frac{1}{1-\alpha}} \xi \lambda}. \tag{1.13}$$

[8] The analysis here applies to interior balanced growth paths where positive proportions of agents choose either occupation, worker, or entrepreneur. More discussion is provided below.

Combining (1.12) and (1.13), recalling that $g = \lambda\xi$, and rearranging terms yields:

$$\chi\left(1-(\psi)^{1-\sigma}\right) = \beta(1+\lambda\xi)^{1-\sigma}\left(\left((1-\alpha)\,\alpha^{\frac{\alpha}{1-\alpha}}\xi\,\frac{2-(2-\psi)\lambda}{1+\alpha^{\frac{1}{1-\alpha}}\xi\lambda}\right)^{1-\sigma}-1\right). \quad (1.14)$$

Here the left-hand side is the (normalized) cost of becoming an entrepreneur in terms of forgone utility when young, and the right-hand side is the (normalized) benefit in terms of higher utility when old. Equation (1.14) pins down the equilibrium fraction of entrepreneurs, λ, which in turn determines the entrepreneurial premium and the rate of economic growth.

Depending on parameters, there can be corner solutions with $\lambda = 0$ or $\lambda = 1$, i.e. there aren't any entrepreneurs or all old agents are entrepreneurs. In addition, the balanced growth path need not be unique. The reason is that on the one hand an increase in the fraction of entrepreneurs lowers the entrepreneurial premium (making entrepreneurship less attractive), but on the other hand it also increases the growth rate (making entrepreneurship, where higher rewards occur later in life, relatively more attractive). To provide a sharp contrast with the case of endogenous preferences, we will focus on parameter configurations where the balanced growth path for fixed preferences is both interior and unique.

Assumption 1. The parameters α, ξ, and ψ satisfy:

$$2\,(1-\alpha)\,\alpha^{\frac{\alpha}{1-\alpha}}\xi > 1 > \frac{(1-\alpha)\,\alpha^{\frac{\alpha}{1-\alpha}}\xi\,\psi}{1+\alpha^{\frac{1}{1-\alpha}}\xi}.$$

Proposition 1. *Under Assumption 1, there exists a* $\bar{\chi}(\alpha,\xi,\psi) > 0$ *such that for all* $\chi < \bar{\chi}(\alpha,\xi,\psi)$ *a unique interior balanced growth equilibrium exists, i.e. there is a unique* $\lambda \in (0,1)$ *that satisfies Equation* (1.14).

1.3. ENDOGENOUS CULTURE I: WEBER AND THE TRANSMISSION OF PATIENCE

The balanced growth analysis in the previous section shows that the growth rate in our economy is determined by both technology parameters (such as the efficiency of the innovation technology ξ) and preference parameters (such as the time discount factor β). Despite this fact, when using similar growth models to address variations in economic growth across time and space, the literature has typically focused on variations in technology as the driving force. Unlike technology, preferences usually are assumed to be exogenous. Deviating from this practice, we now endogenize preferences, and analyze the interaction of preference formation with technology, occupational choice, and ultimately, economic growth.

1.3.1 Endogenizing Patience

We start by focusing on patience, parameterized by the time discount factor β. Since risk is not important for the analysis in this section, we abstract from uncertainty and assume that $\nu = 0$. Adult agents in period t are endowed with a predetermined discount factor, β_t, but they can affect the discount factor of their children, β_{t+1}. For example, in their children's upbringing parents can emphasize the appreciation of future rewards. Given that we assume $\sigma < 1$, a higher β always yields higher utility. However, investing in children's patience is costly, so parents face a tradeoff. More precisely, denoting by l_t the effort a parent of generation t spends on raising her child's patience, the parent's discounted utility is:

$$\chi(l_t)\frac{c_{t,1}^{1-\sigma}}{1-\sigma} + \beta_t \frac{c_{t,2}^{1-\sigma}}{1-\sigma} + zV_{t+1}(\beta_{t+1}(l_t)),$$

where χ is a strictly decreasing, strictly concave, and differentiable function, and effort is bounded by $0 \leq l_t \leq 1$. The structure of preferences is still of the form given in (1.8), although χ and β are now endogenous variables rather than given parameters. The child's patience is given by:

$$\beta_{t+1}(l_t) = (1-\delta)\beta_t + f(l_t), \tag{1.15}$$

where f is an increasing, non-negative, and strictly concave function, and δ satisfies $0 < \delta \leq 1$. Notice that if $\delta < 1$ there is some direct persistence in preferences across generations, which captures children's imitation of their parents and other transmission channels that do not require direct parental effort. In addition to this direct transmission, the function $f(l_t)$ captures the return to parental effort in terms of increasing the child's patience.

1.3.2 Transmission of Patience in the Balanced Growth Path

We now characterize balanced growth paths with endogenous patience. People face a twofold decision problem. First, when young they choose whether to be workers or become entrepreneurs. This decision hinges only on returns within the person's lifetime, and much of the previous analysis for fixed preferences still applies. Second, people choose the investment l_t in instilling patience in their children.

We proceed by analyzing the individual decision problem under the assumption that a balanced growth path has already been reached, so that the entrepreneurial premium is constant, and wages and profits grow at the constant rate g. The decision problem can be analyzed recursively, with the discount factor β serving as the state variable of a dynasty. In principle, the state of technology N_t is a second state variable, because growth in N_t scales up all wages and returns. However, due to the homothetic utility function, in a balanced growth path utility at time t can be expressed as:

$$V_t(\beta_t, N_t) = \left(\frac{N_t w_0^W}{N_0}\right)^{1-\sigma} v(\beta_t),$$

where v is a value function that does not depend on N_t and is scaled so that it gives utility conditional on the worker's wage being equal to one. This value function, in turn, satisfies the following set of Bellman equations:

$$v(\beta) = \max\left\{v^W(\beta), v^E(\beta)\right\}, \tag{1.16}$$

where:

$$v^W(\beta) = \max_{0\le l\le 1}\left\{\chi(l) + \beta\,(1+g)^{1-\sigma} + z\,(1+g)^{1-\sigma}\,v(\beta')\right\}, \tag{1.17}$$

$$v^E(\beta) = \max_{0\le l\le 1}\left\{\chi(l)\psi^{1-\sigma} + \beta\,((1+g)\,\eta)^{1-\sigma} + z\,(1+g)^{1-\sigma}\,v(\beta')\right\}. \tag{1.18}$$

The maximization in (1.17) and (1.18) is subject to the law of motion for patience across generations:

$$\beta' = (1-\delta)\beta + f(l). \tag{1.19}$$

The Bellman equations (1.17) and (1.18) represent the utilities conditional on choosing to be a worker or an entrepreneur, respectively, and (1.16) captures the optimal choice between these two careers.

Given our assumptions on f and l, there is a maximum level of patience, $\beta_{\max}$, that can be attained. The decision problem is therefore a dynamic programming problem with a single state variable in the interval $[0, \beta_{\max}]$, and can be analyzed using standard techniques. The following proposition summarizes the properties of the value function and the associated policy functions for investing in patience and for choosing an occupation.

Proposition 2. *The system of Bellman equations (1.16)–(1.18) has a unique solution. The value function v is increasing and convex in β. The optimal occupational choice is either to be a worker for any β, or there exists a $\bar{\beta}$ such that impatient people with $\beta < \bar{\beta}$ strictly prefer to be workers, patient people with $\beta > \bar{\beta}$ strictly prefer to be entrepreneurs, and people with $\beta = \bar{\beta}$ are indifferent. The optimal investment in patience $l = l(\beta)$ is non-decreasing in β.*

The proof of the proposition is contained in the mathematical appendix. The convexity of the value function follows from two features of the decision problem: the discount factor enters utility linearly, and there is a complementarity between being patient and being an entrepreneur.

To gain intuition, consider the decision problem without the occupational choice, i.e. assume that all members of a dynasty are forced to be either workers or entrepreneurs regardless of their patience. If we vary the discount factor β of the initial generation, while holding constant the investment choices l of all generations, the utility of the initial generation is a linear increasing function of β. This is because initial utility is a linear function of present and future discount factors, and the initial discount factor, in turn, has a linear effect on future discount factors through the term $1-\delta$ in the law of motion (1.19). In addition, if the occupation of all generations is held constant, it is in fact optimal to choose a constant l for all β, because the marginal return to investing in patience depends only on the choice of occupation, and not on β.

Now consider the full model with a choice between the two occupations. The career with the steeper income profile, namely entrepreneurship, is more attractive when β is high. As we increase β, each time either a current or future member of the dynasty switches from being a worker to being an entrepreneur, the value function also becomes steeper in β. The optimal l increases at each step, because the cost of providing patience declines with the steepness of the income profile, while the marginal benefit increases. Since there are only two possible occupations, the value function is piecewise linear, where the linear segments correspond to ranges of β for which the optimally chosen present and future occupations are constant. At each kink of the value function, some member of the dynasty is indifferent between being a worker and an entrepreneur. Since the choice of l depends on the chosen occupation, there may be multiple optimal choices l at a β where the value function has a kink, whereas in between kinks the optimal choice of l is unique. The following proposition summarizes our results regarding the optimal choice of income profiles and investment in patience.

Proposition 3. *The state space $[0, \beta_{\max}]$ can be subdivided into (at most) countably many closed intervals $[\underline{\beta}, \overline{\beta}]$ such that over the interior of any range $[\underline{\beta}, \overline{\beta}]$, the occupational choice of each member of the dynasty (i.e. parent, child, grandchild, and so on) is constant and unique (though possibly different across generations), and $l(\beta)$ is constant and single-valued. The value function $v(\beta)$ is piecewise linear, where each interval $[\underline{\beta}, \overline{\beta}]$ corresponds to a linear segment. Each kink in the value function corresponds to a switch, from being a worker to being an entrepreneur, by a present or future member of the dynasty. At a kink, the optimal choices of occupation and l corresponding to both adjoining intervals are optimal (thus, the optimal policy functions are not single-valued at a kink).*

The proposition implies that the optimal policy correspondence $l(\beta)$ is a non-decreasing step-function, which takes multiple values only at a step. Proposition 3 allows us to characterize the equilibrium law of motion for patience. Since the policy correspondence $l(\beta)$ is monotone, the dynamics of β are monotone as well and converge to a steady state from any initial condition.

Proposition 4. *The law of motion of β is described by the following difference equation:*

$$\beta' = g(\beta) = (1 - \delta)\,\beta + f\,(l\,(\beta))\,,$$

where $l(\beta)$ is a non-decreasing step-function (as described in Proposition 3). Given an initial condition β_0, patience in the dynasty converges to a constant β where parents and children choose the same profession.

Notice that while the discount factor of a dynasty always converges, the steady state does not have to be unique even for a given β_0. For example, if the initial generation is indifferent between the two occupations, the steady state can depend on which one is chosen.

Given the optimal occupational choices of parents and children, the optimal choice of l has to satisfy first-order conditions. This allows us to characterize more sharply the decisions on patience and their interaction with occupational choices. We have already established that both patience β and occupation converge within a dynasty. Thus, the population ultimately divides into worker dynasties and entrepreneur dynasties, and these two types face different incentives for investing in patience. Consider the case in which the solutions for l are interior. For workers, the first-order condition characterizing the optimal effort l^W for investing in patience is given by:

$$-\chi'(l^W) = \frac{z(1+g)^{2(1-\sigma)} f'(l^W)}{1 - z(1+g)^{1-\sigma}(1-\delta)}. \tag{1.20}$$

The corresponding condition for entrepreneurial dynasties is given by:

$$-\chi'(l^E)\psi^{1-\sigma} = \frac{z(1+g)^{2(1-\sigma)} \eta^{1-\sigma} f'(l^E)}{1 - z(1+g)^{1-\sigma}(1-\delta)}. \tag{1.21}$$

In both equations, the left-hand side is strictly increasing in l, and the right-hand side is strictly decreasing. Moreover, for a given l the left-hand side is smaller for entrepreneurial dynasties, and the right-hand side is larger. Therefore, in the balanced growth path we must have $l^E > l^W$: The returns to being patient are higher for entrepreneurs because of their steeper income profile, inducing them to invest more in patience. In the balanced growth path, we therefore also have $\beta^E > \beta^W$, where:

$$\beta^W = \frac{f(l^W)}{\delta},$$
$$\beta^E = \frac{f(l^E)}{\delta}.$$

These findings line up with Max Weber's (1905) view of entrepreneurs as future-oriented individuals who possess a "spirit of capitalism". However, in our theory, differences in patience are not just a determinant of occupational choice (as in Weber), but also a consequence of it. Entrepreneurial dynasties develop patience because of the complementarity between this preference trait and their occupation. In contrast, Weber focused on religion as a key determinant of values and preferences across social groups.

Figure 1.1 provides an example of the characteristics of the value and policy functions analyzed in Propositions 2 and 3.[9] In the example, the value function has two linear segments. Below the threshold of $\beta = 0.65$, the optimal choice is to become a worker, and investment in patience in this range is such that all subsequent generations are workers too. Thus, investment in patience is constant over this range, as displayed in the lower panel. Above the threshold, the optimal choice for both the current and future generations

[9] The parametrization is as in the balanced growth computations in Section 1.3.3 with the equilibrium fraction of entrepreneurs given by $\lambda = 0.35$.

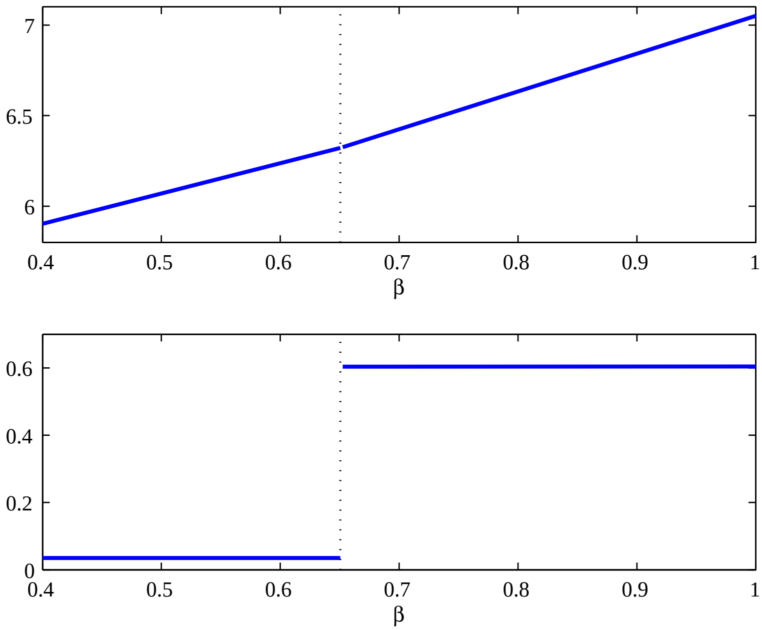


Figure 1.1 Example of value function (upper panel) and policy function for *I* (lower panel).

is to become entrepreneurs. Consequently, investment in patience is constant over this range as well, but considerably higher compared to worker dynasties. The value function has a kink at $\beta = 0.65$ and becomes steeper, because the return to patience is higher for entrepreneurs given their steeper lifetime income profiles. The differential investment results in a substantial gap in patience across occupations in the balanced growth path, with a discount factor $\beta^W = 0.55$ for workers, and $\beta^E = 0.95$ for entrepreneurs.

1.3.3 Multiplicity of Balanced Growth Paths with Endogenous Patience

Given the preceding analysis, it is clear that there is no balanced growth path in which all dynasties have identical preferences, and in which there are positive fractions of both entrepreneurs and workers. The reason is that the entrepreneurs have a steeper income profile, given the need to acquire skills when young and the entrepreneurial return that is received when old. This steeper income profile implies that parents of entrepreneurs have a higher incentive to invest in patience compared to parents of workers. Moreover, in any given period the population will sort such that the more patient individuals become entrepreneurs and the less patient become workers. Finally, because of persistence of patience within dynasties, occupations also will be persistent within dynasties.

Hence, a balanced growth path has the property that the two groups are characterized by different preferences, patient entrepreneurs and impatient workers. Given the patience gap between these groups, at least one of them will strictly prefer their own occupation over the alternative, both for themselves and for their children. In fact, generically there exists a continuum of balanced growth path where both workers and entrepreneurs strictly prefer their own occupation, and where the fraction of entrepreneurs, the entrepreneurial premium, and the equilibrium growth rate vary across growth paths. For given parameters, the balanced growth path that is reached depends on initial conditions. More generally, the multiplicity of balanced growth paths opens up the possibility of history dependence and a persistent impact of policies or institutions on the performance of an economy.

To illustrate these results, we focus on the case where preferences are not persistent, $\delta = 1$. We would like to characterize the set of balanced growth paths in terms of the growth rate g, the entrepreneurial premium η, and the patience levels β^W and β^E of workers and entrepreneurs. From (1.20) and (1.21), we know that the investments in patience l^W and l^E by workers and entrepreneurs have to satisfy:

$$\begin{aligned} -\chi'(l^W) &= z(1+g)^{2(1-\sigma)} f'(l^W), \\ -\chi'(l^E)\psi^{1-\sigma} &= z(1+g)^{2(1-\sigma)} \eta^{1-\sigma} f'(l^E), \end{aligned}$$

and we have $\beta^W = f(l^W)$ and $\beta^E = f(l^E)$. Here, focusing on the $\delta = 1$ case implies that the choice of future patience depends only on today's occupational choice, but not directly on the current patience.

The balanced growth values of the value functions (1.17) and (1.18) are:

$$\begin{aligned} v^W &= \frac{\chi(l^W) + \beta\,(1+g)^{1-\sigma}}{1 - z\,(1+g)^{1-\sigma}}, \\ v^E &= \frac{\chi(l^E)\psi^{1-\sigma} + \beta\,((1+g)\,\eta)^{1-\sigma}}{1 - z\,(1+g)^{1-\sigma}}. \end{aligned}$$

In the balanced growth path, each group has to prefer their own occupation over the alternative, for the present generation and future descendants. In particular, there are four constraints to consider. The first is that a person with patience β^E prefers entrepreneurship for all members of the dynasty over everyone being a worker:

$$v^E \geq \chi(l^W) + \beta^E\,(1+g)^{1-\sigma} + z(1+g)^{1-\sigma} v^W. \tag{1.22}$$

The right-hand side has two components, because the first generation still has patience β^E, with all following generations in the deviation would have patience β^W. The second constraint is that entrepreneurship for all generations is preferred to the first generation being an entrepreneur, but all following generations switching to being workers. This constraint can be written as:

$$\begin{aligned} v^E \geq{}& \chi(l^{EW})\psi^{1-\sigma} + \beta^E\,((1+g)\,\eta)^{1-\sigma} + z(1+g)^{1-\sigma}\left(\chi(l^W) + \beta^{EW}(1+g)^{1-\sigma}\right) \\ &+ z^2(1+g)^{2(1-\sigma)} v^W. \end{aligned} \tag{1.23}$$

Here l^{EW} and β^{EW} are the investment and patience level that are optimal given that path of occupational choices, characterized by:

$$-\chi'(l^{EW})\psi^{1-\sigma} = z(1+g)^{2(1-\sigma)}f'(l^{EW}).$$

and $\beta^{EW} = f(l^{EW})$. The parallel constraints for worker dynasties with patience β^{W} are given by:

$$\chi(l^{E})\psi^{1-\sigma} + \beta^{W}((1+g)\eta)^{1-\sigma} + z(1+g)^{1-\sigma}v^{E} \leq v^{w}. \tag{1.24}$$

and:

$$\begin{aligned}\chi(l^{WE}) + \beta^{W}(1+g)^{1-\sigma} + z(1+g)^{1-\sigma}\left(\chi(l^{E})\psi^{1-\sigma} + \beta^{WE}((1+g)\eta)^{1-\sigma}\right)\\ +z^{2}(1+g)^{2(1-\sigma)}v^{E} \leq v^{W},\end{aligned} \tag{1.25}$$

where l^{WE} and β^{WE} are characterized by:

$$-\chi'(l^{WE}) = z(1+g)^{2(1-\sigma)}\eta^{1-\sigma}f'(l^{WE}),$$

and $\beta^{WE} = f(l^{WE})$. It can now be shown that a continuum of balanced growth paths exists. Because of the gap in balanced growth preferences, when one occupational group is just indifferent between their occupation and the alternative, the other group strictly prefers their own occupation. It is therefore possible to raise the return of the indifferent group in some range so that both groups strictly prefer to stay in their own occupation. The potentially binding constraints are given by (1.23) and (1.25). The following lemma contains the main result underlying the multiplicity of balanced growth paths.

Lemma 1. *When the entrepreneurial premium η in the balanced growth path is such that* (1.23) *holds as an equality, then* (1.22), (1.24), and (1.25) *hold as strict inequalities.*

Building on this lemma, we can now establish the main result:

Proposition 5. *If there exists a balanced growth with path a fraction of entrepreneurs λ such that $0 < \lambda < 1$, there exists a continuum of additional balanced growth paths with different fractions of entrepreneurs and thus different growth rates.*

That is, there are multiple balanced growth paths unless the only feasible balanced growth path features a corner solution with all agents choosing the same profession.

We have focused on the $\delta = 1$ case for analytical convenience. When there is direct persistence in patience across generations ($\delta < 1$), the forces generating multiple balanced growth paths are strengthened even more, and generally a wider range of rates of entrepreneurship and economic growth can be long-run outcomes. Figure 1.2 illustrates this with a computed example. The parameter values used are as follows: $z = 0.5, \sigma = 0.5, \xi = 3, \alpha = 0.3, \psi = 0.5$. The cost function for investing in patience is given by $\chi(l) = 1 - l$, and the law of motion for patience is parameterized as:

$$\beta' = (1-\delta)\beta + \delta\tilde{\beta} + \theta_1 l^{\theta_2},$$

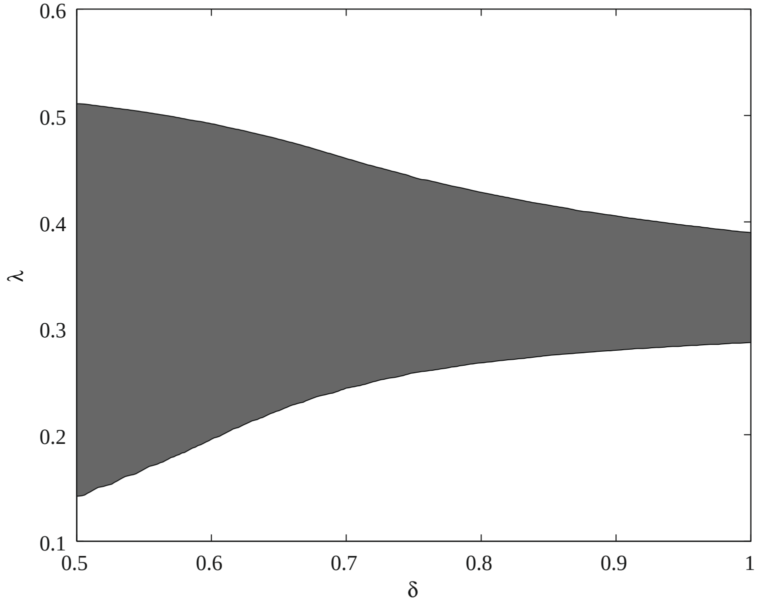


Figure 1.2 Range of balanced growth paths for different δ.

where we set $\tilde{\beta} = 0.5$ and $\theta_2 = 0.8$. We computed outcomes for a variety of values of the persistence parameter δ. For $\delta = 1$, we set $\theta_1 = 1$, and for lower δ the value of θ_1 is adjusted, to hold the impact of investing in patience on utility constant in the balanced growth path (so that changing δ does not lead to a level shift in patience).

For these parameters, Figure 1.2 plots the range for λ (the fraction of entrepreneurs in the population) that can be supported as a balanced growth path. At $\delta = 1$ (no direct persistence in patience across generations), the balanced growth level of λ varies between 0.29 and 0.39, which corresponds to growth rates (per generation) between $g = 0.87$ and $g = 1.27$, or, if a generation is interpreted to last 25 years, between 2.5 and 3.3% per year. As we lower δ and make patience more persistent, the range of balanced growth paths widens. At $\delta = 0.5$, λ can vary between 0.15 and 0.51 in the balanced growth path, which corresponds to annual growth rates between 1.5 and 3.8% per year.

Figure 1.3 demonstrates what the law of motion for patience looks like in the balanced growth path for different values of λ. In all panels, the persistence of patience is set to $\delta = 0.8$. In the top panel, we set $\lambda = 0.26$, which is close to the lowest fraction of entrepreneurs that can be sustained in a balanced growth path. In this growth path, the return to entrepreneurship is high. The law of motion for patience intersects the 45-degree line twice, where the lower intersection corresponds to the long-run patience of workers, and the higher intersection corresponds to entrepreneurs. Given high returns to entrepreneurship, dynasties that start out with patience that is only a little higher

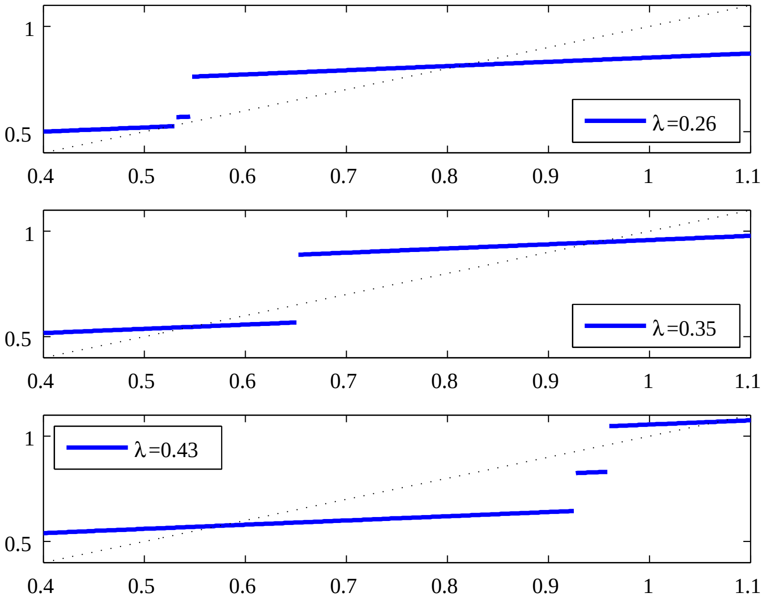


Figure 1.3 Laws of motion for β in balanced growth paths for $\delta = 0.8$ and different values of λ.

than the long-run patience of workers, ultimately converge to entrepreneurship. The law of motion has three linear segments, where the bottom one corresponds to worker dynasties and the top one to entrepreneur dynasties. The (small) middle segment pertains to dynasties where the current generation consists of workers who invest sufficiently in patience for all following generations to switch to entrepreneurship. In the middle panel, we set $\lambda = 0.35$. Here the law of motion has only two segments. All dynasties are either workers or entrepreneurs forever; there are no transitions between the occupations. The bottom panel for $\lambda = 0.43$ corresponds to a low return to entrepreneurship. The law of motion is a mirror image of the top panel. There are three segments, where the middle segment now corresponds to dynasties where the current generation consists of entrepreneurs, but all subsequent ones will be workers. Comparing across the levels of λ, it is apparent that as we move to higher levels of λ the long-run levels of patience (i.e. the intersections with the 45-degree line) increase both for workers and for entrepreneurs. This is because a higher λ implies a higher growth rate, which results in steeper income profiles for both professions, and thus more investment in patience.

1.3.4 Implications of Multiplicity of Balanced Growth Paths

Taken at face value, our finding of multiplicity of balanced growth paths implies that different economies, although characterized by identical technological parameters, can experience permanently different growth rates, driven by cultural differences across their populations. Of course, cultural differences themselves are endogenous in our theory. From this perspective, the theory suggests the possibility of path dependence, that is, a country's success at entrepreneurship and innovation may depend on the cultural and economic makeup of the country at the onset of modern economic growth. This theme is explored in more detail in Doepke and Zilibotti (2008), where we explicitly model the transition of an economy with endogenous preferences from a stagnant, pre-industrial economy to capital-driven growth. In that paper, the distribution of preferences at the onset of modern growth depends on the nature of pre-industrial occupations in terms of lifetime income profiles and the distribution of land ownership. Combining the approach of Doepke and Zilibotti (2008) with the theory outlined here would lead to the prediction that the nature of the pre-industrial economy can have long-term repercussions for economic development.

Another implication of multiplicity of balanced growth paths is that policies or institutions that affect preferences can have a long-term impact on economic growth. Consider a country that imposes high taxes on entrepreneurs or discourages entrepreneurship through other means, as in the centrally planned economies of Eastern Europe during the 20th century. Over time, such policies would shift the culture of the population toward being less future-oriented with a lower propensity for entrepreneurship. Consider now the transition of the economy when the political constraints on entrepreneurship are removed. We would expect to observe a small class of entrepreneurs gaining high returns, but lower rates of entrepreneurship and a lower rate of economic growth compared to a country undergoing a similar transition from more favorable initial cultural conditions.

The model can also be extended to allow for open economies. The simplest case is that of a world economy in which trade across borders is frictionless, so that all goods are traded at the same price, and workers and entrepreneurs get the same returns regardless of where they live. In such an environment, initial cross-country differences would manifest themselves in permanent differences in rates of entrepreneurship and innovation across countries, even though ultimately all countries would benefit from innovation (and experience the same growth rates) because of integrated markets.

1.3.5 The Model with Financial Markets

In the sections above, we showed that workers and entrepreneurs face different incentives for investing in patience, because entrepreneurs face a steeper income profile. However, the difference in the income profile would not matter if people could use financial markets to smooth consumption. A steep income profile directly translates into a steep utility profile only if financial markets are absent or incomplete.

To illustrate this point, consider the opposite extreme of perfect financial markets, i.e. people can borrow and lend at a fixed interest rate R subject to a lifetime budget constraint. For simplicity, we abstract from financial bequests. The only occupations that are chosen in equilibrium are now those that maximize the present value of income,

$y_1 + y_2/R$. Therefore, the lifetime returns of being a worker and being an entrepreneur have to be equalized:

$$\omega^W + \frac{(1+g)\omega^W}{R} = \psi\omega^W + \frac{(1+g)\omega^E}{R},$$

which implies that:

$$\eta = 1 + \frac{(1-\psi)\,R}{1+g} = 1 + \frac{(1-\psi)\,R}{1+\lambda\xi}.$$

The equilibrium condition (1.13) continues to hold, hence:

$$\eta = (1-\alpha)\,\alpha^{\frac{\alpha}{1-\alpha}}\xi\,\frac{2-(2-\psi)\lambda}{1+\alpha^{\frac{1}{1-\alpha}}\xi\lambda}.$$

Combining these equations yields a relationship between the proportion of entrepreneurs, λ(or, alternatively, the growth rate), and the market interest rate:

$$1 + \frac{(1-\psi)\,R}{1+g} = (1-\alpha)\,\alpha^{\frac{\alpha}{1-\alpha}}\frac{2\xi-(2-\psi)g}{1+\alpha^{\frac{1}{1-\alpha}}g}. \tag{1.26}$$

Since workers and entrepreneurs have the same lifetime income, it is sufficient to consider the individual saving decision of one group, e.g. the workers:

$$\max_s \frac{\left(\omega^W - s\right)^{1-\sigma}}{1-\sigma} + \frac{\beta}{\chi}\frac{\left(Rs + \omega^W\,(1+g)\right)^{1-\sigma}}{1-\sigma}.$$

The solution yields a standard Euler equation:

$$\frac{Rs + \omega^W\,(1+g)}{\omega^W - s} = \left(\frac{\beta}{\chi}R\right)^{\frac{1}{\sigma}}.$$

Hence, denoting by c^Y and c^O the consumption of the young and the old, respectively,

$$c^Y = \omega^W \frac{1+g+R}{R + \left(R\frac{\beta}{\chi}\right)^{\frac{1}{\sigma}}},$$

$$c^O = \left(R\frac{\beta}{\chi}\right)^{\frac{1}{\sigma}}\omega^W \frac{1+g+R}{R + \left(R\frac{\beta}{\chi}\right)^{\frac{1}{\sigma}}}.$$

Given this solution to the saving problem, the optimal investment in patience is given by:

$$l(\beta,g) = \underset{0\leq l\leq 1}{\operatorname{argmax}} \left\{ \left(\omega^W\right)^{1-\sigma} \left(\frac{1+g+R}{R+\left(R\frac{\beta}{\chi(l)}\right)^{\frac{1}{\sigma}}} \right)^{1-\sigma} \right.$$
$$\left. \left(\chi\left(l\right) + \beta \left(\frac{\beta}{\chi\left(l\right)} R \right)^{\frac{1-\sigma}{\sigma}} \right) + z\left(1+g\right)^{1-\sigma} v(\beta') \right\}.$$

The policy function, $l(\beta,g)$ determines the equilibrium law of motion of β, and hence the steady-state value of β. This is a function of g and R.

So far we have found two equilibrium conditions for three endogenous variables, g, β, and R. The model is closed by an asset market-clearing condition that pins down the interest rate. We assume that the young cannot borrow from the old, since the latter cannot obtain repayment within their lifetime. Hence, all borrowing and lending takes place between workers and entrepreneurs of a given cohort. The market-clearing condition then yields $s^W + s^E = 0$, or:

$$\left(R\frac{\beta}{\chi}\right)^{\frac{1}{\sigma}} - \left(1+g\right) + \psi \left(R\frac{\beta}{\chi}\right)^{\frac{1}{\sigma}} - \eta\left(1+g\right) = 0$$
$$\left(R\frac{\beta}{\chi}\right)^{\frac{1}{\sigma}} \left(1+\psi\right) = \left(1+g\right)\left(1+\eta\right).$$

This is the third of the conditions that jointly pin down g, β, and R in the balanced growth path.

The next proposition summarizes our main findings for the model with a perfect market for borrowing and lending.

Proposition 6. *When a perfect market exists for borrowing and lending within generations, the only occupations that are chosen in equilibrium are those that maximize the present value of income. The set of optimal occupations is independent of patience β. If both occupations yield the same present value of income, investment in patience l is independent of which occupation is chosen.*

The intuition for this result is simple: with perfect borrowing and lending, every adult will choose the income profile that yields the highest present value of income, regardless of patience.[10] The proposition shows that at least some degree of financial market imperfection is necessary for occupational choice and investments in patience to be interlinked.

[10] In the model of the previous section, general equilibrium forces ensure that there exist equilibria with positive growth where both occupations yield the same present value of income.

A positive implication of this finding is that the degree of discount-factor heterogeneity in a population depends on the development of financial markets. In an economy where financial markets are absent, workers and entrepreneurs face very different incentives for investing in patience, and consequently the gap in patience across occupations is large in the balanced growth path. In contrast, in a modern economy with deeper financial markets we would expect to observe smaller cultural differences across occupations.

1.4. ENDOGENOUS CULTURE II: KNIGHT AND THE TRANSMISSION OF RISK TOLERANCE

In our economic environment, entrepreneurs face not only a steeper income profile than workers; they also face risk, provided that $\nu > 0$. As a result, risk preferences too should be relevant for explaining entrepreneurship, in line with Frank Knight's characterization of risk-taking entrepreneurs (see Knight, 1921, and more recently Kihlstrom and Laffont, 1979; Vereshchagina and Hopenhayn, 2009). In this section, we provide a formal analysis of this possibility.

1.4.1 Endogenizing Risk Preferences

To facilitate our analysis of endogenous risk preferences, we focus on a period utility function with mean-variance preferences. That is, the period utility function evaluating (potentially stochastic) consumption c is given by:

$$U(c) = E(c) - \sigma\sqrt{Var(c)}, \tag{1.27}$$

where $E(c)$ is expected consumption and $Var(c)$ is the variance of consumption, and σ is a measure of risk aversion. The specific functional form is chosen to be consistent with balanced growth.[11] The utility function implies that people are always better off with a lower risk aversion, i.e. a higher risk tolerance. However, as in our analysis of patience, there is a cost of investing in children's preferences. The effort that a parent of generation t spends on raising the child's risk tolerance is denoted by l_t. Total utility is then given by:

$$\chi(l_t)\left(E(c_{t,1}) - \sigma_t\sqrt{Var(c_{t,1})}\right) + \beta\left(E(c_{t,2}) - \sigma_t\sqrt{Var(c_{t,2})}\right) + zV_{t+1}(\sigma_{t+1}(l_t)),$$

where χ is a strictly decreasing, strictly concave, and differentiable function, and effort is bounded by $0 \leq l_t \leq 1$. The child's risk preferences are given by:

$$\sigma_{t+1}(l_t) = (1-\delta)\sigma_t + \delta\sigma_{\max} - f(l_t), \tag{1.28}$$

where f is an increasing and strictly concave function with $f(0) = 0$, and δ satisfies $0 < \delta \leq 1$. Here, $\sigma_{\max}$ denotes the level of risk aversion exhibited by a dynasty that never

[11] While this utility function is not of the expected-utility form, the main results carry over to expected utility as well. For an analysis of the usual CRRA case see Doepke and Zilibotti (2012).

invests in risk tolerance. If $\delta < 1$ there is some direct persistence in preferences across generations.

Let w^W denote the workers' wage, and η the ratio of the expected return of entrepreneurs to this wage. To simplify the analysis, we assume that the risk of entrepreneurship takes the form that with probability κ, the entrepreneur is successful and earns a positive return, whereas with probability $1-\kappa$ the entrepreneur fails and earns zero. That is, in the notation of Section 1.2.1 we have:

$$\nu = \frac{1-\kappa}{\kappa},$$

so that if successful, the earnings are:

$$(1+\nu)\eta w^W = \frac{\eta w^W}{\kappa},$$

whereas with probability $1-\kappa$ entrepreneurial output is zero. The mean return is then ηw^W, and the variance of the return is given by:

$$\begin{aligned} Var(c^E) &= \kappa\left(\frac{\eta w^W}{\kappa} - \eta w^W\right)^2 + (1-\kappa)\left(\eta w^W\right)^2 \\ &= \frac{1-\kappa}{\kappa}\left(\eta w^W\right)^2. \end{aligned}$$

Thus, the old-age felicity of an entrepreneur is given by:

$$E(c^E) - \sigma\sqrt{Var(c^E)} = \eta w^W\left(1 - \sigma\sqrt{\frac{1-\kappa}{\kappa}}\right).$$

1.4.2 Transmission of Risk Preferences in the Balanced Growth Path

We now consider balanced growth paths. People choose both a career, and whether and how much to invest in their child's risk tolerance. We analyze the individual decision problem under the assumption that the economy is in a balanced growth path, so the entrepreneurial premium is constant, and wages and profits grow at the constant rate g. The decision problem admits a recursive representation with the risk aversion parameter, σ, serving as the state variable of the dynasty. As in our analysis of endogenous patience, the state of technology N_t is in principle a second state variable. However, the linear homogeneity of utility in expected consumption allows us to express the value function at time t in a multiplicatively separable form:

$$V_t(\sigma_t, N_t) = \frac{N_t w_0^W}{N_0} v(\sigma_t),$$

where $v(\sigma_t) = V_t(\sigma_t, 1)$ satisfies the following set of Bellman equations:

$$v(\sigma) = \max\left\{v^W(\sigma), v^E(\sigma)\right\}, \tag{1.29}$$

$$v^W(\sigma) = \max_{0\leq l\leq 1} \left\{\chi(l) + \beta(1+g) + z(1+g)\,v(\sigma')\right\}, \tag{1.30}$$

$$v^E(\sigma) = \max_{0\leq l\leq 1} \left\{\chi(l)\psi + \beta(1+g)\,\eta\left(1-\sigma\sqrt{\frac{1-\kappa}{\kappa}}\right) + z(1+g)\,v(\sigma')\right\}, \tag{1.31}$$

the maximizations in (1.30) and (1.31) being subject to:

$$\sigma' = (1-\delta)\sigma + \delta\sigma_{\max} - f(l). \tag{1.32}$$

Here, v^W and v^E are the present-value utilities conditional on choosing to be a worker or an entrepreneur, respectively, and v yields the optimal occupational choice.

Since l is bounded and $\delta > 0$, there is a lower bound $\sigma_{\min}$ for feasible levels of risk aversion. Note that, depending on f and $\delta, \sigma_{\min}$ could be negative, corresponding to risk-loving individuals who would choose a risky lottery over a safe one with the same expected return. For a given growth rate g and average return to entrepreneurship η, the decision problem is a standard dynamic programming problem with a single state variable in the interval $[\sigma_{\min}, \sigma_{\max}]$. The following propositions summarize the properties of the value function and the associated optimal policy functions.

Proposition 7. *The system of Bellman equations (1.29)–(1.31) has a unique solution. The value function v is decreasing and convex in σ. The optimal occupational choice is either to be a worker for any σ, or to be an entrepreneur for any σ, or there exists a $\bar{\sigma}$ such that people with high risk aversion, $\sigma > \bar{\sigma}$, strictly prefer to be workers; people with low risk aversion, $\sigma < \bar{\sigma}$, strictly prefer to be entrepreneurs; and people with $\sigma = \bar{\sigma}$ are indifferent. The optimal investment in risk tolerance $l = l(\sigma)$ is non-increasing in σ.*

Proposition 8. *The state space $[\sigma_{\min}, \sigma_{\max}]$ can be subdivided into (at most) countably many closed intervals $[\underline{\sigma}, \overline{\sigma}]$ such that over the interior of any range $[\underline{\sigma}, \overline{\sigma}]$ the occupational choice of each member of the dynasty (i.e. parent, child, grandchild, and so on) is constant and unique (though possibly different across generations), and $l(\sigma)$ is constant and single-valued. The value function $v(\sigma)$ is piecewise linear, where each interval $[\underline{\sigma}, \overline{\sigma}]$ corresponds to a linear segment. Each kink in the value function corresponds to a switch from being a worker to being an entrepreneur by a present or future member of the dynasty. At a kink, the optimal choices of occupation and l corresponding to both adjoining intervals are optimal (thus, the optimal policy functions are not single-valued at a kink). If there is an interval $[\underline{\sigma}, \overline{\sigma}]$ such that over this interval all present and future members of the dynasty are workers, the value function $v(\sigma)$ is constant over this interval, and there is no investment in risk tolerance: $l(\sigma) = 0$.*

The proofs of the propositions (omitted) are analogous to the proofs of Propositions 2 and 3. The final part of Proposition 8 arises because workers do not face any risk, so that in all-worker dynasties utility is independent of risk preferences, and the return on investing in risk tolerance is zero.

The next proposition characterizes the dynamics of risk aversion within dynasties.

Proposition 9. *The law of motion of σ is described by the following difference equation:*

$$\sigma' = g(\sigma) = (1-\delta)\,\sigma + \delta\sigma_{\max} - f\,(l\,(\sigma))\,,$$

where $l\,(\sigma)$ is a non-increasing step-function (as described in Proposition 8). Given an initial condition σ_0, risk aversion in the dynasty converges to a constant σ where parents and children choose the same profession. If the dynasty ends up as a worker dynasty, the limit for risk aversion is given by $\sigma = \sigma_{\max}$.

The proof (omitted) is analogous to the proof of Proposition 4.

We have already established that in worker dynasties the return to investing in risk tolerance is zero, so that these dynasties do not invest in risk tolerance and hence we have $l^W = 0$ and $\sigma^W = \sigma_{\max}$. For entrepreneurs, in contrast, the return to investing in risk tolerance is positive. If their choice of investment is interior, the investment l^E is characterized by a first-order condition:

$$-\chi'(l^E)\psi = \frac{z(1+g)^2\beta\eta\sqrt{\frac{1-\kappa}{\kappa}}f'(l^E)}{1-z(1+g)(1-\delta)}. \tag{1.33}$$

Here, the left-hand side is strictly increasing in l, and the right-hand side is strictly decreasing. The optimal parental investment in risk tolerance is increasing in the entrepreneurial premium η, the growth rate g, and the entrepreneurial risk $1-\kappa$.

Parallel to our analysis of endogenous patience, the gap in risk preferences between workers and entrepreneurs leads to a multiplicity of balanced growth paths. There can be long-run differences in growth rates across countries, where faster-growing countries are characterized by a larger group of entrepreneurial individuals with low risk aversion. As in the discussion of Section 1.3.4, the multiplicity of balanced growth paths can give rise to path dependence, to persistent effects of institutions and policies that affect risk-taking, and (in an open-economy context) to specialization of certain groups or countries in innovative and risk-taking activities. Also, the development of financial markets once again interacts with endogenous culture and growth, as discussed in Section 1.3.5 for the patience case. For example, for a given distribution of preferences, better risk-sharing institutions (e.g. through insurance markets or tax and transfer policies) can make entrepreneurship more attractive to individuals with high risk aversion, and thereby lead to faster economic growth. However, there is also a downside to the provision of more insurance. In the limit with perfect risk sharing there would be no incentive to invest in risk tolerance, and consequently over time the population would end up more risk averse compared to a country where less insurance is available. Consider now the arrival of a new technology that involves some uninsurable idiosyncratic risk. The population in the well-insured country would be less likely to pick up such new opportunities, and thus might fall back over time compared to a less well-insured, but more risk tolerant and innovative country.

1.5. PATERNALISTIC MOTIVES FOR PREFERENCE TRANSMISSION

Up to this point, in our model of preference transmission parents are motivated solely by altruism, i.e. they evaluate the welfare of the children using the same utility function that drives the children's choices. However, preference transmission could be driven also by paternalistic motives. This is the case when there are potential disagreements between parents and children about optimal choices, and parents use preference transmission as a tool to influence their children's choices.

The paternalistic motive is especially salient in the relationship between parents and adolescent children. It is common for parents to desire to control the tendency of adolescents to take risks parents disapprove of, such as reckless driving, the use of drugs or alcohol, or risky sexual behavior.[12]

1.5.1 Allowing for Conflict Between Parents and Children

To analyze how paternalistic motives affect preference transmission, we extend the model by allowing children to make an additional choice at a young age, denoted by x, that depends on risk preferences. For simplicity, we assume this choice to be orthogonal to the adult occupational choice, i.e. x does not affect the relative return of the adult occupations or the child's ability to enter either occupation. The environment is a simplified version of Doepke and Zilibotti (2012), where we propose a general theory of parenting style related to paternalism.

Children choose from a set of feasible lotteries so as to maximize the felicity function $U_y(x, \sigma)$, whereas their parents evaluate the choice with a different felicity function, $U(x, \sigma)$, where σ denotes the adult's risk aversion parameter. As a concrete example, let the choice of the lottery x result in a random consumption process $c(x)$, and consider parental preferences given by:

$$U(x, \sigma) = E(c(x)) - \sigma\sqrt{Var(c(x))},$$

as in (1.27), whereas the child's preferences are given by:

$$U_y(x, \sigma) = E(c(x)) - (\sigma - \xi)\sqrt{Var(c(x))}.$$

That is, children have intrinsically lower risk aversion (which is consistent with empirical evidence), where $\xi > 0$ captures the gap in risk aversion between the young and the old. For a given σ, children would choose riskier lotteries x than what their parents would prefer.

[12] There is well-documented evidence that children are especially prone to risk-taking. For instance, in a series of laboratory experiments carried out in New Mexico, Harbaugh et al. (2002) it was found that 70–75% of children in the 5–8 year age group chose fair gambles with varying odds over a certain outcome, while only 43–53% of the adults did.

We denote by $x(\sigma)$ optimal choice from the children's standpoint. This choice is given by:

$$x(\sigma) = \underset{x}{\operatorname{argmax}} \left\{ U_y (x, \sigma) \right\}.$$

This choice is static, because the choice of x does not have dynamic consequences. Assuming the choice set to be continuous and differentiable implies:

$$\partial U_y (x(\sigma), \sigma) / \partial x = 0.$$

We now turn to the parents' decision problem. The utility of adult workers and entrepreneurs can be written as:

$$v^W(\sigma) = \max_{0 \le l \le 1} \left\{ \chi(l) + \beta(1+g) + z(1+g) W(\sigma', \sigma) \right\},$$

$$v^E(\sigma) = \max_{0 \le l \le 1} \left\{ \chi(l)\psi + \beta(1+g)\eta \left(1 - \sigma \sqrt{\frac{1-\kappa}{\kappa}} \right) + z(1+g) W(\sigma', \sigma) \right\},$$

where $W(\sigma', \sigma)$ captures the utility that the parents derive from their children. This function is given by[13]:

$$W\left(\sigma', \sigma\right) = U\left(x(\sigma'), \sigma\right) + \beta \max \left\{ v^W\left(\sigma'\right), v^E\left(\sigma'\right) \right\}.$$

Notice that $x(\sigma')$ is written as a function of σ'. This is because the parent cannot control x directly, but must take as given the child's decision based on the child's preference parameter σ'. The choice σ' is constrained by the law of motion:

$$\sigma' = (1-\delta)\sigma + \delta\sigma_{\max} - f(l).$$

1.5.2 Optimal Preference Transmission with Paternalistic Motives

Consider a parent who anticipates her child to become an entrepreneur, and assume, for simplicity, $\delta = 1$. If the optimal l is interior, the following first-order condition obtains:

$$\chi'\left(l^E\right)\psi = z(1+g)\left(\underbrace{\frac{\partial U\left(x(\sigma'), \sigma\right)}{\partial x} \frac{\partial x}{\partial \sigma'}}_{\text{paternalistic motive}} + \beta \frac{\partial v^E}{\partial \sigma} \right) f'\left(l^E\right).$$

Relative to the model of Section 1.4, a new term appears in the first-order condition which captures the paternalistic motive. This terms vanishes whenever there is no disagreement between parents and children, i.e. when $U = U_y$ and $\sigma = \sigma'$, because in

[13] In Doepke and Zilibotti (2012), we consider a formulation with partial paternalism, where the W function takes the form:

$$W(\sigma', \sigma) = qU_y(x(\sigma'), \sigma') + (1-q)U(x(\sigma'), \sigma) + \beta \max \{v^W(\sigma'), v^E(\sigma')\}.$$

this case we have:

$$\frac{\partial U\left(x(\sigma'),\sigma\right)}{\partial x}=0,$$

i.e. the envelope theorem applies. Likewise, the paternalistic motive would also be mute if a fixed choice of x were imposed on the child, because this would imply $\partial x/\partial\sigma' = 0$. In contrast, paternalism does affect the parent's decision problem whenever three conditions are all satisfied: There is disagreement between parent and child regarding the choice of x; the child is free to choose x; and the child's choice depends on the endogenous preference parameter σ'. In this case, it is valuable for the parent to distort the child's preferences in order to induce the child to choose an x that is more to the parent's liking. Alternatively, if the option were available, the parent would impose restrictions on the ability of the child to choose freely. When forming a child's preferences, parents realize that reducing the child's risk tolerance comes at the expense of the child's future utility, implying a tradeoff for the altruistic parents. Thus, in general the parent will strike a compromise, and accept that the child chooses an x that is different from the parents' most preferred option.

The discussion above assumes that the parental choice of σ' (via l) does not affect the child's occupational choice. However, if the paternalistic motive is sufficiently strong, the occupational choice of the child may be affected. More formally, if $\hat{\sigma}$ denotes the risk aversion parameter such that $v^W(\hat{\sigma}) = v^E(\hat{\sigma})$, it is possible that absent the paternalistic motive the parent would choose $\sigma' < \hat{\sigma}$, inducing the child to become an entrepreneur, whereas the paternalistic motive induces a choice $\sigma' > \hat{\sigma}$, implying that the child will choose to be a worker. This scenario is more likely if ξ (i.e. the child's intrinsic risk-loving bias) is large, and if the set of feasible lotteries x among which the child can choose includes choices the parent would strongly disapprove of. In practice, this choice set would depend on various features of the environment in which the adolescent grows up. For instance, adolescents living in areas infested by juvenile gangs are more exposed to risky choices than are children in safe middle-class neighborhoods, where risky choices are limited to more innocuous transgressions. An implication of this analysis, which we explore in more detail in Doepke and Zilibotti (2012), is that families living in areas exposed to acute juvenile risk will emphasize values that are less conducive to an entrepreneurial spirit. When integrated into the general equilibrium model of Section 1.2.1, the theory bears the prediction that countries where juvenile risk is more severe will have a smaller equilibrium proportion of entrepreneurs as well as larger risk premia.

1.6. LITERATURE REVIEW

1.6.1 Cultural Transmission, Human Capital, and Non-cognitive Skills

The theory presented in the previous sections provides a two-way link between the economic environment and preferences. A pioneering contribution to this literature is Becker and Mulligan (1997), which formalizes a model where people choose their own preferences rather than those of their children. In Mulligan (1997), parents choose their own level of altruism toward their children. Along similar lines, in Haaparanta and Puhakka (2004), agents invest in their own patience and in health. Doepke and Zilibotti (2008) (discussed in more detail below) provide the first theory where altruistic parents shape their children's preferences in order to "best prepare" them for the economic environment in which they will operate.

In these studies, as in our model above (except in the extension of Section 1.5), parents evaluate their children's wellbeing using their children's preferences. Namely, parents choose their investments in preference optimally by maximizing their children's utility. There is no explicit desire of parents to preserve their own culture or to instill values that they regard as intrinsically good or moral. In particular, parents may choose to teach their children preferences that differ from their own. In contrast, a number of recent studies postulate that cultural transmission hinges on a form of "imperfect empathy" (see Bisin and Verdier, 2001; Hauk and Saez-Marti, 2002; Gradstein, 2007; Klasing, 2012; Saez-Marti and Sjoegren, 2008; Tabellini, 2008; and Saez-Marti and Zenou, 2011). According to this approach, parents use their own preferences to evaluate the children's utility and are driven by a desire to make the children's values similar to their own. The two approaches and their differences are reviewed in more detail by Saez-Marti and Zilibotti (2008).[14]

In the Beckerian approach, parents transmit traits to their children that are supposed to make them fit for success. Thus, investment in preference transmission resembles a standard human capital investment. From this perspective, preferences are closely related to what the recent labor literature has labeled "non-cognitive skills." These skills determine how well people can focus on long-term tasks, behave in social interactions, and exert self-restraint, and include patience, perseverance, and self-discipline, among others. Recent empirical studies emphasize the importance of such human assets for economic success (see Heckman et al. 2006; Segal, 2013).

Within the realm of non-cognitive skills, we emphasize the role of patience and of the propensity to take risks. The importance of patience for economic success has been documented by experimental studies. A longitudinal study by Mischel et al. (1992) finds that individuals who were more patient as children were subsequently more likely to acquire formal education, to choose market-oriented occupations, and to earn higher income. More recently, Sutter et al. (2013) found that measures of time preferences of young people aged 10–18 elicited through experiments predict saving behavior, smoking and alcohol abuse, BMI, and conduct at school. Reyes-Garcia et al. (2007) study the effect of patience on economic outcomes among the Tsimanes, an Amazonian tribal society that only recently transitioned from self-sufficiency to a market economy. They found

[14] Our analysis in Section 1.5.2 and in Doepke and Zilibotti (2012) provides a bridge between these two approaches. Our analysis proposes an explicit microfoundation of the child-adult preference conflict, whereas in the existing literature imperfect empathy is postulated as a primitive.

that individuals who were already more patient in the pre-market environment (when patience was a latent attribute with no effect on individual success) acquired on average more education and engaged more often in entrepreneurial activity when the society introduced markets.[15]

The importance of the propensity to take risk for entrepreneurship has been emphasized, among others, by Kihlstrom and Laffont (1979). Several studies point to robust evidence that risk tolerant people are more likely to become entrepreneurs; see, e.g. Van Praag and Cramer (2001), Cramer et al. (2002), and Kan and Tsai (2006).

The evidence discussed above leaves open the extent to which patience and risk tolerance hinge on parental effort or on the influence of the environment, as opposed to being genetically inherited. The long-standing debate among anthropologists and population geneticists on the role of nature versus nurture has reached no clear conclusion.[16] Both genes and culture appear to be important, likely in a non-linear interactive fashion. The recent economic literature has explored, in different contexts, both the evolutionary selection and the cultural transmission mechanisms. For instance, recent studies focusing on economic development from a very long-run perspective have emphasized the importance of Darwinian evolution of preferences and of genetic diversity for the process of development (see, e.g. Galor and Michalopoulos, 2012; Ashraf and Galor, 2013). We view the selection and investment in preference approaches to endogenous preference formation as complementary, because they operate on different time horizons.[17]

There is direct evidence that non-cognitive skills are influenced by social factors and family upbringing at a shorter time horizon. Heckman (2000) and Carneiro and Heckman (2003) review the evidence from a large number of programs targeting disadvantaged children. They show that most programs were successful in permanently raising the treated children's non-cognitive skills. These children were more motivated to learn, less likely to engage in crime, and altogether more future-oriented than children of non-treated families. Similar conclusions are reached by studies in child development psychology such as Shonkoff and Philips (2000) and Taylor et al. (2000).

Some studies focus explicitly on preference parameters of economic models. For example, Knowles and Postlewaite (2004) provide evidence of cultural transmission of patience. Using the PSID, they find that parental savings behavior is highly correlated with the education and savings choices of their children's households, after controlling for standard individual characteristics. Moreover, the correlation is stronger between mothers and children than between fathers and children. Since mothers tend to be more actively involved than fathers in the child-rearing process, this observation suggests that there is

[15] These results are consistent with other studies on developing countries.

[16] See, e.g. Cavalli-Sforza and Feldman (1981), Bowles and Gintis (2002), and Richerson and Boyd (2005).

[17] Earlier articles emphasizing the evolutionary selection of preferences include Galor and Moav (2002) and Clark and Hamilton (2006). A recent paper by Baudin (2010) incorporates the interaction of evolutionary forces and cultural transmission in a Beckerian model of endogenous fertility. The interplay between cultural diversity and economic growth is analyzed in Ashraf and Galor (2012).

cultural transmission in patience and propensities to save. In the same vein, Dohmen et al. (2012) document that trust and risk attitudes are strongly correlated between parents and children in the German Socio-Economic Panel. Using the same data set, Zumbuehl et al. (2013) find that parents who invest more in child-rearing efforts are more similar to their children in terms of attitudes toward risk. All these studies concur on the importance of the transmission of non-cognitive skills within families.

1.6.2 Investments in Patience and the Spirit of Capitalism

Doepke and Zilibotti (2008) are closely related to the model discussed in this chapter. The authors propose a dynamic dynastic model rooted in the Beckerian tradition where parents invest in their children's patience and work ethic (modeled as the inverse of the marginal utility of leisure).[18] Preferences are treated as a human-capital-like state variable: parents take their own preferences as given, but can invest in those of their children. The focus of the theory is on the interaction of this accumulation process with the choice of an occupation and savings.

The authors show that the endogenous accumulation of "patience capital" can lead to the stratification of a society into social classes, characterized by different preferences and occupational choices. This occurs even if all individuals initially are identical. In the presence of such endogenous differences in preferences, episodes of technological change can trigger drastic changes in the income distribution, including the leapfrogging of a lower class over the existing elite. The theory is applied to the changes in the distribution of income and wealth that occurred during and after the Industrial Revolution in Britain. Before the onset of industrialization, wealth and political power were associated with the possession of land. Over the 19th century, a new class of entrepreneurs and businessmen and women emerged as the economic elite, replacing the landed elite.

From a theoretical standpoint, the focal point of Doepke and Zilibotti (2008) is an association between occupations and consumption profiles, similar to the model presented in this chapter. In some professions, lifetime earnings are relatively flat, while in others, in particular those requiring the acquisition of skills, high returns are achieved only late in life. These differences affect the incentive of altruistic parents for investing in their children's patience capital: the steeper the consumption profile faced by their children, the stronger the incentive for parents to teach them to be patient. The converse is also true: patient agents have a higher propensity to choose professions entailing steep earnings and consumption profiles.

In the historical application they consider, the pre-industrial middle class had accumulated patience capital, and consequently was better prepared to exploit the new economic opportunities than was the existing elite. The differences in patience, in turn, had their roots in the nature of pre-industrial professions. For centuries, artisans, craftsmen, and

[18] Doepke and Zilibotti (2005) developed a simplified model that focuses only on patience.

merchants were used to sacrificing consumption and leisure in their youth to acquire skills. Consequently, middle-class parents had the strongest incentive to instill into their children a patience and work ethic, that is, a "spirit of capitalism" in Weberian terms. In contrast, the landed elite had accumulated little patience, but a strong appreciation for leisure. The preference profile of the elite arises because the traditional aristocratic sources of income were mostly rents, which neither grew steeply over time, nor hinged on labor effort.

Doepke and Zilibotti (2008) differ from the model presented in this chapter insofar as it abstracts from innovation. In that model, cultural differences that were formed in pre-industrial times explain why different classes responded differently to the new technological opportunities arising at the outset of the Industrial Revolution. However, technology is exogenous, whereas in this chapter cultural transmission is linked explicitly to a theory of endogenous technical change.[19] The theory discussed in this chapter rationalizes why some individuals become entrepreneurs and innovators, and how this affects the speed of technical change and long-run growth.[20]

An implication shared by both Doepke and Zilibotti (2008) and the model presented in this chapter is that cultural transmission makes dynasties facing steeper income profiles more patient. This prediction is consistent with the evidence from a field experiment conducted on Danish households by Harrison et al. (2002). Using monetary rewards, they show that highly educated adults have time discount rates (which are inversely related to the discount factor) as low as two-thirds of those of less educated agents. Since spending time on education typically steepens people's income profile, this finding is in line with the prediction of the theory. A positive correlation between steep income profiles and patience has also been documented at the macro level (see Carroll and Summers, 1991; Becker and Mulligan, 1997). The former documents that in both Japan and the United States consumption-age profiles are steeper when economic growth is high. The latter paper shows that consumption grows faster for richer families and adult consumption grows faster for children of the rich.

1.6.3 Religious Beliefs and Human Capital

Another set of papers studies culture as a system of beliefs affecting people's choices, and ultimately economic development. Significant attention has been paid to religion. Barro and McCleary (2003) show that economic growth is higher in countries with a more widespread belief in hell and heaven. Guiso et al. (2003) come to similar conclusions. Cavalcanti et al. (2007) develop a theoretical model with the possibility of beliefs in

[19] In addition, the model discussed here considers the cultural transmission of risk aversion as well as the possibility of paternalism. Neither feature is covered in Doepke and Zilibotti (2008). Conversely, in that paper we consider the interaction between patience and work ethic, a dimension from which we abstract here.

[20] In this regard, our analysis is related to Klasing (2012) and Klasing and Milionis (2013). However, these papers use a different growth model (related to Acemoglu et al. 2006) and a different cultural transmission mechanism (related to Bisin and Verdier, 2001).

rewards in afterlife. They argue that the model can quantitatively explain cross-country differences in the takeoff from pre-industrial stagnation to growth.

Some influential recent studies point to a close connection between the transmission of religious beliefs and human capital investment. In particular, Botticini and Eckstein (2005, 2006, 2007) examine the cultural roots of the economic success of the Jewish population through a theory of specialization in trade-related activities. They conclude that the key factor was not the system of beliefs of the Jewish religion per se. Rather, it is the extent to which religious beliefs led to human capital accumulation. They document that a religious reform introduced in the second century B.C. caused an increase in literacy rates among Jewish farmers, which, in turn, led to increasing specialization in occupations with a high return to literacy, such as artisanship, trade, and finance. High literacy also led to increased migration into towns, where occupations that reward literacy are concentrated. In a similar vein, Becker and Woessmann (2009) documented that in 19th century Prussia, Protestant counties were more prosperous than Catholic ones, but the effect was entirely due to differences in literacy and education. They conclude that the main channel of the effect of religion on economic performance is human capital.[21]

In the literature discussed so far, religious beliefs are exogenous. In contrast, in Fernández-Villaverde et al. (2010) social norms and beliefs mediated by religious institutions are instead endogenous. They construct a theory where altruistic parents socialize children about sex, instilling a stigma against pre-marital sex in order to reduce the risk of out-of-wedlock births. Religious beliefs and institutions operate as enforcement mechanisms. Similar to Doepke and Zilibotti (2008), cultural transmission responds to changes in the underlying environment. In particular, when modern contraceptives reduce the risk associated with pre-marital sex, they reduce the need for altruistic parents and religious authorities to inculcate sexual mores. The equilibrium effect of technology on culture yields the surprising implication that the number of out-of-wedlock births initially grows significantly in response to new contraceptive technology, due to the higher cultural tolerance for pre-marital sex.

While Doepke and Zilibotti (2008) and Fernández-Villaverde et al. (2010) emphasize the process of cultural transmission, Fernández (2013) and Fogli and Veldkamp (2011) describe culture as a process of Bayesian learning from public and private signals. Those

[21] The finding that the main channel through which Protestantism led to higher economic prosperity was higher literacy and human capital is interpreted by Becker and Woessmann (2009) as evidence against Max Weber's hypothesis that Protestant work ethic had a causal effect of economic success. The distinction is, to some extent, semantic. Their findings are consistent with the broader interpretation of Weber provided by Doepke and Zilibotti (2008) who abstract from religion, but argue that the cultural transmission of patience induces the middle class to undertake human capital investments. In this perspective, one can interpret religious beliefs (e.g. Protestantism) as a complementary driver of patience and work ethic. To the extent to which patience is a constituent of the spirit of capitalism, the evidence of Becker and Woessmann (2009) would be actually consistent with a broad interpretation of Max Weber's theory.

papers explain the sharp increase in female labor supply during the 20th century.[22] Doepke and Tertilt (2009) focus on an earlier period and provide a theory of the expansion of women's rights in the 19th century. The authors argue that rising demand for human capital changed cultural attitudes regarding the proper role of women in society, and ultimately triggered political reform.[23]

1.6.4 Beliefs and Social Norms

Many recent studies link culture and beliefs with the process of development through the effects these have on institutions. For instance, Aghion et al. (2010) and Aghion et al. (2011) argue that trust determines the demand for regulation, especially in labor markets.[24] Heterogeneous beliefs about the effect of redistributive policies are the focus of Piketty (1995). A number of papers also consider the feedback effect from institutions to culture. For instance, Hassler et al. (2005) argue that a generous unemployment benefits system induces low geographic mobility of workers in response to labor market shocks. Low mobility, in turn, increases over time the attachment of workers to their location (modeled as a preference trait), sustaining a high demand of social insurance. A similar argument is developed by Michau (2013), who incorporates his theory in a model of cultural transmission. Lindbeck and Nyberg (2006) argue that public transfers weaken parents' incentives to instill a work ethic in their children. The relationship between trust, efficiency, and size of the welfare state is emphasized by Algan et al. (2013).[25]

Culture, trust, and beliefs have also been argued to have first-order effects on institutional stability and on the ability of societies to foster economic cooperation among its citizens. Rohner et al. (2013) construct a theory where persistent civil conflicts are driven by the endogenous dynamics of inter-ethnic trade and inter-ethnic beliefs about the nature and intentions of other ethnic groups. Inter-ethnic trade hinges on reciprocal trust. The theory predicts that civil wars are persistent (as in Acemoglu et al. 2010), and that societies can plunge into a vicious cycle of recurrent conflicts, low trust, and scant

[22] The learning process can be related to the observation of different family models. Fernández et al. (2004) show that the increase in female labor force participation over time was associated with a growing share of men who grew up in families where mothers worked. They test their hypothesis using differences in mobilization rates of men across states during World War II as a source of variation in female labor supply. They show that higher male mobilization rates led to a higher fraction of women working not only for the generation directly affected by the war, but also for the next generation.

[23] Doepke et al. (2012) provide a more extensive discussion of the relationship between cultural and economic explanations for the historical expansion of women's rights.

[24] For a recent survey of the relationship between trust and economic performance, see Algan and Cahuc (2013).

[25] A related argument is provided by the politico-economic theory of Song et al. (2012) arguing that in countries characterized by inefficient public provision voters are more prone to support high public debt. Although debt crowds out future public expenditure, this is a smaller concern to (young) voters in countries whose governments are inefficient.

inter-ethnic trade (a “war trap”) even though there are no fundamental reasons for the lack of cooperation. Long-run outcomes are path dependent: economies with identical fundamentals may end up in either good or bad equilibria depending on the realization of stochastic shocks that cement or undermine cohesion and inter-group cooperation.[26] Rohner et al. (2013) also provide evidence that the onset and incidence of civil wars are affected significantly by a lagged measure of trust from the World Values Survey. There is also evidence of the opposite channel, i.e. exposure to civil conflict affecting preferences and trust. Using data from a field experiment in rural Burundi, Voors et al. (2012) document that exposure to violence encourages risk-taking but reduces patience, hence depressing saving and investments. Rohner et al. (2012) document survey evidence from the civil conflicts in Uganda that war destroys trust, strengthens ethnic identity, and harms future growth in ethnically divided communities.

In the empirical literature, beliefs and social norms are often difficult to disentangle from the effects of the local economic and institutional environment. Studying the behavior of immigrants and expatriates has proven useful to achieve identification. A noteworthy example is Giuliano (2007), which shows that second-generation southern European male immigrants in the United States behave similarly to their counterparts in their country of origin, and live with their parents much longer than young Americans do. Similarly, Fernández and Fogli (2006, 2009) document that the country of origin explains fertility and work behavior of second-generation American women. Fisman and Miguel (2007) finds that diplomats from more corrupted countries tend to incur significantly more parking violations in the United States (diplomats are generally immune, so fines are not enforced). Bruegger et al. (2009) compare unemployment across Swiss communities with different languages (French versus German). The language border separates cultural groups, but not labor markets or political jurisdictions. They find that cultural differences (identified by language differences) can explain differences in unemployment duration of about 20%.

A number of papers have emphasized the persistence of cultural factors. Culture may respond to changes in the institutional environment, but cultural shifts may take time. This is consistent with the view that adults’ preferences are by and large fixed, as opposed to those of children, whose beliefs, non-cognitive skills, and preferences can be shaped by cultural transmission and the surrounding environment. Even with these influences, cultural changes can take several generations to reach a new steady state after institutions have changed. Alesina and Fuchs-Schuendeln (2007) focus on the fall of the Berlin Wall. After the end of communism, East Germans became subject to the same institutions as West Germans, but carried with them the cultural heritage of the communist experience. Their study documents that several years after unification, East Germans (compared to

[26] In a related paper, Acemoglu and Wolitzky (2012) propose a theory where mistaken signals can trigger belief-driven conflict between two groups.

West Germans) are more supportive of redistribution and believe that social conditions are a more important determinant of individual success. Voigtlaender and Voth (2012) go much further and document evidence that a particular form of cultural trait, namely anti-Semitism in German local communities, has persisted for more than 600 years.[27]

Finally, exogenous sources of variation for culture can be found in historical data. Using data for European regions,Tabellini (2010) finds evidence that culture has a significant causal effect on economic development. The identification relies on two historical variables, the literacy rate and past political institutions.

1.7. OUTLOOK AND CONCLUSIONS

Explaining the vast variation in rates of economic growth and living standards around the world remains one of the main challenges in economics. Growth-theoretic explanations for these observations have focused on variation in factor endowments, technology, or institutions as explanatory variables, while abstracting from the potential role of differences in culture, values, and preferences. In contrast, in this chapter we have developed a theory in which culture (modeled as endogenous preferences) and economic growth are endogenous and affect each other. Economic growth feeds back into the preference formation and transmission process of families, and conversely the existing distribution of preferences in the population determines the potential for economic growth. The theory predicts that countries can reach different balanced growth paths, in which some countries grow fast and others more slowly. Fast-growing countries are the ones with larger shares of the population exhibiting a "spirit of capitalism" (i.e. preferences conducive to innovative activities). Institutions, the development of financial markets, and government policies affecting risk sharing all feed back into preferences and culture, giving rise to long-term changes in economic development that can long outlast the underlying institutions and policies.

In the past, economists generally have shied away from explaining economic phenomena with variation in culture or preferences. A common concern is that such explanations put little discipline on the data. However, this criticism does not apply to explicit models of intergenerational preference transmission that generate specific testable implications, which is the route that we have taken here. In this sense, this chapter is in the spirit of Stigler and Becker (1977), who also analyzed phenomena that at first sight suggest an important role for variation in preferences (such as addiction; customs and tradition; and fashion and advertising).

Of course, for testable implications to be meaningful, researchers need data allowing them to evaluate the restrictions imposed by the theory in practice. From this perspective, an important change in recent years is the increased availability of data sets that permit

[27] They document that cities where Jews were victims of medieval pogroms during the plague era were also very likely to experience anti-Semitic violence in the 20th century, before and during the Nazi rule.

empirical analyses of the transmission of preference traits from parents to children as well as the mutual interaction between cultural preferences and the economic environment (we review a number of such studies in Section 1.6). We expect that combining these new empirical insights with theoretical analyses of the interaction of culture, entrepreneurship, and growth of the kind developed in this chapter will, over time, greatly enhance our understanding of the development process.

A PROOFS OF PROPOSITIONS AND LEMMAS

Proof of Proposition 1. Given Equation (1.14), the zero growth ($\lambda = 0$) steady state exists, if and only if:

$$\chi\left(1-(\psi)^{1-\sigma}\right) \geq \beta\left(\left(2(1-\alpha)\,\alpha^{\frac{\alpha}{1-\alpha}}\,\xi\right)^{1-\sigma}-1\right).$$

Conversely, the balanced growth path features $\lambda = 1$, if and only if:

$$\chi\left(1-(\psi)^{1-\sigma}\right) \leq \beta(1+\xi)^{1-\sigma}\left(\left(\frac{(1-\alpha)\,\alpha^{\frac{\alpha}{1-\alpha}}\,\xi\,\psi}{1+\alpha^{\frac{1}{1-\alpha}}\,\xi}\right)^{1-\sigma}-1\right).$$

An interior balanced growth path with positive fractions of workers and entrepreneurs exists if (1.14) is satisfied as an equality for some λ with $0 < \lambda < 1$. A steady state has to exist (either corner or interior) because (1.14) is continuous in λ. The first inequality in Assumption 1 guarantees that the right-hand side of (1.14) is positive for $\lambda = 0$. The second inequality guarantees that the right-hand side of (1.14) reaches zero for a $\tilde{\lambda}$ with $0 < \tilde{\lambda} < 1$. This also implies that the right-hand side of (1.14) is strictly decreasing in λ for $\lambda \leq \tilde{\lambda}$ sufficiently close to $\tilde{\lambda}$. Let $\hat{\lambda}$ denote the lower bound of the monotonic region. The right-hand side of (1.14) is bounded strictly away from zero for $0 \leq \lambda \leq \hat{\lambda}$. By choosing χ sufficiently small, we can guarantee that (1.14) is not satisfied for a λ in this region. This implies that (1.14) is satisfied for a λ that lies in this monotonic region, which then has to be unique, resulting in a unique, interior balanced growth path. □

Proof of Proposition 2. The system of Bellman equations (1.16)–(1.18) defines a mapping T on the space of bounded continuous functions on the interval $[0, \beta_{\max}]$, endowed with the sup norm, where the mapping is given by:

$$\begin{aligned} Tv(\beta) = \max_{I\in\{0,1\},0\leq l\leq 1} & \left\{(1-I)\left[\chi(l)+\beta\,(1+g)^{1-\sigma}\right]\right. \\ & \left.+I\left[\chi(l)\psi^{1-\sigma}+\beta\,((1+g)\,\eta)^{1-\sigma}\right]+z\,(1+g)^{1-\sigma}\,v(\beta')\right\}, \end{aligned} \qquad (1.34)$$

where the maximization is subject to:

$$\beta' = (1-\delta)\beta + f(l).$$

I is an indicator variable for the occupational choice, and $\beta_{\max} = f(1)/\delta$. Since we imposed assumptions that guarantee $0 < z(1+g)^{1-\sigma} < 1$, this mapping is a contraction by Blackwell's sufficient conditions, and it therefore has a unique fixed point by the Contraction Mapping Theorem. This proves the first part of the proposition.

The proof that the value function is increasing and convex is an application of Corollary 1 to Theorem 3.2 in Stokey and Lucas (1989). Using this result, we can establish the result by establishing that the operator T preserves these properties. To establish that the value function is increasing, let v be a non-decreasing bounded continuous function. We need to show that Tv is a strictly increasing function. To do this, choose $\overline{\beta} > \underline{\beta}$. We now need to establish that $Tv(\overline{\beta}) > Tv(\underline{\beta})$. Since the right-hand side of (1.34) is the maximization of a continuous function over a compact set, the maximum is attained. Let $\underline{l}$ and $\underline{I}$ be choices attaining the maximum for $\underline{B}$. We then have:

$$\begin{aligned} Tv(\overline{\beta}) \geq {} & (1-\underline{I})\left[\chi(\underline{l}) + \overline{\beta}(1+g)^{1-\sigma}\right] \\ & + \underline{I}\left[\chi(\underline{l})\psi^{1-\sigma} + \overline{\beta}((1+g)\eta)^{1-\sigma}\right] + z(1+g)^{1-\sigma}\, v((1-\delta)\overline{\beta} \\ & + f(\underline{l})) > (1-\underline{I})\Big[\chi(\underline{l}) + \underline{\beta}(1+g)^{1-\sigma}\Big] \\ & + \underline{I}\Big[\chi(\underline{l})\psi^{1-\sigma} \\ & + \underline{\beta}((1+g)\eta)^{1-\sigma}\Big] + z(1+g)^{1-\sigma}\, v((1-\delta)\underline{\beta} + f(\underline{l})) = Tv(\underline{\beta}), \end{aligned}$$

which is the desired result. Here the weak inequality follows because the choices $\underline{l}, \underline{I}$ may not be maximizing at $\overline{\beta}$, and the strict inequality follows because v is assumed to be increasing, and we have that $\overline{\beta} > \underline{\beta}$ and $\eta > 0$.

To establish convexity of the value function, let v be a (weakly) convex bounded continuous function. We need to establish that Tv is also a convex function. To show this, choose a number θ such that $0 < \theta < 1$, let $\overline{\beta} > \underline{\beta}$, and let $\beta = \theta\overline{\beta} + (1-\theta)\underline{\beta}$. We now need to show that $\theta Tv(\overline{\beta}) + (1-\theta)Tv(\underline{\beta}) \geq Tv(\beta)$. Let l and I be choices attaining the maximum for β. Since these are feasible, but not necessarily optimal choices at $\overline{\beta}$ and $\underline{\beta}$, we have:

$$\begin{aligned} Tv(\overline{\beta}) \geq {} & (1-I)\left[\chi(l) + \overline{\beta}(1+g)^{1-\sigma}\right] \\ & + I\left[\chi(l)\psi^{1-\sigma} + \overline{\beta}((1+g)\eta)^{1-\sigma}\right] + z(1+g)^{1-\sigma}\, v((1-\delta)\overline{\beta} + f(l)), \\ Tv(\underline{\beta}) \geq {} & (1-I)\Big[\chi(l) + \underline{\beta}(1+g)^{1-\sigma}\Big] \\ & + I\Big[\chi(l)\psi^{1-\sigma} + \underline{\beta}((1+g)\eta)^{1-\sigma}\Big] + z(1+g)^{1-\sigma}\, v((1-\delta)\underline{\beta} + f(l)). \end{aligned}$$

Working toward the desired condition, we therefore have:

$$\begin{aligned}
&\theta\, Tv(\overline{\beta}) + (1-\theta)\, Tv(\underline{\beta}) \\
&\quad \geq (1-I)\left[\chi(l) + \beta\,(1+g)^{1-\sigma}\right] + I\left[\chi(l)\psi^{1-\sigma} + \beta\,((1+g)\,\eta)^{1-\sigma}\right] \\
&\qquad + z\,(1+g)^{1-\sigma}\left[\theta v((1-\delta)\overline{\beta} + f(l)) + (1-\theta)v((1-\delta)\underline{\beta} + f(l))\right] \\
&\quad \geq (1-I)\left[\chi(l) + \beta\,(1+g)^{1-\sigma}\right] + I\left[\chi(l)\psi^{1-\sigma} + \beta\,((1+g)\,\eta)^{1-\sigma}\right] \\
&\qquad + z\,(1+g)^{1-\sigma}\, v((1-\delta)\beta + f(l)) = Tv(\beta),
\end{aligned}$$

which is the required condition. Here, the last inequality follows from the assumed convexity of v. The operator T therefore preserves convexity, and thus the fixed point must also be convex. Notice that linearity is key to this result: the discount factor enters utility linearly, and the parental discount factor has a linear effect on the discount factor of the child.

Regarding the optimal occupational choice, the difference between the utility of being a worker and an entrepreneur for given β and l is given by:

$$\chi(l)\left(1-\psi^{1-\sigma}\right) - \beta\,(1+g)^{1-\sigma}\left(\eta^{1-\sigma}-1\right),$$

where the first term is always positive, and the second term is negative as long as $\eta > 1$. Given that the second term is weighted by β, it follows that being a worker is always optimal for β sufficiently close to zero. Since the utility derived from entrepreneurship relative to being a worker is strictly increasing in β, there is either a cutoff $\bar{\beta}$ such that entrepreneurship is chosen for $\beta \geq \bar{\beta}$, or being a worker is always the preferred choice (when the required cutoff would be larger than $\beta_{\max}$).

As the last step, we would like to show that the optimal investment in patience $l = l(\beta)$ is non-decreasing in β. Fix two discount factors $\underline{\beta} < \overline{\beta}$. Let $\underline{u}_1 = 1$ if at $\underline{\beta}$ the optimal choice is to be a worker, and $\underline{u}_1 = \psi^{1-\sigma}$ otherwise. Similarly, for the second period we define $\underline{u}_2 = (1+g)^{1-\sigma}$ for workers and $\underline{u}_2 = ((1+g)\,\eta)^{1-\sigma}$ for entrepreneurs. $\bar{u}_1$ and $\bar{u}_2$ are defined in the same way. Now let $\underline{l}$ and $\bar{l}$ denote the optimal investments in patience at $\underline{\beta}$ and $\overline{\beta}$. The optimal choice of l the implies the following inequalities:

$$\begin{aligned}
&\chi(\underline{l})\underline{u}_1 + \underline{\beta}\underline{u}_2 + z(1+g)^{1-\sigma} v((1-\delta)\underline{\beta} + f(\underline{l})) \\
&\quad \geq \chi(\bar{l})\underline{u}_1 + \underline{\beta}\underline{u}_2 + z(1+g)^{1-\sigma} v((1-\delta)\underline{\beta} + f(\bar{l})) \\
&\chi(\underline{l})\bar{u}_1 + \overline{\beta}\bar{u}_2 + z(1+g)^{1-\sigma} v((1-\delta)\overline{\beta} + f(\underline{l})) \\
&\quad \leq \chi(\bar{l})\bar{u}_1 + \overline{\beta}\bar{u}_2 + z(1+g)^{1-\sigma} v((1-\delta)\overline{\beta} + f(\bar{l})).
\end{aligned}$$

Subtracting the two inequalities yields:

$$\begin{aligned}
&\chi(\underline{l})\left(\underline{u}_1 - \bar{u}_1\right) + z(1+g)^{1-\sigma}\left(v((1-\delta)\overline{\beta} + f(\bar{l})) - v((1-\delta)\underline{\beta} + f(\bar{l}))\right) \\
&\quad \geq \chi(\bar{l})\left(\underline{u}_1 - \bar{u}_1\right) + z(1+g)^{1-\sigma}\left(v((1-\delta)\overline{\beta} + f(\underline{l})) - v((1-\delta)\underline{\beta} + f(\underline{l}))\right).
\end{aligned}$$

Now there are two possibilities. If the optimal occupational choices at $\underline{\beta}$ and $\overline{\beta}$ are the same, we have $\underline{u}_1 = \overline{u}_1$ and the inequality reads:

$$\begin{aligned} &v((1-\delta)\overline{\beta} + f(\bar{l})) - v((1-\delta)\underline{\beta} + f(\bar{l})) \\ &\quad \geq v((1-\delta)\overline{\beta} + f(\underline{l})) - v((1-\delta)\underline{\beta} + f(\underline{l})). \end{aligned}$$

Since we have already shown that v is convex, this implies $\bar{l} \geq \underline{l}$. The second possibility is that at $\underline{\beta}$ it is optimal to be a worker, and at $\overline{\beta}$ it is optimal to be an entrepreneur, so that we have $\underline{u}_1 - \overline{u}_1 > 0$. Rearranging the expression gives:

$$\begin{aligned} \left(\chi(\underline{l}) - \chi(\bar{l})\right)\left(\underline{u}_1 - \bar{u}_1\right) \geq z(1+g)^{1-\sigma} \Big[& v((1-\delta)\bar{\beta} + f(\underline{l})) - v((1-\delta)\underline{\beta} + f(\underline{l})) \\ & - \left(v((1-\delta)\bar{\beta} + f(\bar{l})) - v((1-\delta)\underline{\beta} + f(\bar{l})) \right) \Big]. \end{aligned}$$

Due to the convexity of v, if we have $\underline{l} > \bar{l}$, the left-hand side would be negative and the right-hand side positive; we therefore must have $\underline{l} \leq \bar{l}$, which completes the proof. □

Proof of Proposition 3. In Proposition 2, we can subdivide the state space $[0, \beta_{\max}]$ into (at most) two closed intervals (they are closed because of our continuity assumptions), where each interval corresponds to the choice of a given occupation (worker or entrepreneur). The agent is just indifferent between the occupations at the boundary between the intervals, and strictly prefers a given occupation in the interior of an interval. The intervals can be further subdivided according to the occupational choice of the child. Since $l(\beta)$ may not be single-valued, there may be multiple optimal β' corresponding to a given β today. Nevertheless, since the β' are strictly increasing in β (because of Proposition 3 and $\delta < 1$) and given that there are only two occupations, we can once again subdivide today's state space into at most two closed intervals, each one corresponding to a specific occupational choice of the child. Continuing this way, the state space $[0, \beta_{\max}]$ can be divided into a countable number of closed intervals (there are two possible occupations in each of the countably many future generations), where each interval corresponds to a specific occupational choice of each generation. Let $[\underline{\beta}, \overline{\beta}]$ be such an interval. We want to establish that the value function is linear over this interval, and that the optimal choice of patience $l(\beta)$ is single-valued and constant over the interior of this interval.

It is useful to consider the sequential formulation of the decision problem. Taking the present and future occupational choices as given and writing the resulting first and second period utilities net of cost of investing in patience as $u_{1,t}$ and $u_{2,t}$, we can substitute for β_t and write the remaining decision problem over the l_t on the interval $[\underline{\beta}, \overline{\beta}]$ as:

$$v(\beta) = \max \left\{ \chi(l_0) u_{1,0} + \beta u_{2,0} \right. \\ \left. + \sum_{t=1}^{\infty} z^t \left[\chi(l_t) u_{1,t} + \left((1-\delta)^t \beta + \sum_{s=0}^{t-1} (1-\delta)^{t-s-1} f(l_s) \right) u_{2,t} \right] \right\}. \quad (1.35)$$

For given current and future occupations, (1.35) is strictly concave in l_t for all t, since χ is concave and f is strictly concave. Moreover, the discount factor β and all expressions involving l_t appear in separate terms in the sum. Therefore, it follows that, given the optimal income profiles, for all t the optimal l_t is unique, and independent of β. Since on the interior of $[\underline{\beta}, \overline{\beta}]$, the current and future optimal occupations are unique, the optimal policy correspondence $l(\beta)$ is single-valued. By construction of the intervals, at the boundary between the two intervals both occupations are optimal choices for at least one generation, hence $l(\beta)$ may take on more than one optimal value, one corresponding to each optimal set of income profiles.

The optimal value function v over the interval $[\underline{\beta}, \overline{\beta}]$ is given by (1.35) with occupations and investment in patience l_t fixed at their optimal (and constant) values. Equation (1.35) is linear in β; it therefore follows that the value function is piecewise linear, with each kink corresponding to the boundary between two of the intervals. □

Proof of Proposition 4. Since f is an increasing function and we assume that $\delta < 1$, the law of motion is strictly increasing in β. Notice that $l(\beta)$ may not be single-valued for all β. Strictly increasing here means that $\overline{\beta} < \underline{\beta}$ implies $\overline{\beta}' < \underline{\beta}'$ for all optimal $\overline{\beta}' \in g(\overline{\beta})$ and $\underline{\beta}' \in g(\underline{\beta})$, even if $g(\overline{\beta})$ or $g(\underline{\beta})$ is a set. For a given β_0, the law of motion g defines (potentially multiple) optimal sequences of discount factors $\{\beta_t\}_{t=0}^{\infty}$. Any such sequence is a monotone sequence on the compact set $[0, \beta_{\max}]$, and must therefore converge. Notice, however, that since $l(\beta)$ is not single-valued everywhere, different steady states can be reached even from the same initial β_0. □

Proof of Lemma 1. Assume that (1.23) holds with equality:

$$v^E = \chi(l^{EW})\psi^{1-\sigma} + \beta^E ((1+g)\eta)^{1-\sigma} + z(1+g)^{1-\sigma} \left(\chi(l^W) + \beta^{EW} (1+g)^{1-\sigma} \right) \\ + z^2 (1+g)^{2(1-\sigma)} v^W. \quad (1.36)$$

Now replacing l^{EW} and β^{EW} on the right-hand side with l^E and β^E lowers utility, because these are not the optimal choices given the chosen occupations. We therefore have:

$$\chi(l^E)\psi^{1-\sigma} + \beta^E ((1+g)\eta)^{1-\sigma} + z(1+g)^{1-\sigma} v^E \\ > \chi(l^E)\psi^{1-\sigma} + \beta^E ((1+g)\eta)^{1-\sigma} + z(1+g)^{1-\sigma} \left(\chi(l^W) + \beta^E (1+g)^{1-\sigma} \right) \\ + z^2 (1+g)^{2(1-\sigma)} v^W,$$

where we also rewrote the left-hand side to explicitly show the first-generation utility. Now subtracting the (identical) first-generation terms on both sides and dividing by $z(1+g)^{1-\sigma}$ we get:

$$v^E > \left(\chi(l^W) + \beta^E (1+g)^{1-\sigma}\right) + z(1+g)^{1-\sigma} v^W,$$

which is (1.22) as a strict inequality.

Moving on, replacing the β^E of the initial generation on both sides of (1.36) with β^W leaves the equality intact, because the discount factor enters both sides in the same way:

$$\begin{aligned}&\chi(l^E)\psi^{1-\sigma} + \beta^W ((1+g)\,\eta)^{1-\sigma} + z(1+g)^{1-\sigma} v^E \\ &\quad = \chi(l^{EW})\psi^{1-\sigma} + \beta^W ((1+g)\,\eta)^{1-\sigma} + z(1+g)^{1-\sigma}\left(\chi(l^W) + \beta^{EW}(1+g)^{1-\sigma}\right) \\ &\qquad + z^2(1+g)^{2(1-\sigma)} v^W. \end{aligned} \tag{1.37}$$

Now switching the first-generation occupational choice from entrepreneurship to work yields the following strict inequality:

$$\begin{aligned}&\chi(l^{WE}) + \beta^W (1+g)^{1-\sigma} + z(1+g)^{1-\sigma}\left(\chi(l^E)\psi^{1-\sigma} + \beta^{WE}((1+g)\,\eta)^{1-\sigma}\right) \\ &\quad + z^2(1+g)^{2(1-\sigma)} v^E < v^W.\end{aligned}$$

The strict inequality arises because $l^{EW} < l^E$, implying that the increase in the first-period utility from being a worker is larger on the right-hand side. This still applies after investment in patience is reoptimized (to l^{WE} on the left-hand side and l^W on the right-hand side) due to the envelope theorem. The resulting inequality is a strict version of (1.25).

Finally, again starting with (1.37), replacing the initial investment in patience with l^{EW} (and plugging in the corresponding discount factor in the next generation) lowers utility on the left-hand side, so that we have:

$$\begin{aligned}&\chi(l^{EW})\psi^{1-\sigma} + \beta^W ((1+g)\,\eta)^{1-\sigma} + z(1+g)^{1-\sigma}\left(\chi(l^E)\psi^{1-\sigma} + \beta^{EW}((1+g)\,\eta)^{1-\sigma}\right) \\ &\quad + z^2(1+g)^{2(1-\sigma)} v^E \\ &< \chi(l^{EW})\psi^{1-\sigma} + \beta^W ((1+g)\,\eta)^{1-\sigma} + z(1+g)^{1-\sigma}\left(\chi(l^W) + \beta^{EW}(1+g)^{1-\sigma}\right) \\ &\quad + z^2(1+g)^{2(1-\sigma)} v^W.\end{aligned}$$

Subtracting the identical first-generation terms and dividing by $z(1+g)^{1-\sigma}$ yields:

$$\begin{aligned}&\chi(l^E)\psi^{1-\sigma} + \beta^{EW}((1+g)\eta)^{1-\sigma} + z(1+g)^{1-\sigma} v^E \\ &\quad < \chi(l^W) + \beta^{EW}(1+g)^{1-\sigma} + z(1+g)^{1-\sigma} v^W.\end{aligned}$$

Now changing the initial discount factor from β^{EW} to $\beta^{W} < \beta^{EW}$ lowers the left-hand side yet again more than the right-hand side (because $\eta > 1$), so that the inequality stays intact:

$$\chi(l^E)\psi^{1-\sigma} + \beta^W((1+g)\eta)^{1-\sigma} + z(1+g)^{1-\sigma}\nu^E < \nu^W,$$

which is a strict version of (1.24). □

Proof of Proposition 5. The fraction of entrepreneurs λ in the balanced growth path can be mapped into an entrepreneurial premium η and a growth rate g given the analysis in Section 1.2.3 above. The entrepreneurial premium is continuous in λ. Hence, if there exists a fraction of entrepreneurs λ that satisfies $0 < \lambda < 1$ and such that conditions (1.22)–(1.25) hold as strict inequalities, there has to be a range of λ and associated η and g such that the conditions continue to hold. If at the initial λ condition (1.23) holds with equality, then given Lemma 1 we know that the remaining constraints hold as strict inequalities. Given continuity it is then possible to raise η (by changing λ) within some range and have all conditions hold as strict inequalities, implying that a continuum of balanced growth paths exists. The same argument can be applied reversely to the point where (1.25) holds as an equality. The highest entrepreneurial return that is consistent with balanced growth is characterized by (1.25) holding as an equality. □

Proof of Proposition 6. Since the financial market allows for an arbitrary allocation of consumption across the two periods, an occupation that is dominated in terms of the present value of income is also dominated in terms of consumption, and therefore is never chosen. Hence, the set of optimal occupations is independent of patience β, because the present value of income in the two occupations does not depend on β. When both occupations yield the same present value of income, they also lead to the same consumption profile. The cost of investing in patience depends only on first-period consumption, which therefore does not depend on the chosen occupation. Likewise, the return to investing in patience is independent of the occupation of the current generation. Investment in patience therefore does not depend on which occupation is chosen. □

ACKNOWLEDGMENTS

Financial support from the National Science Foundation (grant SES-0820409, Doepke) and from the European Research Council (ERC Advanced Grant IPCDP-229883, Zilibotti) is gratefully acknowledged. Veronika Selezneva provided excellent research assistance. We thank Mariko Klasing for helpful comments.

REFERENCES

Acemoglu, Daron, Aghion, Philippe, Zilibotti, Fabrizio, 2006. Distance to frontier, selection and economic growth. Journal of the European Economic Association 4 (1), 37–74.

Acemoglu, Daron, Ticchi, Davide, Vindigni, Andrea, 2010. Persistence of civil wars. Journal of the European Economic Association 8 (4), 664–676.

Acemoglu, Daron, Wolitzky, Alexander, 2012. Cycles of Distrust: An Economic Model. NBER Working Paper 18257.

Aghion, Philippe, Howitt, Peter, 1992. A model of growth through creative destruction. Econometrica 60 (2), 323–351.

Aghion, Phillipe, Algan, Yann, Cahuc, Pierre, 2011. Civil society and the state: the interplay between cooperation and minimum wage regulation. Journal of the European Economic Association 9 (1), 3–42.

Aghion, Phillipe, Algan, Yann, Cahuc, Pierre, Shleifer, Andrei, 2010. Regulation and distrust. Quarterly Journal of Economics 125 (3), 1015–1049.

Alesina, Alberto, Fuchs-Schuendeln, Nicola, 2007. Good bye Lenin (or not?): the effect of communism on people's preferences. American Economic Review 97 (4), 1507–1528.

Alesina, Alberto, Giuliano, Paola, 2009. Preferences for Redistribution. NBER Working Paper 14825.

Alesina, Alberto, Glaeser, Edward, 2004. Fighting Poverty in the U.S. and Europe. Oxford University Press, Oxford, UK.

Algan, Yann, Cahuc, Pierre, 2013. Trust and growth. Annual Review of Economics 5 (1), 521–549.

Algan, Yann, Cahuc, Pierre, Sangnier, Marc, 2013. Efficient and Inefficient Welfare States. Unpublished Manuscript, Sciences Po, Paris.

Ashraf, Quamrul, Galor, Oded. 2012. Cultural Diversity, Geographical Isolation, and the Origin of the Wealth of Nations. IZA Discussion Paper 6319.

Ashraf, Quamrul, Galor, Oded, 2013. The "Out-of-Africa" hypothesis, human genetic diversity, and comparative economic development. American Economic Review 103 (1), 1–46.

Barro, Robert J., McCleary, Rachel M., 2003. Religion and economic growth across countries. American Sociological Review 68 (5), 760–781.

Baudin, Thomas, 2010. A role for cultural transmission in fertility transitions. Macroeconomic Dynamics 14 (4), 454–481.

Beauchamp, Jonathan, Cesarini, David, Johannesson, Magnus, 2011. The Psychometric Properties of Measures of Economic Risk Preferences. Unpublished Manuscript, Harvard University.

Becker, Gary S., Mulligan, Casey B., 1997. The endogenous determination of time preference. Quarterly Journal of Economics 112 (3), 729–758.

Becker, Sascha, Woessmann, Ludger, 2009. Was Weber wrong? A human capital theory of protestant economic history. Quarterly Journal of Economics 124 (2), 531–596.

Bisin, Alberto, Verdier, Thierry, 2001. The economics of cultural transmission and the dynamics of preferences. Journal of Economic Theory 97 (2), 298–319.

Botticini, Maristella, Eckstein, Zvi, 2005. Jewish occupational selection: education, restrictions, or minorities? Journal of Economic History 65 (4), 922–948.

Botticini, Maristella, Eckstein, Zvi, 2006. Path dependence and occupations. In: Durlauf, Steven N., Blume, Lawrence (Eds.), New Palgrave Dictionary of Economics, Palgrave Macmillan, New York.

Botticini, Maristella, Eckstein, Zvi, 2007. From farmers to merchants, voluntary conversions and diaspora: a human capital interpretation of jewish history. Journal of the European Economic Association 5 (5), 885–926.

Bowles, Samuel, Gintis, Howard, 2002. The inheritence of inequality. Journal of Economic Perspectives 16 (3), 3–30.

Bruegger, Beatrice, Lalive, Rafael, Zweimueller, Josef, 2009. Does Culture Affect Unemployment? Evidence from the Roestigraben. CESifo Working Paper Series No 2714.

Carneiro, Pedro, Heckman, James J., 2003. Human capital policy. In: Heckman, James J., Krueger, Alan B. (Eds.), Inequality in America: What Role for Human Capital Policies. MIT Press, Cambridge, pp. 77–240.

Carroll, Christopher D., Summers, Lawrence H., 1991. Consumption growth parallels income growth: some new evidence. In: Bernheim, B. Douglas, Shaven, John B. (Eds.), National Saving and Economic Performance. University of Chicago Press, Chicago, pp. 305–343.

Cavalcanti, Tiago V., Parente, Stephen L., Zhao, Rui, 2007. Religion in macroeconomics: a quantitative analysis of Weber's thesis. Economic Theory 32 (1), 105–123.

Cavalli-Sforza, Luigi Luca, Feldman, Marcus W., 1981. Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press.

Clark, Gregory, Hamilton, Gillian, 2006. Survival of the richest: the malthusian mechanism in pre-industrial England. Journal of Economic History 66 (3), 707–736.

Coen-Pirani, Daniele, 2004. Effects of differences in risk aversion on the distribution of wealth. Macroeconomic Dynamics 8 (5), 617–632.

Cozzi, Marco, 2011. Risk Aversion Heterogeneity, Risky Jobs and Wealth Inequality. Unpublished Manuscript, Queen's University.

Cramer, J.S., Hartog, Joop, Jonker, Nicole, van Praag, C.M., 2002. Low risk aversion encourages the choice for entrepreneurship: an empirical test of a truism. Journal of Economic Behavior and Organization 48 (1), 29–36.

De Nardi, Mariacristina, 2004. Wealth inequality and intergenerational links. Review of Economic Studies 71 (3), 743–768.

Doepke, Matthias, Tertilt, Michèle, 2009. Women's liberation: what's in it for men? Quarterly Journal of Economics 124 (4), 1541–1591.

Doepke, Matthias, Tertilt, Michèle, Voena, Alessandra, 2012. The economics and politics of women's rights. Annual Review of Economics 4, 339–372.

Doepke, Matthias, Zilibotti, Fabrizio, 2005. Social class and the spirit of capitalism. Journal of the European Economic Association 3 (2–3), 516–524.

Doepke, Matthias, Zilibotti, Fabrizio, 2008. Occupational choice and the spirit of capitalism. Quarterly Journal of Economics 123 (2), 747–793.

Doepke, Matthias, Zilibotti, Fabrizio, 2012. Parenting with Style: Altruism and Paternalism in Intergenerational Preference Transmission. IZA Discussion Paper 7108.

Dohmen, Thomas, Falk, Armin, Huffman, David, Sunde, Uwe, 2012. The intergenerational transmission of risk and trust attitudes. Review of Economic Studies 79 (2), 645–677.

Fehr, Ernst, Hoff, Karla, 2011. Introduction: tastes, castes and culture: the influence of society on preferences. The Economic Journal 121 (556), F396–F412.

Fernández, Raquel, 2013. Cultural change as learning: the evolution of female labor force participation over a century. American Economic Review 103 (1), 472–500.

Fernández, Raquel, Fogli, Alessandra, 2006. Fertility: the role of culture and family experience. Journal of the European Economic Association 4 (2–3), 552–561.

Fernández, Raquel, Fogli, Alessandra, 2009. Culture: an empirical investigation of beliefs, work, and fertility. American Economic Journal: Macroeconomics 1 (1), 147–177.

Fernández, Raquel, Fogli, Alessandra, Olivetti, Claudia, 2004. Mothers and sons: preference formation and female labor force dynamics. Quarterly Journal of Economics 119 (4), 1249–1299.

Fernández-Villaverde, Jesús, Greenwood, Jeremy, Guner, Nezih, 2010. From shame to game in one hundred years: an economic model of the rise in premarital sex and its de-stigmatization. Journal of the European Economic Association 11.

Fisman, Raymond, Miguel, Edward, 2007. Corruption, norms and legal enforcement: evidence from diplomatic parking tickets. Journal of Political Economy 115 (6), 1020–1048.

Fogli, Alessandra, Veldkamp, Laura, 2011. Nature or nurture? Learning and the geography of female labor force participation. Econometrica 79 (4), 1103–1138.

Galor, Oded, Michalopoulos, Stelios, 2012. Evolution and the growth process: natural selection of entrepreneurial traits. Journal of Economic Theory 147 (2), 759–780.

Galor, Oded, Moav, Omer, 2002. Natural selection and the origin of economic growth. Quarterly Journal of Economics 117 (4), 1133–1191.

Giuliano, Paola, 2007. Living arrangements in western europe: does cultural origin matter? Journal of the European Economic Association 5 (5), 927–952.

Gorodnichenko, Yuriy, Roland, Gerard. 2010. Culture, Institutions and the Wealth of Nations. NBER Working Paper 16368.

Gradstein, Mark, 2007. Endogenous Reversals of Fortune. Unpublished Manuscript, Ben Gurion University.

Greif, Avner, 1994. Cultural beliefs and the organization of society: a historical and theoretical reflection on collectivist and individualist societies. Journal of Political Economy 102 (5), 912–950.

Grosjean, Pauline, 2013. A history of violence: the culture of honor and homicide in the US south. Journal of the European Economic Association.

Guiso, Luigi, Paiella, Monica, 2008. Risk aversion, wealth, and background risk. Journal of the European Economic Association 6 (6), 1109–1150.

Guiso, Luigi, Sapienza, Paola, Zingales, Luigi, 2003. People's opium? Religion and economic attitudes. Journal of Monetary Economics 50 (1), 225–282.

Guiso, Luigi, Sapienza, Paola, Zingales, Luigi, 2006. Does culture affect economic outcomes? Journal of Economic Perspectives 20 (2), 23–49.

Guvenen, Fatih, 2006. Reconciling conflicting evidence on the elasticity of intertemporal substitution: a macroeconomic perspective. Journal of Monetary Economics 53 (7), 1451–1472.

Haaparanta, Pertti, Puhakka, Mikko, 2004. Endogenous Time Preference, Investment and Development Traps. BOFIT Discussion Paper No. 4/2004, Bank of Finland.

Harbaugh, William T., Krause, Kate, Vesterlund, Lise, 2002. Risk attitudes of children and adults: choices over small and large probability gains and losses. Experimental Economics 5 (1), 53–84.

Harrison, Glenn W., Lau, Morten I., Williams, Melonie B., 2002. Estimating individual discount rates in Denmark: a field experiment. American Economic Review 92 (5), 1606–1617.

Hassler, John, Rodriguez, Jose V., Mora, Kjetil Storesletten, Zilibotti, Fabrizio, 2005. A positive theory of geographic mobility and social insurance. International Economic Review 46 (1), 263–303.

Hauk, Esther, Saez-Marti, Maria, 2002. On the cultural transmission of corruption. Journal of Economic Theory 107 (2), 311–335.

Heckman, James J., 2000. Policies to foster human capital. Research in Economics 54 (1), 3–56.

Heckman, James J., Stixrud, Jora, Urzua, Sergio, 2006. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. Journal of Labor Economics 24 (3), 411–482.

Hendricks, Lutz, 2007. How important is discount rate heterogeneity for wealth inequality? Journal of Economic Dynamics and Control 31 (9), 3042–3068.

Inglehart, Ronals, Bashkirova, Elena, Basanez, Miguel, Chiu, Hei-yuan, Diez-Nicolas, Juan, Esmer, Yilmaz, Halman, Loek, Klingemann, Hans-Dieter, Nwabuzor, Elone, Petterson, Thorleif, Siemienska, Renata, Yamazaki, Seiko, 2000. World Values Surveys and European Values Surveys, 1981-1984, 1990-1993 and 1995-1997. Inter-university Consortium for Political and Social Research ICPSR 2790.

Kan, Kamhon, Tsai, Wei-Der, 2006. Entrepreneurship and risk aversion. Small Business Economics 26, 465–474.

Kihlstrom, Richard E., Laffont, Jean-Jacques, 1979. A general equilibrium entrepreneurial theory of firm formation based on risk aversion. Journal of Political Economy 87 (4), 719–748.

Klasing, Mariko J., 2012. Cultural Change, Risk-Taking Behavior, and the Course of Economic Development. Unpublished Manuscript, Carleton University.

Klasing, Mariko J., Milionis, Petros, 2013. Cultural constraints on innovation-based growth. Economic Inquiry.

Knight, Frank H., 1921. Risk, Uncertainty, and Profit. Houghton Mifflin, Boston and New York.

Knowles, John A., Postlewaite, Andrew, 2004. Do Children Learn to Save from Their Parents? Unpublished Manuscript, University of Pennsylvania.

Krusell, Per, Smith Jr., Anthony A., 1998. Income and wealth heterogeneity in the macroeconomy. Journal of Political Economy 106 (5), 867–896.

Lindbeck, Assar, Nyberg, Sten, 2006. Raising children to work hard: altruism, work norms, and social insurance. Quarterly Journal of Economics 121 (4), 1473–503.

Matsuyama, Kiminori, 1999. Growing through cycles. Econometrica 67 (2), 335–347.

Michau, Jean-Baptiste, 2013. Unemployment insurance and cultural transmission: theory and application to European unemployment. Journal of the European Economic Association 11 (5).

Mischel, Walter, Yuichi Shoda, Monica L. Rodriguez. 1992. Delay of gratification in children. In: Loewenstein, George, Elster, Jon (Eds.), Chapter 6 of Choice Over Time. Russell Sage Foundation, New York.

Mokyr, Joel, 2011. Cultural Entrepreneurs and Economic Development. Unpublished Manuscript, Ben-Gurion University.

Mulligan, Casey B., 1997. Parental Priorities and Economic Inequality. University of Chicago Press.

Piketty, Thomas, 1995. Social mobility and redistributive politics. Quarterly Journal of Economics 110, 551–584.

Reyes-Garcia, Victoria, Godoy, Ricardo, Huanca, Tomas, Leonard, William R., McDade, Thomas, Tanner, Susan, Vadez, Vencent, 2007. The origin of monetary income inequality: patience, human capital, and the division of labor. Evolution and Human Behavior 28, 37–47.

Richerson, Peter J., Boyd, Robert, 2005. Not by Genes Alone. The University of Chicago Press.

Rohner, Dominic, Thoenig, Mathias, Zilibotti, Fabrizio, 2012. Seeds of Distrust: Conflict in Uganda. CEPR Discussion Paper No. 8741.

Rohner, Dominic, Thoenig, Mathias, Zilibotti, Fabrizio. 2013. War signals: a theory of trade, trust and conflict. Review of Economic Studies 80 (3), 1114–1147.

Romer, Paul M. 1990. Endogenous technological change. Journal of Political Economy 98 (5, Part 2), S71–S102.

Saez-Marti, Maria, Sjoegren, Anna, 2008. Peers and culture. Scandinavian Journal of Economics 110, 73–92.

Saez-Marti, Maria, Zenou, Yves, 2011. Cultural transmission and discrimination. Journal of Urban Economics 72 (2–3), 137–146.

Saez-Marti, Maria, Zilibotti, Fabrizio, 2008. Preferences as human capital: rational choice theories of endogenous preferences and socioeconomic changes. Finnish Economic Papers 21 (2), 81–94.

Schumpeter, Joseph A., 1942. Capitalism, Socialism, and Democracy. Harper, New York.

Segal, Carmit, 2013. Misbehavior, Education, and labor market outcomes. Journal of the European Economic Association 11 (4), 743–779.

Shonkoff, Jack, Philips, Deborah (Eds.), 2000. From Neurons to Neighborhoods: The Science of Early Childhood Development. National Academy Press, Washington, D.C.

Smith, Adam, 1776. An inquiry into the nature and causes of the wealth of nations. In: Cannan, Edwin (Eds.), The University of Chicago Press, Chicago (published 1976).

Song, Zheng, Storesletten, Kjetil, Zilibotti, Fabrizio, 2012. Rotten parents and disciplined children: a politico-economic theory of public expenditure and debt. Econometrica 80 (6), 2785–2803.

Stigler, George J., Becker, Gary S., 1977. De Gustibus Non Est Disputandum. American Economic Review 67 (2), 76–90.

Stokey, Nancy L., Lucas Jr., Robert E., 1989. Recursive Methods in Economic Dynamics. Harvard University Press, Cambridge.

Sutter, Matthias, Kocher, Martin G., Glaetze-Ruetzler, Daniela, Trautmann, Stefan T., 2013. Impatience and uncertainty: experimental decisions predict adolescents field behavior. American Economic Review 103 (1), 510–531.

Tabellini, Guido, 2008. The scope of cooperation: norms and incentives. The Quarterly Journal of Economics 123 (3), 905–950.

Tabellini, Guido, 2010. Culture and institutions: economic development in the regions of Europe. Journal of the European Economic Association 8 (4), 677–716.

Taylor, J., McGue, M., Iacono, W.G., 2000. Sex differences, assortative mating, and cultural transmission effects on adolescent delinquency: a twin family study. Journal of Child Psychology and Psychiatry 41 (4), 433–440.

van Praag, C.M., Cramer, J.S., 2001. The roots of entrepreneurship and labour demand: individual ability and low risk aversion. Economica 68 (269), 45–62.

Vereshchagina, Galina, Hopenhayn, Hugo A., 2009. Risk taking by entrepreneurs. American Economic Review 99 (5), 1808–1830.

Voigtlaender, Nico, Voth, Hans-Joachim, 2012. Persecution perpetuated: the medieval origins of anti-semitic violence in Nazi Germany. Quarterly Journal of Economics 127 (2), 1339–1392.

Voors, Maarten J., Eleonora, E.M., Nillesen, Philip Verwimp, Bulte, Erwin H., Lensink, Robert, Van Soest, Daan P., 2012. Violent conflict and behavior: a field experiment in burundi. American Economic Review 102 (2), 941–964.

Weber, Max, 1905. The Protestant Ethic and the Spirit of Capitalism. (Translated by Talcott Parsons); with a foreword by R. H. Tawney. Charles Scribner's Sons, New York, 1958. Republished by Dover, New York, 2003.

Zumbuehl, Maria, Dohmen, Thomas, Pfann, Gerard, 2013. Parental investment and the intergenerational transmission of economic preferences and attitudes. Unpublished Manuscript, University of Bonn.

CHAPTER TWO

Trust, Growth, and Well-Being: New Evidence and Policy Implications

Yann Algan[*] and Pierre Cahuc[†]
[*]Sciences Po, France
[†]ENSAE-CREST, Ecole Polytechnique, France

Abstract

This survey reviews the recent research on trust, institutions, and economic development. It discusses the various measures of trust and documents the substantial heterogeneity of trust across space and time. The conceptual mechanisms that explain the influence of trust on economic performance and the methods employed to identify the causal impact of trust on economic performance are reviewed. We document the mechanisms of interactions between trust and economic development in the realms of finance, innovation, the organization of firms, the labor market, and the product market. The last part reviews recent progress to identify how institutions and policies can affect trust.

Keywords

Trust, Growth, Economic development, Institutions

JEL Classification Codes

O11, O43, Z13

> *There are countries in Europe ... where the most serious impediment to conducting business concerns on a large scale, is the rarity of persons who are supposed fit to be trusted with the receipt and expenditure of large sums of money.*
>
> (***Mill, 1848, p. 132***)

2.1. INTRODUCTION

The debate about the roots of economic development and the origins of income inequality across the globe has deeply evolved over time. Early researches focused on the proximate factors of growth, stressing the role of technological progress and the accumulation of human and physical capital. A decade ago, the focus shifted to the role of formal institutions, considered as the endogenous incentives to accumulate and innovate (Acemoglu et al. 2001); and to what extent those institutions could be distinguished from factors like human capital (Glaeser et al. 2004). More recently, the attention has been gradually evolving toward deeper factors, ingrained in culture or long-term history.

Handbook of Economic Growth, Volume 2A
ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53538-2.00002-2

This survey reviews some strands of the recent research on the role of cultural values in economic development (see Nunn, 2009; Spolaore and Wacziarg, 2013 for surveys on long-term history). In particular, we investigate the role of one of the most fundamental cultural values that could explain economic development: trust. Since the path breaking work of Banfield (1958), Coleman (1990), and Putnam (2000), trust, broadly defined as cooperative attitude outside the family circle, was considered as a key element of many economic and social outcomes by social scientists. Yet, while praised in other social sciences, the role of trust in the mainstream economic literature has long been disputed.

The potential role of trust in economic development had naturally attracted some interest decades ago, no doubt for the reason stated by Arrow (1972): "virtually every commercial transaction has within itself an element of trust, certainly any transaction conducted over a period of time. It can be plausibly argued that much of the economic backwardness in the world can be explained by the lack of mutual confidence." Arrow's intuition was straightforward. In a complex society, it is impossible to write down and enforce detailed contracts that encompass all the states of nature for economic exchanges. Ultimately, in the absence of informal rules like trusting behavior, markets are missing, gains from economic exchanges are forgone, and resources are misallocated. To that respect, trust and the informal rules shaping cooperation could explain differences in economic development.

But the theoretical and empirical foundations of the relationship between trust and growth have long been considered as weak, at best. A good illustration of the state of the art one decade ago is given by the former issue of the Handbook of Economic Growth in 2005. In the chapter devoted to social capital, Durlauf and Fafchamps (2005) outlined powerfully all the conceptual and statistical flaws raised by the notion of trust in the economic literature. The concept of social capital, a buzzword according to Solow, raised a lot of ambiguity by encompassing vague concepts as norms, networks, or cooperation. Besides, the authors documented forcefully the identification issued raised by the few cross-country or cross-regional correlations between social capital and growth (see also Durlauf for a critical assessment of the empirical literature on social capital, 2002).

In this chapter, we show that decisive and substantial progress has been made on the different dimensions that give trust a central role in mainstream economics, and more importantly, for explaining economic development. This chapter has five main goals. First, we outline a unified conceptual framework for thinking about how trust and cooperation can increase economic efficiency. We distinguish the specific role of trust, relative to reputation incentives, to overcome market failures. Second, we review the various methods to measure trust and cooperation empirically. The recent development of experimental economics, combined with an increasing number of social surveys, has helped to clarify what trust is and how it differs from other beliefs and preferences. Third, we document the empirical relationship between trust, income per capita, and growth. We review the recent advances to identify a causal impact of trust on economic outcomes. Recent empirical work confirms what Arrow posited: trust does indeed appear to constitute a decisive determinant of growth. This observation is buttressed at present by a range of contributions that not only have shed light on the correlations between these two variables, but have also elaborated strategies for detecting the ways in which trust may affect growth. Fourth, we review the burgeoning literature that focuses on the channels of influence of trust: from financial, product, and labor markets to innovation and the organization of firms. Finally, we document more recent research looking at how institutions and trust co-evolve, and how public policy could boost pro-social behaviors.

Several surveys to date have analyzed the role of social capital and trust in economics (see Guiso et al. 2008b, 2011; Tabellini, 2008a; Fehr, 2009; Bowles and Polania-Reyes, 2012, among others). The present addition to the literature is specific in three ways. First, we focus on the relations between trust, growth, and institutions and we utilize the most recent assemblages of data on values, which allow us to cover more than 90% of the world population. Second, we take full account of the progress made during the last decade in identifying the impact of trust, or inherited trust, by deploying as instruments, events of an essentially historical kind. Recent research allows us to pinpoint more closely the mechanisms by which transmission of trust affects the economy, and to distinguish its various channels. Lastly, we present a synthesis of research on how political and economic institutions interact with trust. We also review the various factors and policies that have been found to affect trust, such as the transparency of institutions, the extent of inequality or education, and early childhood intervention.

The remainder of the chapter is organized as follows. The first part outlines the theoretical mechanisms that explain the influence of trust on economic performance. The second part discusses the various measures of trust and documents the international and interregional heterogeneity of trust, using surveys that furnish rich sets of data going back to the start of the 1980s. The third part is a presentation of the dynamics of trust, stressing that in general it evolves slowly from one generation to the next. This inertia, which may nevertheless be perturbed by major historical events such as wars, is observable both at the individual level and at the macro-social level. Part four presents the methods employed to identify the causal impact of trust and provides an empirical illustration of the relation between trust and economic development. Part five describes the mechanisms by which trust has an impact on growth. Part six analyzes the interaction of trust with formal institutions and policies and discusses how trust can be built. And, part seven concludes this chapter by discussing the new perspectives provided by recent research showing that well-being depends not only on income but also, and foremost, on the quality of social relationships.

2.2. THEORETICAL FOUNDATIONS

We begin by providing a conceptual framework that rationalizes the relationship between trust and economic performance. We then document the theoretical channels through which trust interacts with the institutional environment and can emerge as a stable equilibrium.

For trust to have an economic impact and to improve efficiency, one has first to consider the reasons why the economy would depart from the first-best allocation in absence of trust. In his analysis of the limits of organization, Arrow (1972) considers trust as co-substantial to economic exchange in the presence of transaction costs that impede information and contracts. Fundamentally, the economic efficiency of trust flows from the fact that it favors cooperative behavior and thus facilitates mutually advantageous exchanges in presence of incomplete contracts and imperfect information. In Arrow's terms, trust would act as a lubricant to economic exchange in a second-best allocation.

This remark raises various questions. How can we rationalize the impact of trust on economic exchange? How can trust emerge and be sustained in economic exchanges? Why should we expect trust rather than institutions to overcome these market imperfections?

To address those issues, we start from a simple example inspired from the trust game of Berg et al. (1995), where each participant is an investor. We show that cooperation cannot emerge in absence of reputation, which is at odds with the insights of behavioral economics, which documents that individuals do often cooperate with anonymous others in a one-shot exchange. It is thus necessary to include trust as an additional characteristic to rationalize cooperation. We then discuss how trust evolves and is transmitted to become a stable equilibrium. We also document the interaction between trust and institutions to explain economic exchanges.

2.2.1 Cooperation and Reputation

Let us consider two individuals, both of whom are free to invest—or not—an irrecoverable sum $I > 0$ that will enable them to produce jointly. Only by mutual agreement do they invest. Once they do, the incompleteness of contracts, arising out of the complexity of the association which makes it impossible for a third party to verify that everything promised is performed, gives each player the chance to profit from the association at the expense of the other. Hence, each player has the option of investing or not at the outset, and of cooperating or defecting subsequently. Production is positive only if the two individuals invest. If the two players cooperate, their investment yields production amounting to $2(Y + I) > 0$, divided into equal shares such that each obtains a gain, net of the cost of the investment, amounting to $Y > 0$. If neither cooperates, production is zero and the sum each invested is entirely forfeited. Finally, if one cooperates while the other defects, the one who defects preempts the production to his advantage and obtains a net gain of $2Y + I$, while the one who cooperated forfeits his initial investment entirely. The gains are represented in Table 2.1. The Nash equilibrium of this game is an absence of cooperation entailing that the players have no interest in participating, since the anticipated gains are

Table 2.1 Payoff matrix

***P*1/*P*2**	**Cooperation**	**Defection**
Cooperation	(Y, Y)	$(-I, 2Y + I)$
Defection	$(2Y + I, -I)$	$(-I, -I)$

Notes: This table shows the payoff matrix of the prisoner's dilemma game. Player 1 chooses row strategies, Player 2 plays columns.

systematically negative. This model illustrates the fact that the absence of cooperation may prevent mutually advantageous exchanges from coming about.

The possibilities of cooperation arising between individuals interacting in this type of game have been explored through random matching games based on purely rational individuals encountering one another at random (Kandori, 1992; Ellison, 1994). The horizon of these random matching games is infinite: in each interval each player takes part in a prisoner's dilemma game with a fresh partner drawn at random from the population. Anonymity is retained to the horizon of the game. It is demonstrable that cooperative solutions can emerge as subgame perfect equilibria if the population and the players' preferences for the present are sufficiently small. Equilibrium strategies consist of no longer cooperating, or of cooperating less often, in all future encounters, once a player has participated in a game in which cooperation was chosen by neither partner. It is the threat of a future surge of non-cooperative behavior that may act as an incentive to cooperation at each interval. These results tell us that the spontaneous emergence of cooperative behavior in populations of large size is improbable if each individual is a pure *homo economicus* and they all interact anonymously.

In this setting, cooperation can only emerge as a reputation device and in the presence of punishment. Greif (1993, 1994), in his analysis of the Maghribi and Geneose traders, has shown that the transmission of information, and the coordinated implementation of strategies intended to punish those caught defecting, might facilitate cooperation. Cooperation may exist in the absence of any formal institution defining legal rules if the size of the population and the preference for the present are sufficiently small. If these conditions are unmet, however, formal institutions explicitly laying down legal rules and sanctions are needed in order to sustain cooperation.

The value of such analyses is that they illuminate the role of coordination and of formal institutions. But they cannot account for the cooperative behavior often experimentally observed to arise in anonymous, non-repetitive games. In particular, Henrich et al. (2001) showed that individuals from various societies display cooperation in games absent of any reputational considerations (see the synthesis of Fehr, 2009; Bowles and Gintis, 2007).

2.2.2 Cooperation and Other-Regarding Preferences

To rationalize the existence of cooperation in absence of reputation, the economic literature has incorporated the insights from research in psychology, social science, and behavioral economics, showing the existence of an intrinsic motivation linked to cooperation (see the synthesis by Bowles and Polania-Reyes, 2012; Kahneman and Tversky, 2000). Individuals are motivated by more than material payoffs and value the act of cooperating per se. They have "warm glow preferences" or concerns for reciprocity that favor cooperation.

To modelize this behavior, Francois and Zabojnik (2005), Tabellini (2008b), Algan and Cahuc (2009), Bidner and Francois (2011), Michau (2012), and others, suppose that from non-cooperation there may flow psychological costs. A variant consists of supposing a preference for reciprocity: individuals are altruistic with others who display cooperative behavior, but may sanction those who do not respect cooperative norms (Fehr and Schimdt, 1999; Fehr and Gatcher, 2000; Gintis et al. 2005; Hoff et al. 2011). In all these settings, individuals are assumed to have other-regarding preferences and not just self-regarding preferences, which allow cooperation to emerge in large, anonymous groups.

On the assumption that psychological costs from non cooperation exist, we can modify the payoffs of the trust game described above by adding a cost for non cooperation. In this setting, cooperation becomes a Nash equilibrium, in the previous game described by the payoff matrix above, if the costs from non cooperating, denoted C, are superior to the net individual gain from non cooperation $Y + I$. The term C may be influenced by social and cultural norms, by education, or by the social distance between individuals. For example, Tabellini (2008b) assumes that the psychological costs from non cooperation decrease with social distance: all those sufficiently close cooperate among themselves, but they adopt non-cooperative strategies with those more distant. This assumption is consistent with evidence that individuals tend to distrust more those who are dissimilar to themselves (see Alesina and La Ferrara, 2002).

In this setting, to trust another individual at any one iteration is to embrace the belief that the others taking part in the game are choosing cooperation; that they are, in other words, trustworthy. It is possible to analyze the role of trust in a random matching game where a portion of the population is trustworthy. The trustworthy persons cooperate systematically. Each person knows whether he himself is trustworthy or untrustworthy, but this private information is not available to the others. When two persons meet up, they may decide to go ahead and invest, or pass on the opportunity, in which case they get a payoff equal to zero. If they do go ahead, the trustworthy partners systematically cooperate since not to do so is too costly for them. Conversely, the untrustworthy and purely opportunistic persons always choose to defect.

This modified game can rationalize the existence of cooperation, that is trust, as a Nash-equilibrium. To demonstrate, let us denote by s the portion of trustworthy persons in the population. The expected gain of a trustworthy person who invests amounts to $sY - (1 - s)I$, which implies that such persons invest if the trustworthy portion of the population is superior to $s > I/(Y + I)$. If this condition is unmet, no one has a reason to invest, as all persons who do want to go ahead and invest are necessarily untrustworthy. There are in consequence two possible equilibria depending on the values of s. Either no one invests, if $s < I/(Y + I)$, or in the other eventuality, everyone does. Investment, production, and exchange thus increase with the portion of trustworthy persons in the population, and consequently with trust in others.

Assuming that trust emerges because certain persons are spontaneously cooperative has the advantage of explaining with simplicity why it is that cooperation may arise out of anonymous, non-repetitive interactions. This explanation provides a simple framework to analyze the determinants of trust and its role in the functioning of the economy.

2.2.3 Dynamics of Cooperation

How does cooperation evolve over time? How can cooperative values persist in certain environments and disappear in others? To address this issue, recent works endogenize the transmission of values, along with the seminal work of Bisin and Verdier (2001) stressing the role of family transmission. Parents may inculcate moral values into their children, but these child-rearing choices pose coordination problems, for being honest only pays if others are being honest too. The more other parents are inculcating moral values into their children that will render them trustworthy as adults, the better an option it becomes to raise your children that way too. Building on Hauk and Saez-Marti (2002), Francois and Zabojnik (2005), Tabellini (2008b), Aghion et al. (2010), and Bidner and Francois (2011), we show how such a mechanism might work by introducing education into our model.

Let us assume that the parents get psychological gains, denoted by $G > 0$, an expression of utility, for inculcating honesty-based values into their children and thus ensuring that, as adults, they will systematically be cooperative. In this context, trustworthy adults bear, as before, a psychological cost $C > Y + I$, when they behave dishonestly. Parents get the psychological gain only if their children do behave cooperatively, i.e. do invest. When children do not invest, or in other words, do not display their cooperative behavior, parents do not derive any gain from the values that have been inculcated.

Parents opt for values that maximize the expected utility of their offspring plus their utility gains obtained from inculcating honesty-based values, in the knowledge that each of those children will in turn be randomly encountering others and having to decide whether to go ahead and invest with them or not. The parents' payoff to inculcate honesty-based value equals $G+sY-(1-s)I$, if $s > I/(Y+I)$ and zero otherwise, since their children invest when adults only if $s > I/(Y + I)$. Parents who do not inculcate such values get $s(2Y+I)-(1-s)I$ if $s > I/(Y+I)$ and zero otherwise. The expected gains of education depend on the proportion of trustworthy persons in the generation of the children. It is optimal to bring your children up honestly if the offsetting gains are expected to be equal to or greater, i.e. if $G > s(Y+I)$ and $s > I/(Y+I)$. If this condition is not fulfilled, parents have no incentives to inculcate honesty-based values into children. There will thus be no investment: an economy populated with persons rendered untrustworthy by their upbringing will arrive at a "bad" and feebly productive equilibrium. On the other hand, if one is convinced that the upbringing the other children are receiving from their parents is honesty-based, there may be utility in bringing one's own up the same way. In this case, the economy arrives at a "good" equilibrium, with trustworthy persons and augmented investment and production.

The array of equilibria arrived at in the models of Francois and Zabojnik (2005), Tabellini (2008b), Aghion et al. (2010), and Bidner and Francois (2011) highlights the fragility of the mutual confidence that flows from settling at a good equilibrium. This approach also brings into focus the interaction between moral values and institutions. For example, Aghion et al. (2010) assume that a government elected by majority vote may lay down regulations meant to facilitate mutually advantageous exchange, for the purpose of countering the low levels of spontaneous cooperation that are a concomitant of populations with a relatively small proportion of trustworthy persons in their midst. But these regulations give rise to significant corruption precisely because the proportion of trustworthy persons is small, which keeps distrust alive. Distrust and corruption nourish each other and lead to bad equilibria characterized by weak production and highly burdensome regulation.

Let us enrich this perspective by introducing a dynamic dimension. Let us assume that the gains from inculcating honesty-based values increase as the proportion of trustworthy parents rises. This might be because children are influenced not only by the upbringing they received from their parents, but also by that received from others encountered outside the family circle. Cavalli-Sforza and Feldman (1981) distinguish three modes in which values may be imparted: vertical, oblique, and horizontal. The vertical mode corresponds to transmission from parents to children. The transmission is oblique when the influence comes from adults other than the parents. Horizontal transmission is what those of the same generation have in common. Guiso et al. (2008b) set forth a model that represents several simultaneous modes of transmission, assuming that parents impart beliefs to their children as to the trustworthiness of others, and that children revise this belief set as a function of those whom they encounter. The economy may then be stuck in a bad equilibrium without production, if the beliefs imparted by the parents are too pessimistic, for mutual distrust may impede all exchange (in the game above: everyone passes on the opportunity to invest), and thus stifle all possibility of testing and revising inherited beliefs. Such dynamic sequences have the merit of accounting for the intergenerational transmission of trust empirically observed (Dohmen et al. 2012). They may also explain not only the persistent effect of trust-destroying shocks like the onset of the slave trade in west Africa (Nunn and Wantchekon, 2011), bad colonial institutions (Acemoglu et al. 2001), and legal origins (La Porta et al. 2008), but also the persistent effects of positive shocks like the presence of participatory institutions in the free communes of the Italian Middle Ages (Putnam et al. 1993; Guiso et al. 2008a).

2.3. EMPIRICAL MEASURES OF TRUST

To measure the impact of cooperative values on economic development and institutions, one has to define the empirical counterpart of the trusting behavior at play in the previous theoretical games.

2.3.1 Definition of Trust

Research on the relationship between trust and growth focuses essentially on generalized trust, in other words, on relations among individuals who are not bound by the kind of personal ties that bind members of the same family, or fellow workers. In this context, the generally used definition of trust is taken from Coleman (1990), according to whom "an individual trusts if he or she voluntary places resources at the disposal of another party without any legal commitment from the latter, but with the expectation that the act of trust will pay off." One of the advantages of this approach is to define trust as a behavior that can be directly measured with experimental games, as shown by Fehr (2009). Defined this way, trust is also linked to the notion of social capital utilized by Fukuyama (1995), Putnam (2000), and Guiso et al. (2011), for whom social capital is the ensemble of "those persistent and shared beliefs and values that help a group overcome the free rider problem in the pursuit of socially valuable activities."

2.3.2 Measures of Trust

Trust can be measured by using surveys and laboratory experiments. Empirical research investigating the link between growth and trust usually draws on answers from survey questions. The reason for this is the availability of surveys, which cover a large number of countries since the beginning of the 1980s. Nevertheless, these surveys evoke difficulties in interpretation. Besides the polysemy of questions and responses, it is not sure that the individuals who declare to have strong trust in others actually behave in a more cooperative way. For that reason, researchers have undertaken laboratory experiments as well as field experiment paired with surveys, in order to better capture their scope.

2.3.2.1 Surveys

In surveys, the measure of trust is most often measured with the "generalized trust question" first introduced by Almond and Verba (1963) in their study of civil society in post-war Europe. This question runs as follows: "Generally speaking, would you say that most people can be trusted, or that you can't be too careful when dealing with others?" Possible answers are "Most people can be trusted" or "Need to be very careful." The same question is used in the European Social Survey, the General Social Survey, the World Values Survey, Latinobarómetro, and the Australian Community Survey. Surveys generally include other questions related to trust. For instance, the WVS asks the "fair question": "Do you think most people would try to take advantage of you if they got the chance, or would they try to be fair?" The GSS includes the trust question, the fair question, and adds the "help question": Would you say that most of the time people try to be helpful, or that they are mostly just looking out for themselves? These different questions are sometimes used to build indexes that intend to provide alternative measures of trust or get an average indicator of moral values or civic capital (Tabellini, 2010; Guiso et al. 2011).

The resulting survey data supply us with subjective information that certainly demands cautious interpretation. These questions raise concerns about interpretation. In particular, individuals who respond that you need to be very careful to the trust question could be motivated by a strong aversion against risk (see for these topics, Fehr, 2009; Bohnet and Zeckhauser, 2004). However, most important for investigating empirically the relation between growth and trust is to know whether the responses to the trust question are linked to actual cooperative behavior.

2.3.2.2 Experimental Games in the Lab

Contributions have analyzed the relationship between responses to the trust questions or to connected questions and the behavior in experimental games. In general, these works use variants of the "investment game," known also as the "trust game," of Berg et al. (1995) presented above. In laboratory experiments, this game is played as follows. In stage 1, the subjects in rooms A and B are each given 10 dollars as a show-up fee. While subjects in room B pocket their show-up fee, subjects in room A must decide how much of their 10 dollars to send to an anonymous counterpart in room B. The amount sent, denoted by M, is tripled resulting in a total return $3M$. In stage 2, a counterpart in room B is given the tripled money and must decide how much to return. One measures "trust in others", as defined by Coleman (1990), by the amount sent initially by the sender. Trustworthiness is measured by the amount sent back by the player in room B.

The first contributions that analyzed the relationship between survey-answer from the generalized trust question and the amount sent in the trust game found mixed results. Glaeser et al. (2000) measured the relation between questions related to trust in surveys and the behavior of participants in trust games. This study was carried out at Harvard University, where 274 students were asked the trust question before they played the trust game either in the role of sender or receiver. The authors find that although questions about trusting attitudes do not predict trusting behavior, such questions do appear to predict trustworthiness. Holm and Danielson (2005) find a positive correlation between behavior in games and answers to the trust question in Sweden, but not in Tanzania. Lazzarini et al. (2005) find a correlation in face-to-face, non-anonymous trust games in Brazil. Other experiments have been run on representative surveys, with also contrasting results. While Fehr et al. (2002) find that the trust question does predict trusting behavior but not trustworthiness, Ermisch et al. (2009) find exactly the opposite on a representative sample of the British population.

These results are difficult to compare, as the designs of the games are not perfectly identical between the different experiments. While in the game organized by Glaeser et al.

the second movers do not receive any initial payment, in the game of Berg et al. all participants get a show-up fee. This could explain, why a great fraction (70%) of first movers send all their initial endowment to the second movers, in the experiment of Glaeser et al. To measure the level of trust, it is therefore necessary to distinguish this component from other attitudes, such as risk aversion, altruism, and reciprocal behavior. In addition, does trusting behavior measured during those different experiments really capture deep-seated preferences? Or do they just relate to beliefs about the level of civility of others, which can be quickly revised?

This kind of behavior observed in experiments might be as much motivated by altruism as by trust, in the sense of the definition by Coleman. With regard to the positive correlation between the responses to the trust questions and the amounts sent back by the second mover, this correlation could be the consequence of a concern about reciprocity, characterizing the individuals who declare themselves to trust strongly. Thus, the absence of a correlation between the responses to the trust question and the amounts sent by the first movers in the study of Glaeser et al. does not necessarily imply that the trust questions are not correlated with trust in the sense of Coleman, because the amounts sent by the senders are probably strongly influenced by motivations of altruism.

Cox (2004) has proposed an experimental design with the goal of identifying the relative contributions of trust and altruism to the amounts sent in the first stage of the trust game. To achieve this, he compares the results of a trust game, as described above, with those of a dictator game, in which the only difference to the trust game is the absence of a decision by the second movers: thus, they do not have an opportunity to return any money that they receive. The dictator game serves to measure altruism, whereas trust is measured by the difference between the amount sent during the first stage of the trust game and the amount sent in the dictator game. The experiments conducted by Cox show that the trust motive in fact exists, in addition to altruism. The experimental design created by Cox also allows us to identify motives of reciprocity, by comparing the amounts returned in the second stage of the trust game with those sent in a game which differs from the trust game. Here too, the experiments realized by Cox shows the existence of reciprocity.

Cox's design allows us to distinguish between motives of altruism, trust, and reciprocity. Capra et al. (2008) used this design to analyze the relationship between those motives as defined above and attitudes gained from answers to survey questions, by conducting experiments with students from Emory University. They find the same results as Glaeser et al. concerning the trust question, that is, that the responses are not correlated with the amounts sent by the first movers, but with the amounts sent back by the second movers, who sent back more depending on how trusting in others they declared themselves to be in the survey.

However, this correlation disappears as soon as the level of altruism is controlled for. Besides, the amounts sent by the first movers are well correlated with the responses to the "help question" or the "fair question" when altruism is controlled for. Responses to the trust question are not correlated significantly with the amounts sent by the first movers, but the sign of the coefficient indicates an increasing relation between declared trust and the amounts sent. It is possible that the absence of a significant relation results from the low number of observations (62), which is especially problematic for the trust question, whose wording is particularly vague. In short, this contribution suggests an experimental design which distinguishes the motives of trust, altruism, and reciprocity, allowing to identify coherent relations between attitudes declared in answers to survey questions and actual behavior in trust games.

Other studies have also made use of neurobiological methods to measure, with greater precision, the role of trust in comparison with other individual characteristics in the behavior of participants of the trust game. It is known that oxytocin, a hormone released especially during breast-feeding and delivery, is associated with sentiments of affinity and socialization. In particular, research in neurobiology has shown that this hormone plays a central role in behavior related to social connectivity, such as parental and couple relations. Additionally, this hormone significantly reduces stress and anxiety in situations of social interaction. It is known for deactivating the transmission of feelings of anxiety related to the belief of being betrayed. Kosfeld et al. (2005) had the ingenious idea to evaluate the effect of oxytocin on pro-social behavior of individuals participating in trust games. The authors also proposed additional experimental designs to distinguish the pro-social preferences from risk-taking behavior and from beliefs like the level of optimism of the participants. The participants in this study were randomly allocated into two groups. The first group inhaled oxytocin through a spray, the second inhaled a placebo and served as the control group. The results of this experiment are illuminating. Those individuals who received oxytocin tended to display stronger trust behavior. What is even more remarkable, is that those individuals continued to behave trustingly in the exchange with the others, even if the latter didn't reciprocate. By contrast, other attitudes, such as prudence and risk-aversion, or even other beliefs such as optimism in the actions of the others, are not affected. Kosfeld et al. (2005) conclude that the trust game measures veritable preferences for cooperation, and not risk-aversion or anticipation of the others' actions (see Fehr, 2009, for a survey on experimental measures of trust).

2.3.2.3 Experimental Games in the Field

Obviously, the presence of a relationship between survey answers and behavior in trust games does not imply that answers to survey questions allow us to predict daily behavioral patterns, insofar as the latter can be different from those observed in laboratory experiments. We still know very little, however, about whether, and to what extent, the experimental results established in the laboratory carry over to field situations. At this stage, it thus seems key to investigate the relationship between the experimental measures usually elicited in the laboratory and the *field* outcomes of interest, if we are to rely on the experimental method to make inferences about the real world.

In his pioneering work, Karlan (2005) uses the trust game to obtain individual measures of taste for reciprocity, and shows that it can be used to predict loan repayment among participants, up to one year later, in a Peruvian microcredit program. Oliveira et al. (2009) elicited subjects' taste for cooperation in the laboratory using a traditional public goods game. They show that the results are correlated with subjects' contributions to local charities in a donation experiment and with whether they self-report contributing time and/or money to local charitable causes. Similarly, Laury and Taylor (2008) use public goods games to elicit their subjects' taste for cooperation and show that it is associated with the probability to contribute to a field public good in a donation experiment. One prominent limitation of these two studies is that they both obtain information about "field" behavior in the laboratory itself, either through contextualized experiments or self-reports. In this case, one might worry about possible spurious correlations caused by demand effects and/or individuals' willingness to remain self-consistent. Still relying on highly contextualized donation experiments, Benz and Meier (2008) address part of this concern by collecting field data about their subjects' behavior in a charitable giving situation prior to conducting a charitable giving experiment in the classroom, and obtain a significant correlation between both measures.

A promising avenue of research is to extend experimental games to online economics or wikinomics. In particular, the emergence of large organizations based on cooperation and non-monetary incentives, such as Wikipedia and open software, provide a perfect field experiment to test the relationship between experimental measures and field behavior.

In a recent contribution, Algan et al. (2012a) explore this question in one of the most successful contemporary instances of massive voluntary contributions to a public good: the online encyclopedia Wikipedia. Using an Internet-based experimental economics platform, the author elicited preferences for cooperation, altruism, and reciprocity among a sample of 850 Wikipedians directly in the field (i.e. online, in interaction with other Internet users who are not Wikipedia contributors) and related those measures to their real-world contribution records. They find that contributions to Wikipedia—as measured by subjects' number of edits to the encyclopedia—are related to their propensity to cooperate in a traditional public good game and to the level of reciprocity that they exhibit both in a conditional public good game and in a trust game. Moving from the position of a non-contributor with a registered Wikipedia account to that of an experienced Wikipedia contributor is associated with a 10–13% rise in public good contribution levels and with a 7–10% rise in reciprocity levels.

2.3.3 Correlation Between Generalized Trust and Limited Trust

We stressed that most of the research about the economic consequences of trust deals with generalized trust. But what is the relationship between the various forms of trust? Since the seminal work of Banfield (1958) and Coleman (1990), social scientists make a distinction between limited versus generalized morality. Societies with limited morality only promote codes of good conduct within small circles of related persons (kin), whereas selfish behavior is regarded as morally acceptable outside the small network. This behavior was famously described as "amoral familism" by Banfield (1958) in his ethnographic description of a rural village. Societies with generalized morality promote good conduct outside the small family/kin network, offering the possibility to identify oneself with a society of abstract individuals or abstract institutions. Coleman (1990) proposes a similar distinction between strong ties, defined as the quality of the relationship among family members, and weak ties, defined as the strength of social relationships outside the family circle.

Ermisch and Gambetta (2010), using trust games with a representative sample of the British population, find that people with strong family ties have a lower level of trust in strangers than people with weak family ties, and argue that this association is causal. They show that the explanation for this opposition comes from the level of outward exposure: factors that limit exposure, limit subjects' experience, as well as motivation to deal with strangers.

Greif and Tabellini (2010) provide an historical analysis of this opposition by comparing the bifurcation of societal organization between pre-modern China and medieval Europe. Pre-modern China sustained cooperation within the clan, e.g. a kinship-based hierarchical organization in which strong moral ties and reputation among clan members played the key role. By contrast, in medieval Europe, the main example of a cooperative organization is the city, whereby cooperation is across kinship lines with weak ties, and external enforcement played a bigger role.

2.3.4 Heterogeneity of Trust Across Space

As early as the 18th century, Adam Smith (1997 [1766]) was already alluding to substantial differences across nations in what he called the "probity" and "punctuality" of their populations. For example, the Dutch "are the most faithful to their word." Similarly John Stuart Mill observed: "There are countries in Europe ... where the most serious impediment to conducting business concerns on a large scale, is the rarity of persons who are supposed fit to be trusted with the receipt and expenditure of large sums of money" (Mill, 1848, p. 132).

Recent advances in international social survey technique have yielded further evidence of the enormous differences in trust level that may exist across countries. In social survey data there is to be observed a sizable variation in the extent to which people trust others across countries as well as within countries.

Figure 2.1a and 2.1b show average levels of generalized trust for 111 countries, generated from responses to the World Values Survey, the European Values Survey, and the Afrobarometer.[1] These surveys ask the trust question, and the trust variable takes on the

[1] The data set is constructed by combining the five waves of the WVS (1981–2008) with the four waves of the EVS (1981–2008), and adding the third wave of the Afrobarometer (2005).

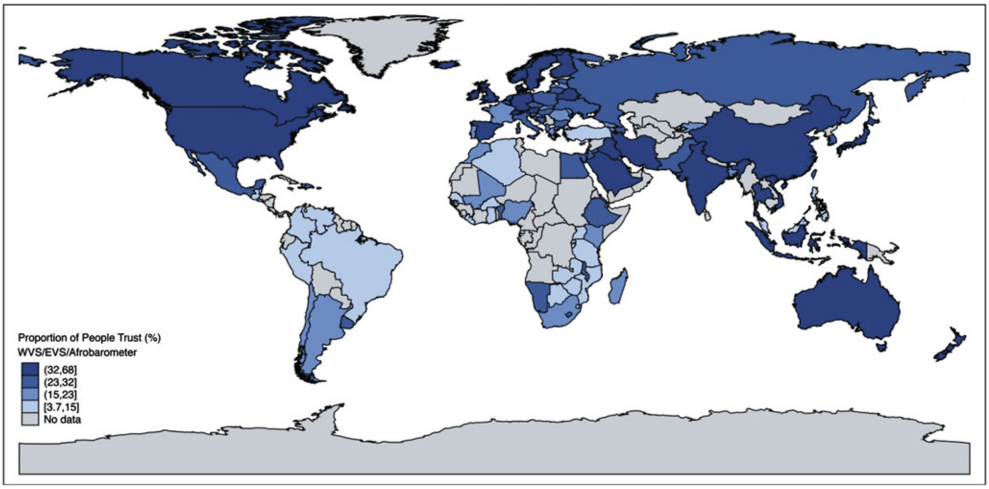


Figure 2.1a World distribution of trust. *Sources: Trust is computed as the country average from responses to the trust question in the five waves of the World Values Survey (1981–2008), the four waves of the European Values Survey (1981–2008), and the third wave of the Afrobarometer (2005). The trust question asks "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" Trust is equal to 1 if the respondent answers "Most people can be trusted" and 0 otherwise.*

value 1 if the respondent answers that "Most people can be trusted" and 0 if he or she thinks that one "Needs to be very careful." Trust levels vary very considerably from one country to another. In Norway, the country with the highest level of trust in the sample, more than 68% of the population trusts others. At the opposite end of the ranking lies Trinidad and Tobago, where only 3.8% of the population exhibits interpersonal trust. The United States ranks in the top quarter, with an average trust level of more than 40%. In general, northern European countries lead the ranking with high average levels of interpersonal trust, while populations in African and South American countries seem not to trust others very much.

The extent to which people trust other, however, varies not only across countries, but also across regions belonging to the same country. Figure 2.2 shows average trust levels for 69 European regions used in Tabellini (2010); the source is the World Values Survey (1990–1997). As we see from the figure, trust levels vary remarkably between regions lying not very far apart. While in the Dutch region of Oost Nederland more than 64.1% trust is shown, in the French Bassin Parisien region this figure is only 14.2%. There is wide divergence between regions within European countries. In Italy, the trust level is almost twice as high in Trento (49%) as it is in Sicilia (26%). In France, trust is 13% points higher in the Sud Ouest region compared to the Nord region. Finally, a divergence in trust levels is also observable in federations. Figure 2.3 displays mean trust levels for 49

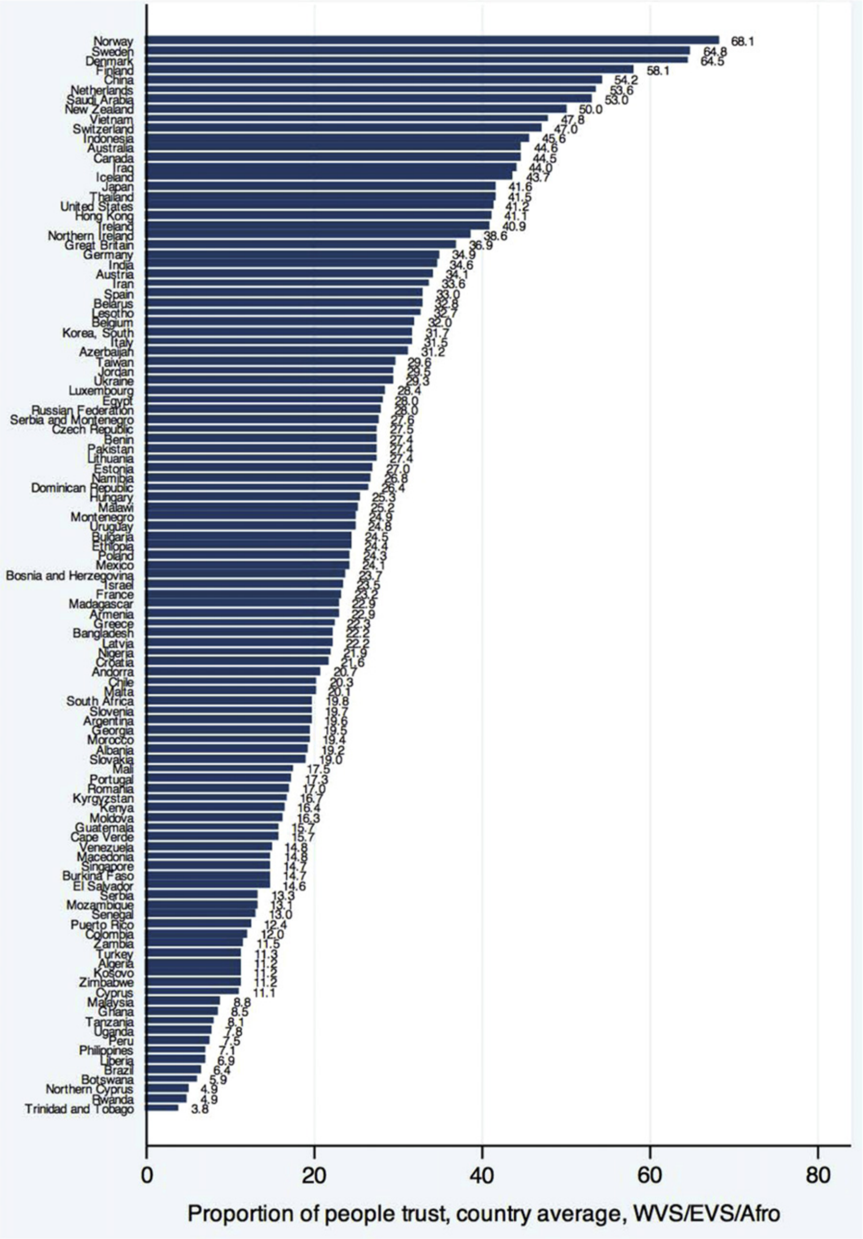


Figure 2.1b Average trust levels in 111 countries. *Sources: Trust is computed as the country average from responses to the trust question in the five waves of the World Values Survey (1981–2008), the four waves of the European Values Survey (1981–2008) and the third wave of the Afrobarometer (2005). The question asks "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" Trust is equal to 1 if the respondent answers "Most people can be trusted" and 0 otherwise.*

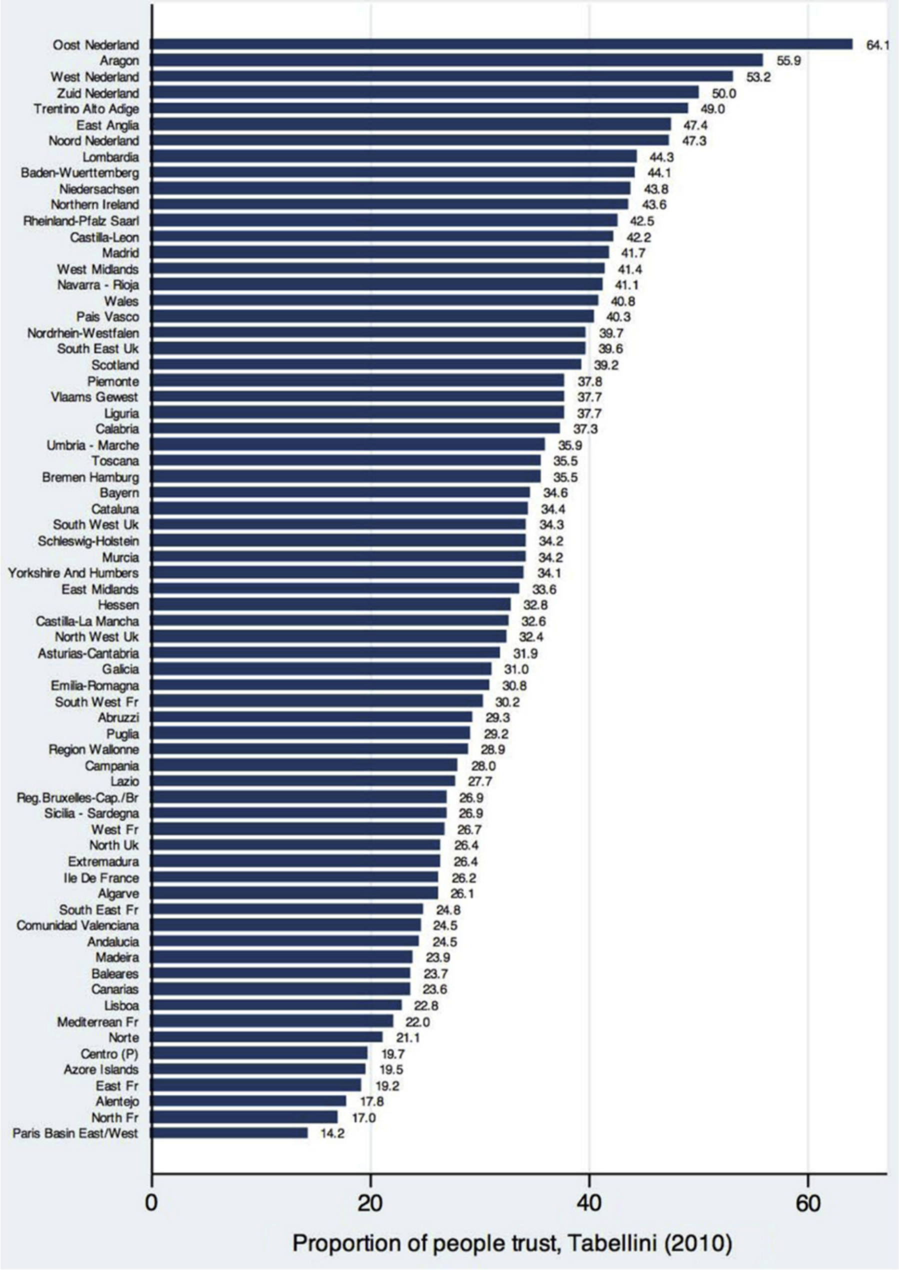


Figure 2.2 Average trust levels in 69 European regions. *Source: The proportion of people that trust is taken from Tabellini (2010). The trust measure is computed as the regional average from responses to the question "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" Trust is equal to 1 if the respondent answers "Most people can be trusted" and 0 otherwise.*

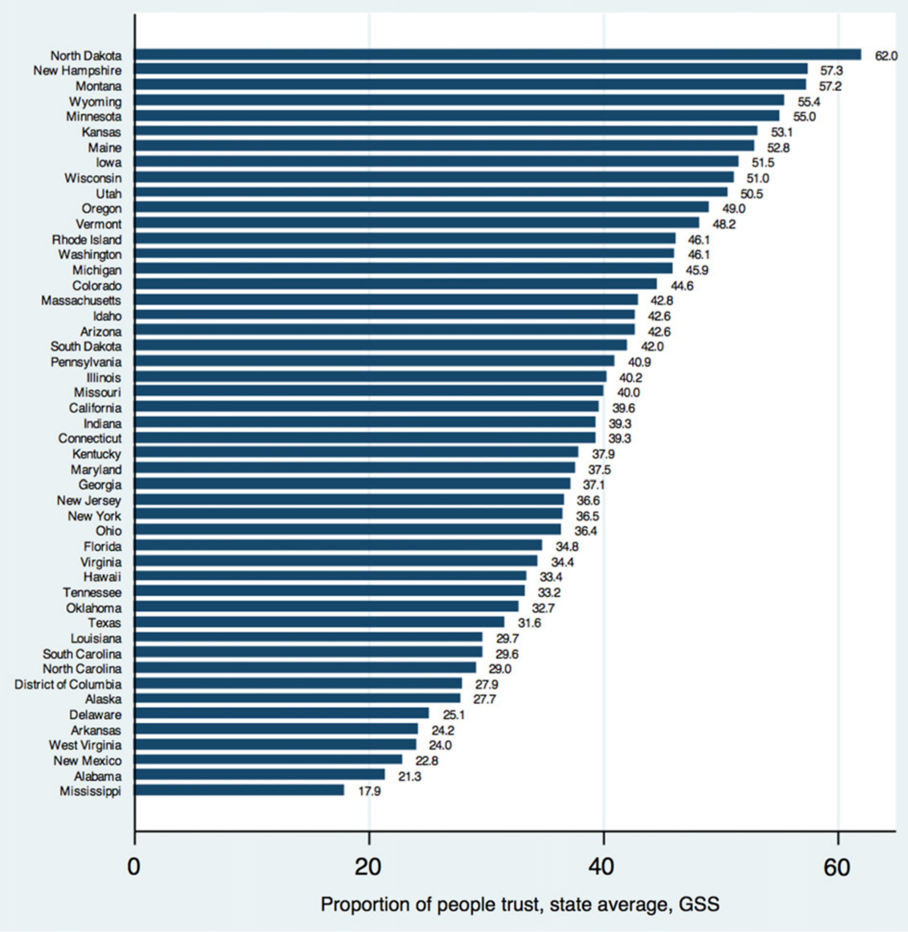


Figure 2.3 Average trust levels in 49 US states. *Sources: The proportion of people that trust is taken from the General Social Survey (1973–2006). The trust measure is computed as the state average from responses to the question "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" Trust is equal to 1 if the respondent answers "Most people can be trusted" and 0 otherwise.*

US states, computed by averaging individual responses from the General Social Survey (GSS, 1973–2006) of the United States. We note wide differences in the degree of trust the citizens of these States have in others. While in North Dakota more than 60% of the respondents trust others, in California less than 40%, and in Mississippi not even 20%, of the respondents think that they can trust people in general.

2.3.5 An Heterogeneity Linked to National Specificities

What are the reasons for the divergence in trust levels across countries? Besides individual characteristics (e.g. age, social status, gender, education, income, and religion), time-invariant country characteristics can account for a large share of the disparity of trust levels around the world.

Table 2.2 reports a micro-regression of individual trust on age, age squared, gender, education, income level, and various types of religious affiliation. Some of these individual characteristics are highly correlated with individual trust. Maleness correlates positively with trust, and age displays a hump-shaped relationship with trust. More educated individuals have significantly higher trust, a relationship documented at length by Helliwell and Putnam (2007). A one standard deviation increase in education (roughly 2.2 years) increases trust by 11% of its sample mean. Trust also correlates positively with income: a one standard deviation increase in income (roughly 0.79) increases trust by 6% of its sample mean. In a seminal paper on the determinants of trust, Alesina and La Ferrara (2002) document the role of additional characteristics negatively correlated with trust, such as a recent history of traumatic experiences or belonging to a group that historically felt discriminated against, such as women or ethnic minorities.

But the feature that especially stands out in Table 2.2 is the very weak predictive power of individual characteristics for explaining cross-country heterogeneity in trust compared to country fixed effects. Including country fixed effects in this regression increases the coefficient of determination, R sq. by about 10% from 0.027 to 0.12. Furthermore, the correlation between average country trust levels and the predicted mean trust is of a magnitude 0.52 without fixed effects, and rises to an almost perfect correlation of 0.99 when country fixed effects are included in the micro-regression.

Figure 2.4 displays country fixed effects in relation to Norway, the country with the highest mean trust in the sample, taken from the above-described micro-regression. The figure thus documents the % point reduction in trust flowing from the fact of living in a country other than Norway, with all individual characteristics (age, gender, education, income, and religion) held constant. In comparison to Norway, trust would be reduced by more than 60 pp (percentage points) in Uganda, Peru, Kosovo, or Algeria; by more than 50 pp in Greece or France; and by around 40 pp in Italy, Germany, or the United States. The country fixed effects thus differ by an order of magnitude from the effects of individual characteristics. This result suggests that it is necessary to look at national characteristics (institutions, history, geography, public policy…) in order to understand how trust is built.

2.4. THE DYNAMICS OF TRUST

International surveys underline how important the heterogeneity of average levels of trust across countries is, for identical characteristics of the inhabitants, such as age,

Table 2.2 Determinants of trust: micro estimates

	Trust	
	(1)	**(2)**
Age	0.003***	0.001***
	(.000)	(.000)
Age sq.	−0.000**	−0.000
	(.000)	(.000)
Gender	0.009**	0.004
	(.003)	(.003)
Education	0.019***	0.015***
	(.004)	(.003)
Protestant	0.165***	0.013
	(.051)	(.009)
Catholic	−0.011	−0.004
	(.200)	(.006)
Hindu	0.107**	0.023
	(.053)	(.023)
Buddhist	0.057	0.010
	(.042)	(.013)
Muslim	0.034	0.021*
	(.047)	(.011)
Jew	−0.030	0.045
	(.018)	(0.032)
Income level	0.020***	0.023***
	(.004)	(.003)
Country FE	No	Yes
Observations	136105	136105
R^2	0.027	0.123

Notes: The dependent variable is *Trust*. It is calculated from answers to the question *"Generally speaking, would you say that most people can be trusted, or that you need to be very careful in dealing with people?"*. Trust is equal to 1 if the respondent answers *"Most people can be trusted"* and 0 otherwise.

Control variables include age in years, Gender (1 = Male), Education (from 1 = No elementary school to 7 = Graduate studies), Income (1 = Below national average, 2 = Average, 3 = Above national average), and dummy variables indicating the religious denomination of the respondent.

Column (2) includes country fixed effects. OLS regressions with robust standard errors clustered at the country level in parentheses.

Sample (79 countries): Albania, Algeria, Argentina, Armenia, Austria, Azerbaijan, Bangladesh, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Canada, Chile, China, Croatia, Cyprus, Czech Republic, Denmark, Egypt, Estonia, Finland, France, Georgia, Germany, Great Britain, Greece, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Japan, Jordan, Kosovo, Kyrgyzstan, Latvia, Lithuania, Luxembourg, Macedonia, Malta, Mexico, Moldova, Montenegro, Morocco, Netherlands, Nigeria, Northern Cyprus, Northern Ireland, Norway, Pakistan, Peru, Philippines, Poland, Portugal, Puerto Rico, Romania, Russian Federation, Saudi Arabia, Serbia, Singapore, Slovakia, Slovenia, South Africa, South Korea, Spain, Sweden, Switzerland, Tanzania, Turkey, Uganda, Ukraine, United States, Venezuela, Vietnam, Zimbabwe.

Sources: World Values Survey (1981–2008) and European Values Survey (1981–2008).

*Coefficient is statistically different from 0 at the .10 levels.

**Coefficient is statistically different from 0 at the .05 levels.

***Coefficient is statistically different from 0 at the .01 levels.

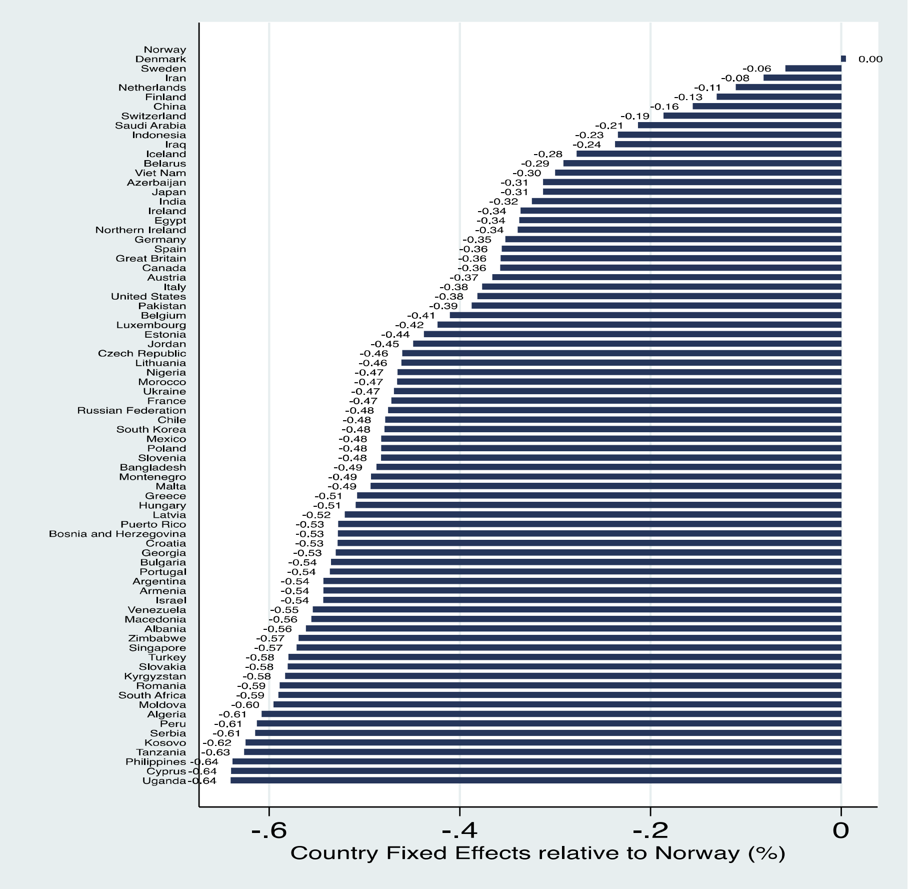


Figure 2.4 Country fixed effects relative to Norway (%). *Interpretation:* Holding individual characteristics constant, living in Uganda rather than in Norway reduces trust by 64%. *Additional controls:* Age, age (squared), gender, education, income, and religion. *Sources: Trust is computed as the country average from responses to the trust question in the five waves of the World Values Survey (1981–2008), the four waves of the European Values Survey (1981–2008), and the third wave of the Afrobarometer (2005). The question asks "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" Trust is equal to 1 if the respondent answers "Most people can be trusted" and 0 otherwise.*

income, education, and religion. These surveys also show that average trust changes little over the course of time: the countries with the weakest levels of trust at present also had weak trust at the beginning of the 1980s. This observation, though, tells us little. For one thing, it is confined to the relatively short period for which survey data are available. For another, it says nothing about the causal factors that may explain the persistence or the evolution of trust. A cluster of recent studies make it their goal to seek these out.

2.4.1 Climate

Four centuries before our era, Aristotle underlined the influence of climate on attitudes: "The nations that live in cold regions and those of Europe are full of spirit, but somewhat lacking in skill and intellect; for this reason, while remaining relatively free, they lack political cohesion and the ability to rule over their neighbors. On the other hand the Asiatic nations have in their souls both intellect and skill, but are lacking in spirit; so they remain enslaved and subject. The Hellenic race, occupying a mid-position geographically, has a measure of both, being both spirited and intelligent" (Politics 7.7, 1327b18–1328a21, trans. Sinclair and Saunders).

When Aristotle wrote the above, sampling was unknown, and there was no way to establish a statistical relationship between climate and attitudes; today it is at least feasible to contemplate doing so. Durante (2010) posits that the inhabitants of Europe's regions are today more trusting to the extent that they were subjected to significant climatic variations between 1500 and 1750. The explanation advanced by Durante is that greater climatic variability, which heightens the undependability of harvests, makes it necessary to stock larger reserves, manage them collectively, and develop trade between regions affected by differing and therefore offsetting climatic shocks. All this favors cooperation and leaves an imprint on the overall social structure. Family bonds are less binding in regions where the amplitude of climatic variation is greater. Young people leave the family nest earlier, since they cannot count on family solidarity to meet their needs when harvests are poor, as they frequently are. Experiments in cooperation induced by climatic harshness may thus have effects persisting across a span of centuries, even as societies are profoundly transformed by the passage from the agricultural stage to the industrial stage.

In a similarly oriented contribution Ostrom (1990) found that trust is high in upland regions where farmers must cultivate scattered plots irrigated by communally maintained ditches. In such regions, mutual trust and cooperation in all facets of life are more frequent than on flatland that can be farmed with much less coordination.

Natural catastrophes can also influence trust, sometimes in unforeseen ways. A portion of those who survive experience a post-traumatic phase during which they turn to others, show altruistic behavior, and invest in communal action. This "catastrophe syndrome" (Valent, 2000; Wallace, 1956) may last a long time and have a durable effect. Castillo and Carter (2011) and Zylberberg (2011) have shown that destructive hurricanes may favor cooperation and trust over a period of years.

2.4.2 The Weight of History

The traffic in slave labor to work plantations in the Americas began in the 16th century, when West African men and women were captured and enslaved during raids led from the coast by Europeans, or sold as slaves to the Europeans after being captured in the course of military conflicts among African belligerents. But the system underwent evolution, for some inhabitants of West Africa found they could survive and even thrive by capturing and selling other humans—passing travelers, neighbors, even members of their own families—to the slave merchants. It may be surmised that these practices, widespread at the time, instilled profound mistrust in the population. Nunn and Wantchekon (2011) have shown that it is still present three centuries later. The inhabitants of these regions still reveal greater mistrust of others, including their neighbors, the members of their ethnic group, and even their own families, than the inhabitants of neighboring regions. The slaves may of course have been captured and sold primarily in areas of conflict, where distrust would have been higher to start with, and the task of the slave merchant correspondingly easier. Nunn and Wantchekon have shown, however, that dwellers in regions more remote from the Atlantic coast, whose ancestors were relatively more sheltered from the slave trade, are less distrustful than those who dwell nearer the coast. They also show that this pattern of diminishing distrust with increasing distance from the coast is not observed in other regions of the globe. This would tend to show that the regions where the slave trade flourished are the ones with distrustful inhabitants, not the converse.

Thus, even across a span of many generations, history may have the effect of shaping trust in ways that we can still perceive. Rohner et al. (2013) provide a theory for the long-run impact of war and conflicts on distrust. Accidental conflicts, e.g. conflicts that do not represent economic fundamentals, might still lead to a permanent breakdown of trust, since agents observe the history of conflicts to update their beliefs and to transmit them over generations. Becker et al. (2011) have studied the imprint left by the Habsburg Empire, which dominated much of central Europe from the 18th century to the beginning of the 20th, and employed administrators who, with respect to the norms of the age, were better educated and less corrupt. The borders of the countries that have come into existence since the collapse of the Empire at the end of World War One may have altered more than once in the interval, as a result of conflicts and political events. Yet in regions that once lay within the boundaries of the Empire, the administration is still more transparent, less corrupt, and better trusted by the population. The improved administrative practices of the Habsburgs left traces that have survived well beyond the dissolution of their Empire.

The weight of this example is more than anecdotal. Numerous circumstances of European history reveal that political decisions can affect trust over the course of many centuries. Today the inhabitants of Italian cities that in the Middle Ages achieved a form of participatory self-government, the communal regime, comparable to that of the city-states of antiquity, and whose ancestors were thus deeply engaged in civic/political life, participate more in elections, give more blood, and are more likely to join associations than the inhabitants of other Italian cities (Guiso et al. 2008a). Regions of Europe endowed with higher levels of education and a more democratic or participatory state form at the end of the eighteenth century today have more trusting and civic-minded inhabitants (Tabellini, 2010). This line of research suggests that education and democracy shape civic behavior in ways that last for centuries.

In the same vein, Jacob and Tyrell (2010) have shown that the activities of the *Stasi*, the state security agency of the former DDR or East Germany as it was known, which by 1989 employed more than 90,000 permanent members and had more than 170,000 informers, have left a durable mark on the civic attitudes of the inhabitants of East Germany. Everyone knew that, in every building and factory, they were being watched by informers among them, and that electronic eavesdropping was in widespread use. Anything one said about the regime might be reported, and twisted in such a way as to ruin one's life. Jacob and Tyrell show that this climate of delation shredded the social fabric. Two decades after the wall came down, the inhabitants of regions in which the *Stasi* were once particularly active are less inclined to do their civic duty: their rate of voter turnout, their rate of participation in voluntary associations, and their rate of voluntary organ donation are all measurably lower than those in the rest of the *Bundesrepublik*.

More generally, Aghion et al. (2010) highlight a steep decline of trust in the former Soviet bloc countries at the time of their conversion to capitalism. The market liberalization at the turn of the 1990s, with its attendant corruption, in this Eastern bloc setting of pervasive distrust and minimal transparency, seems to have degraded any trust the citizens might have had in their state, their justice system, or their fellow citizens, even further. The effect was most detectable in regions where trust was already low at the time the wall came down.

Another potential long-term cause of trust is related to genetic diversity. In a fascinating recent contribution, Ashraf and Galor (2013) show that distance from the cradle of humankind in East Africa is associated with lower genetic diversity within ancient indigenous settlements across the globe. As subgroups of the populations of parental colonies left to establish new settlements, they carried with them only a subset of the overall genetic diversity of their parental colonies. As a result, the migratory distance from East Africa has an adverse effect on genetic diversity in the different ethnic groups populating the globe. Ashraf and Galor then show that genetic diversity affects significantly trust and cooperation, leading to an optimal level of diversity for economic development. On one hand, genetic heterogeneity increases the likelihood of mis-coordination and distrust, reducing cooperation and lowering total factor productivity. On the other hand, diversity has a beneficial effect on the expansion of society's production possibility frontier by widening the spectrum of complementary traits.

2.4.3 Inherited Trust

Studies of how immigrant attitudes evolve as a function of their country of origin and country of arrival shed an interesting light on the malleability of trust. They show that the beliefs and behaviors of immigrants are influenced by their countries of origin; that football players who grew up in countries undergoing civil war are more violent than other players, that they get yellow-flagged or red-flagged more often (Miguel et al. 2011). Fisman and Miguel (2007) observed that UN diplomats from countries with low levels of trust and civic spirit frequently violate the New York City parking laws, from which diplomats are legally immune, whereas those from Scandinavian and Anglophone countries make it a point not to, although they enjoy the same immunity.

Still, the attitudes and beliefs of immigrants are not carved in stone but are influenced by their countries of residence. As a general rule, trust rises among immigrants right from the first generation, if they have moved from a low-trust country to a high-trust one. The converse holds true as well. This phenomenon has been observed in both the US and Europe (Algan and Cahuc, 2010; Dinesen, 2012; Dinesen and Hooghe, 2010). In fact, it is detectable in cases of internal migration too: the civic spirit of Italians who move from southern Italy to the north tends to ameliorate and converge gradually on the prevailing local norm. Conversely, the civic spirit of Italians who move from the north to the south shows some signs of degrading (Ichino and Maggi, 2000; Guiso et al. 2004). Algan et al. (2011) illustrate this pattern with the evolution of trust among the first and second generation of immigrants in European countries. In the European Social Survey, the level of trust of first generation immigrants correlates significantly with the level of trust in their country of origin. By contrast, the level of trust of second generation immigrants is more correlated with the average level of generalized trust and trust in institutions in their new country of residence than with trust in their home country.

Individual distrust, therefore, is not something poured and set for eternity. The environment can modify it. But it is something systematically characterized by the kind of inertia that can leave its mark on at least one and perhaps more generations.

2.5. TRUST, INCOME PER CAPITA, AND GROWTH

To what extent can the above-mentioned cross-sectional heterogeneity in trust level account for cross-sectional differences in income per capita? To what extent can a boost in trust explain economic success within a country? This section first documents the evidence on the strong correlation observed between trust and economic outcome. We then document the main issues raised by the identification of the causal impact of trust, and the recent attempts in the literature to address them.

2.5.1 Cross-Section Correlation

The interest of the economic literature in social capital is fueled by the strong positive correlation between income per capita and average trust levels across countries or regions, first illustrated by the seminal work of Knack and Keefer (1997). The classic book by Putnam et al. (1993) also suggested the existence of such a relationship across regions in Italy by arguing that the northern regions developed faster than the southern ones because the former had a higher stock of social capital measured by association membership.

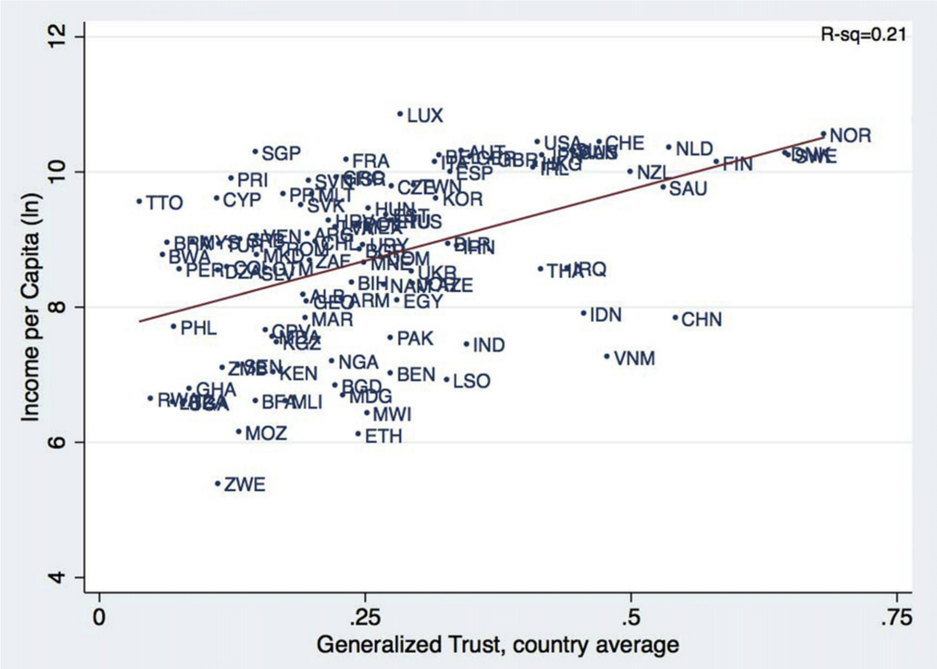


Figure 2.5 Cross-country correlation between average (ln)-income per capita and trust. *Sources: Average income per capita (1980–2009) has been obtained from the Penn World Tables 7.0. Trust is computed as the country average from responses to the trust question in the five waves of the World Values Survey (1981–2008), the four waves of the European Values Survey (1981–2008), and the third wave of the Afrobarometer (2005). The question asks "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" Trust is equal to 1 if the respondent answers "Most people can be trusted" and 0 otherwise.*

Figure 2.5 plots the average (ln) income per capita between 1980 and 2009 against average trust between 1981 and 2008 for a sample of 106 countries. Countries with higher levels of trust also display higher income levels. The correlation is steady; one fifth of the cross-country variation in income per capita is related to differences in generalized trust.

Table 2.3 shows the regressions of income per capita (ln) on trust. A one standard deviation increase in trust, about 0.14, increases (ln) income per capita by 0.59, or 6.8% of its sample mean. When additional controls for education, ethnic fractionalization, and population are included (column 2), the coefficient for trust remains significant but decreases in magnitude. Increasing trust by one standard deviation leads to a rise in income per capita of 0.18, or 2% of the sample mean. As a comparison, increasing fractionalization by one standard deviation (2.5) decreases income by 0.225 or 2.5% of the mean. We additionally control for several institutional measures, such as legal origins (column 3) and political institutions (column 4). Trust remains significant at the 5 or 10% level, while the institutional variables are insignificant.

Table 2.3 Trust and income: cross-country correlation

	Ln GDP per capita (1980–2009)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Generalized trust	4.231***	1.308**	1.526*	1.407**			
	(.718)	(.617)	(.849)	(.669)			
Trust in family					.418		
					(.485)		
Trust in neighbors						.295	
						(.311)	
Trust "people we know"							.176
							(.179)
Education		0.294***	0.302***	0.249***	0.307***	0.348***	0.359***
		(.034)	(.040)	(.047)	(.034)	(.034)	(.033)
Ethnic segmentation		−0.911**	−0.802*	−0.908**	−1.03***	−0.824**	−0.786*
		(.360)	(.404)	(.368)	(.351)	(.387)	(.396)
Population (ln)		−0.015	−0.024	0.037	0.018	0.060	0.057
		(.051)	(.506)	(.058)	(.046)	(.056)	(.054)
French LO			0.275				
			(.233)				
German LO			0.100				
			(.224)				
Scandinavian LO			0.007				
			(.367)				
Political institutions				0.0377			
				(.029)			
Observations	106	93	93	89	61	56	56
R^2	0.218	0.642	0.651	0.653	0.692	0.782	0.782

Notes: The dependent variable is *income per capita (ln)*, averaged over the years 1980–2009, taken from the Penn World Tables. Generalized *Trust* is calculated from answers to the question *"Generally speaking, would you say that most people can be trusted, or that you need to be very careful in dealing with people?"* Trust is equal to 1 if the respondent answers *"Most people can be trusted"* and 0 otherwise. Average trust in family, neighbors, and people you know, is calculated from the question *"Could you tell me for each whether you trust people from this group completely, somewhat, not very much or not at all?"* and the variable takes on the value 4, if the respondent answers *"Trust completely"*, 3 for *"Somewhat"*, 2 for *"Not very much,"* and 1 for *"No trust at all."*

Sample (106 countries): Albania, Algeria, Argentina, Armenia, Australia, Austria, Azerbaijan, Bangladesh, Belarus, Belgium, Benin, Bosnia and Herzegovina, Botswana, Brazil, Bulgaria, Burkina Faso, Canada, Cape Verde, Chile, China, Colombia, Croatia, Cyprus, Czech Republic, Denmark, Dominican Republic, Egypt, El Salvador, Estonia, Ethiopia, Finland, France, Georgia, Germany, Ghana, Great Britain, Greece, Guatemala, Hong Kong, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Japan, Jordan, Kenya, Kyrgyzstan, Latvia, Lesotho, Liberia, Lithuania, Luxembourg, Macedonia, Madagascar, Malawi, Malaysia, Mali, Malta, Mexico, Moldova, Montenegro, Morocco, Mozambique, Namibia, Netherlands, New Zealand, Nigeria, Norway, Pakistan, Peru, Philippines, Poland, Portugal, Puerto Rico, Romania, Russian Federation, Rwanda, Saudi Arabia, Senegal, Serbia, Singapore, Slovakia, Slovenia, South Africa, South Korea, Spain, Sweden, Switzerland, Taiwan, Tanzania, Thailand, Trinidad and Tobago, Turkey, Uganda, Ukraine, United States, Uruguay, Venezuela, Vietnam, Zambia, Zimbabwe.

Sources: The trust data comes from the five waves of the World Values Survey (1981–2008), the four waves of the European Values Survey (1981–2008), and the third wave of the Afrobarometer (2005). Education measures average years of schooling between 1950 and 2010 and is taken from Barro and Lee (2010). Ethnic fractionalization measures the degree of ethnic fractionalization and is taken from Alesina et al. (2003). Population is the average population (ln) between 1980 and 2009, taken from the Penn World Tables 7.0. Legal Origins are taken from La Porta et al. (2007). Political Institutions are measured by the Polity2 index averaged over 2000–2010, taken from the Polity IV database. OLS regressions with robust standard errors in parentheses.

*Coefficients are statistically different from 0 at the 10% level.

**Coefficients are statistically different from 0 at the 5% level.

***Coefficients are statistically different from 0 at the 1% level.

To compare the importance of generalized trust for income relative to other measures of trust, we run regressions replacing the measure of generalized trust by measures of limited trust, controlling for education, ethnic fractionalization, and population. As Table 2.3 makes clear, only generalized trust is significantly associated to income per capita. Limited trust (such as trust in family, neighbors, people one knows personally) is positively associated to income levels, but not significantly (columns 5–7). This result suggests that it is only the ability to cooperate outside the inner circle of family and relatives that is associated to economic performance, and is consistent with Banfield's analysis of the poor performance of Italian villages characterized by amoral familism. This result explains why the economic literature has made generalized trust the primary focus of analysis.

The same steady positive correlation between generalized trust and income per capita holds when we look at more local variations across regions in Europe or across states in the US. Figure 2.6 shows the correlation between generalized trust and average income per capita (ln) in 69 European regions using data taken from Tabellini (2010). Some European countries show a high degree of regional variation both in generalized trust and income per capita. In particular, northern Italy and northern Spain are high-trust regions and

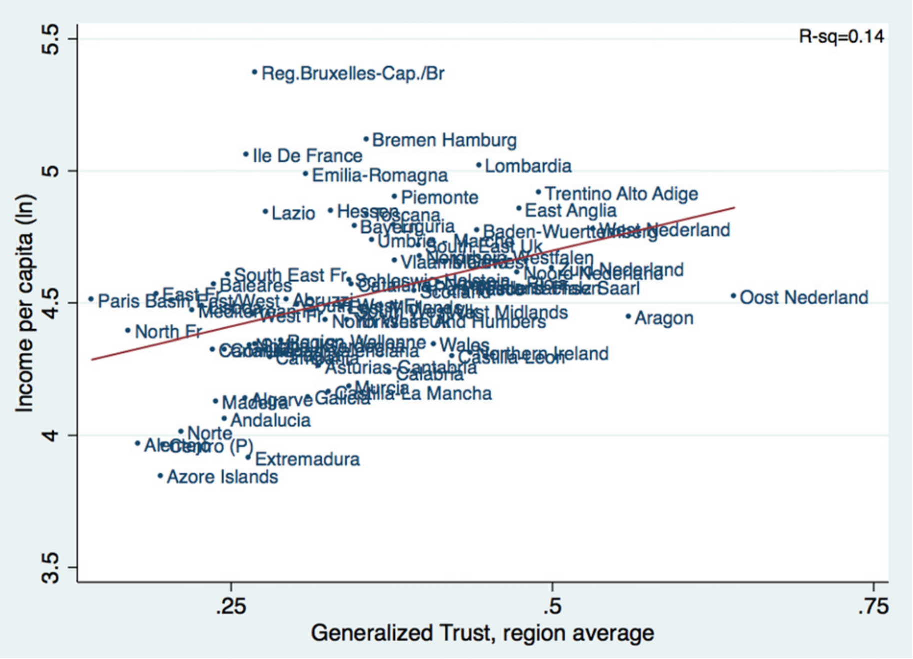


Figure 2.6 Income per capita (ln) and generalized trust in 69 European regions. *Source:* *Tabellini (2010). The trust measure is computed as the regional average from responses to the question "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" Trust is equal to 1 if the respondent answers "Most people can be trusted" and 0 otherwise.*

have high income per capita while southern Spain and southern Italy fare much worse on both dimensions. Figure 2.7 shows that the same positive correlation between trust and income per capita holds across US states. The southern states, in particular the former French colonies, have weak levels of trust and are also outperformed economically by the states of the north-eastern US.

Finally, using novel income data for more than 800 regions around the world collected by Gennaioli et al. (2013), we can observe that trust correlates with GDP at the region level around the world. Figure 2.8 displays the cross correlation of (ln) GDP per capita and trust for three different samples. Table 2.4 gives the associated regression output. Trust correlates positively with per capita income in 771 regions around the world, even stronger when the sample is restricted to regions belonging to groups of high income countries such as the EU27 (including Norway, but excluding Cyprus, Malta, and Luxembourg) and the OECD. Table 2.4 also displays regression results, when additionally education is controlled for. Since the number of individuals polled varies greatly between region,

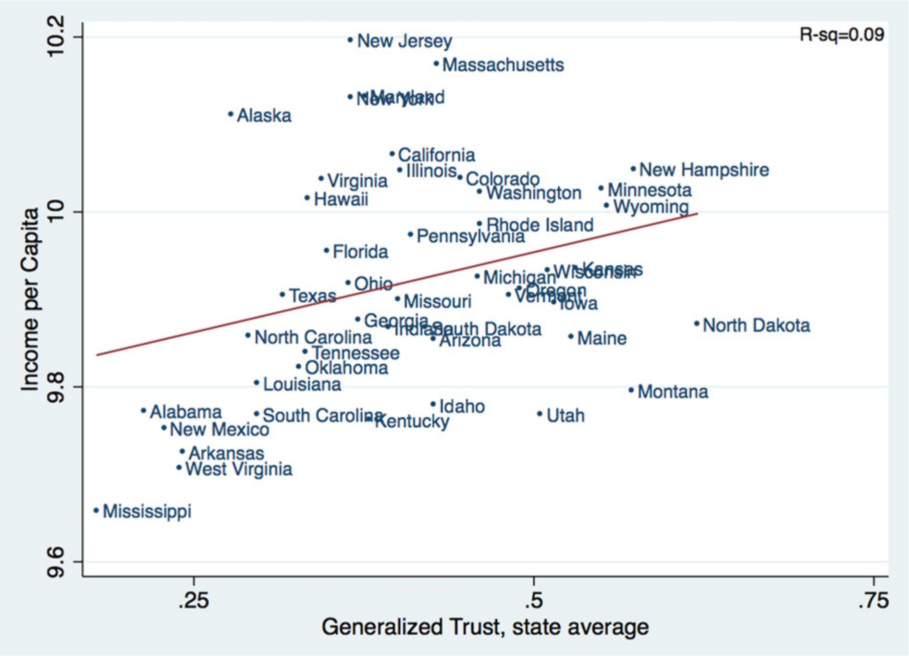
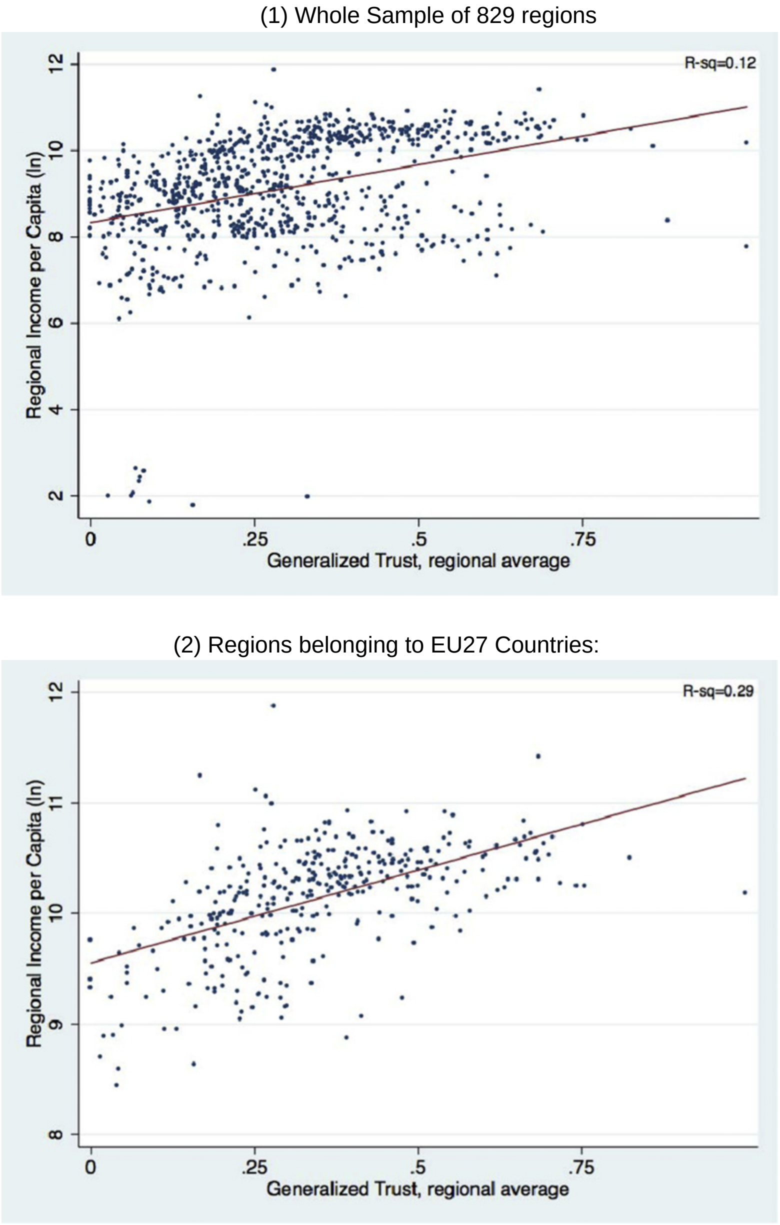


Figure 2.7 Income per capita (ln) and generalized trust in 49 US states. *Sources: Income data is taken from the US Census Bureau and averaged for the years 1972–2011. The proportion of people that trust is taken from the General Social Survey (1973–2006). The trust measure is computed as the state average from responses to the question "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" Trust is equal to 1 if the respondent answers "Most people can be trusted" and 0 otherwise.*



(1) Whole Sample of 829 regions

(2) Regions belonging to EU27 Countries:

Figure 2.8 Regional income per capita (ln) and trust in 829 regions around the world. *Sources: Income data is taken from the US Census Bureau and averaged for the years 1972–2011. The proportion of people that trust is taken from the General Social Survey (1973–2006). The trust measure is computed as the state average from responses to the question "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" Trust is equal to 1 if the respondent answers "Most people can be trusted" and 0 otherwise.*

(3) Regions belonging to the OECD

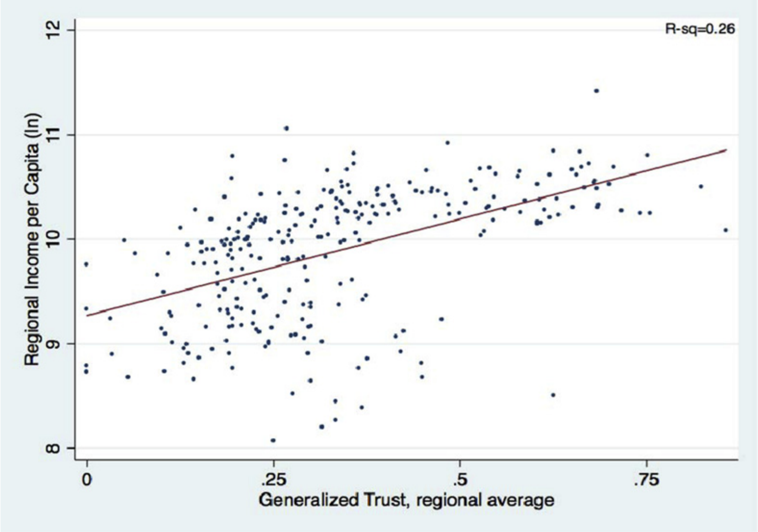


Figure 2.8 (*Continued*).

we account for this by running weighted regressions using precisely this number as our weight. No matter which sample is used, trust is positively and significantly associated with a higher regional per capita income across regions. However, as soon as we introduce country fixed effects, we do not observe any significant correlation between trust and GDP. This result shows that the cross-country heterogeneity in trust and income per capita is much more substantial than the within country variation, and drives the result.

Not only is trust positively correlated with income per capita, but also with growth. This point was first documented by Knack and Keefer (1997, 1999). Their study is based on 29 countries, mostly western European countries, between 1980 and 1992. Table 2.5 enlarges their result on the relation between trust and economic growth to cover 52 countries, regressing average annual growth between 1990 and 2009 on average trust between 1981 and 1990. We control for initial income and initial education. Trust is positively associated with economic growth. The correlation between trust and growth is statistically significant at the 10% level. A one standard deviation increase in trust, about 0.14, increases growth by 0.5% points or 20% of its sample mean. Column 2 controls for the initial level of investment and the correlation becomes statistically significant at the 5% level. Column 3 includes an interaction term between trust and initial income per capita. This interaction term captures the fact that trust should have a stronger effect on growth in poor countries that lack credit markets and appropriate rule of law. Both trust and trust interacted with initial income are statistically significant. The interaction term

Table 2.4 Trust and regional GDP per capita

	Ln GDP per capita					
	Full sample		EU		OECD	
	(1)	(2)	(3)	(4)	(5)	(6)
Trust	1.134**	0.313	1.345***	0.616	1.180***	0.867
	(0.497)	(0.211)	(0.369)	(0.719)	(0.341)	(0.625)
Education	0.306***	0.342***	0.113**	0.327***	0.080**	0.277**
	(0.030)	(0.031)	(0.053)	(0.106)	(0.033)	(0.110)
Country FE	No	Yes	No	Yes	No	Yes
Observations	771	771	278	278	350	350
R^2	0.603	0.964	0.321	0.834	0.298	0.755

Notes: The dependent variable is *ln GDP per capita*, which measures the log of regional income per capita, taken from Gennaioli et al. (2013).

Trust is calculated from answers to the question *"Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?"* Trust is equal to 1 if the respondent answers *"Most people can be trusted"* and 0 otherwise.

Sample: Columns (1) and (2) use the full sample of regions, as in Gennaioli et al. (2013). Columns (3) and (4) restrict the sample to regions belonging to a country being a member of the EU27 (including Norway, but excluding Malta, Cyprus, and Luxembourg). Columns (5) and (6) restrict the sample to regions belonging to a country being a member of the OECD.

OLS regressions with robust standard errors, clustered at the country level, in parentheses. All regressions are weighted by the number of individuals polled in each region.

Sources: The trust data comes from the five waves of the World Values Survey (1981–2008), the four waves of the European Values Survey (1981–2008), and all waves of the US GSS (1973–2006). Education measures the average years of schooling.

*Coefficients are statistically different from 0 at the 10% level.

**Coefficients are statistically different from 0 at the 5% level.

***Coefficients are statistically different from 0 at the 1% level.

is strongly negative, which provides support for the view that trust is more important when enforcement of formal institutions is weak.

2.5.2 Identification Issues

The previous section documents a strong correlation between trust and economic outcomes across countries or regions. However, how can we identify the causal impact of trust on economic performance? To answer this question, we must confront the various identification issues raised by the estimation of the following equation:

$$Y_c = a_0 + a_1 T_c + a_2 X_c + e_c, \tag{2.1}$$

where Y_c denotes economic performance in the geographic location c (country or region); T_c denotes trust; X_c is a vector of characteristics of the location, including the educational level of the population, current and past institutions, and past economic development in the locality; and e_c is an unobserved error term.

Table 2.5 Trust and growth: cross-country correlation

	Growth 1990–2009		
	(1)	**(2)**	**(3)**
Trust 1980–1990	0.0396*	0.0273**	0.480***
	(0.021)	(0.010)	(0.078)
Income p.c. 1990	−0.014***	−0.012***	0.002
	(0.003)	(0.002)	(0.002)
Education 1990	0.002**	0.001*	0.002***
	(0.001)	(0.001)	(0.001)
Investment		0.001***	
		(0.000)	
Trust × Income p.c. 1990			−0.048***
			(0.008)
Observations	52	52	52
R^2	0.491	0.658	0.706

Notes: The dependent variable measures average *GDP per capita growth* between 1990 and 2009, computed from Penn World Tables 7.0.

Trust is calculated from answers to the question *"Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?"* Trust is equal to 1 if the respondent answers *"Most people can be trusted"* and 0 otherwise.

OLS regressions with robust standard errors in parentheses.

Sample (52 countries): Albania, Argentina, Australia, Austria, Bangladesh, Belgium, Brazil, Bulgaria, Canada, Chile, China, Colombia, Croatia, Czech Republic, Denmark, Dominican Republic, El Salvador, Estonia, Finland, France, Germany, Great Britain, Hungary, Iceland, India, Ireland, Italy, Japan, Malta, Mexico, Netherlands, New Zealand, Norway, Pakistan, Peru, Philippines, Poland, Portugal, Romania, Russian Federation, Slovakia, Slovenia, South Africa, South Korea, Spain, Sweden, Switzerland, Taiwan, Turkey, United States, Uruguay, Venezuela.

Sources: The trust data comes from the waves 1–3 of the World Values Survey (1981–1995). Additional Controls: Income p.c. 1990 measures income per capita in 1990 (ln), Penn World Tables 7.0. Education 1990 measures average years of schooling in 1990, taken from Barro and Lee (2010).

*Coefficients are statistically different from 0 at the 1% level.

**Coefficients are statistically different from 0 at the 5% level.

***Coefficients are statistically different from 0 at the 10% level.

The identification of Equation (2.1) raises two main issues. The first is reverse causality: contemporaneous trust is likely to be influenced by the current state of economic development in locality c. The second issue is that of omitted variables that might co-determine both trust and economic performance. Specifically, institutions (Hall and Jones, 1999; Acemoglu et al. 2001; Rodrik, 1999), geography (Sachs, 2003), and more recently deep historical events (Nunn, 2009) and biology (Ashraf and Galor, 2013; Spolaore and Wacziarg, 2013), have been found to affect economic performance. However, as pointed out above, those factors also shape trust. In principle it might be possible to control for institutional quality, but such variables are well known to present difficulties of measurement, and in any case cannot capture informal norms. Worse, if Equation (2.1) is estimated in cross-section, it is impossible to include in the regression a fixed effect at the geographic location level c. This implies that trust and the unobserved error term can be correlated: cov (T_c, e_c) is different from zero and the OLS estimates of Equation (2.1) lead to biased estimates of the effect of trust. This opens up the possibility of a confounding factor: it is impossible to isolate the impact of trust from other time invariant characteristics of location c, such as other cultural values or local institutions. The most recent research in economic development precisely tries to find good strategies to control for any time invariant features at the local level. For instance, to measure the role of institutions in Africa, Michalopoulos and Papaioannou (2013) look at within-ethnic variation in economic development by controlling for ethnicity-fixed effects. They show that a very same ethnic group that belongs to different countries turns out to have similar contemporary income per capita, despite the institutional heterogeneity across countries. This result suggests that inherited traits specific to each ethnic group would explain much better economic development than institutions do.

In the following, we discuss the two main strategies proposed so far in the literature to address these identification issues to single out the causal impact of trust on economic development.

2.5.3 Identification Using Historical Events

A first strategy is to search for historical events as an exogenous variation in trust that could be used as instruments. To rationalize the use of historical events, the literature draws on the theory of the transmission of values. Studies by Bisin and Verdier (2001), Guiso et al. (2008b), and Tabellini (2010) stress the role of two main forces. A portion of current values is shaped by the contemporaneous environment (horizontal transmission of values), and another portion is shaped by beliefs inherited from earlier generations (vertical transmission of values). These theories suggest estimating the following equation for the formation of trust:

$$T_{ct} = b_0 + b_1 T_{ct-1} + b_2 X_{ct} + G_c + G_t + r_{ct}, \tag{2.2}$$

where contemporaneous trust T_c in locality c is explained by the initial trust present in the previous generation $T_{c,0}$, initial economic performance, and the initial and current other characteristics of the locality X_c. r_c is a random residual.

The two-step estimation of Equations (2.1) and (2.2) raises two main concerns. First, we do not have any information on initial trust $T_{c,0}$ since standardized cross-country databases on the level of trust present in earlier generations are not available. At best, it is possible to go back only to the 1980s to get a measure of trust in a cross-section of countries using the World Values Survey. Second, even if we could get a good proxy for initial trust $T_{c,0}$, the correlation between initial trust and contemporaneous economic outcomes may be interpreted as a causal effect from initial trust to contemporaneous outcomes only if these two variables are not codetermined by common factors.

Tabellini (2010) addresses these two issues in the following way. He estimates the causal impact of culture on regional economic development in Europe, where culture is broadly defined as moral values of good conduct, including trust. Importantly, Tabellini estimates the impact of trust within European countries, across regions. This means that it is possible to include country fixed effects in the vector X_c and control for national specificities. Tabellini uses two historical variables as an instrument for contemporaneous trust: past education and past political institutions. The political and social history of Europe ensures that these do vary widely at the regional level. He measures past education by the literacy rate around 1880, and early political institutions by constraints on executive power in the years 1600–1850. Tabellini shows in first-step estimates that contemporaneous trust is strongly correlated with these two instruments. Historically more backward regions, with higher illiteracy rates and worse political institutions, tend to have less generalized trust today. In the second step estimate, Tabellini shows that this historical variation in trust is strongly correlated with current regional development: regions with lower trust also have lower income per capita and lower growth rates, after controlling for country fixed effects, contemporaneous regional education, and past urbanization rates. The relationship is substantial: variation in trust could explain half of the observed income difference between Lombardy and southern Italy.

Tabellini's strategy is very insightful but raises two main concerns. The first one is how validly the instrument satisfies the exclusion restriction. The key assumption is that education and political institutions from the distant past do not directly affect contemporaneous output, after controlling for contemporaneous education and institutions. This assumption is likely to be violated. The literacy rate in the past is likely to have persistent effects on the formation of human capital, a key determinant of output. Similarly, there is much evidence that past institutions do have long-term effects on economic performance (Acemoglu et al. 2001). The second issue is linked to omitted variables. Since the author estimates cross-regional income per capita, he can control for country fixed effects. Thus, he can exclude that trust picks up time invariant characteristics at the country level. However, since the estimates draw on cross-sectional regressions at the regional level, it is impossible to include regional fixed effect in Equation (2.1). Thus, trust can pick up any other time invariant regional characteristics such as local geography or local formal and informal institutions.

Guiso et al. (2008a) follow a similar strategy to identify the impact of trust on income per capita in Italy. However, they look at more disaggregated historical variation in trust across cities within the same regions to exclude the influence of regional invariant characteristics. To estimate Equation (2.2) with historical variables, Guiso et al. revisit Putnam's conjecture that today's difference in trust between the north and the south of Italy is due to the history of independence that certain cities experienced in the north after the turn of the second millennium. They thus instrument today's trust (and more generally civic capital) with the past history of independence of certain cities. Additionally, they can exploit historical variation in the degree of independence of cities belonging to the same region: the communally governed cities were clustered in north central Italy, but not every city between the Apennine and the Alps experienced that form of regime. This strategy has one main advantage compared to Tabellini. Guiso et al. can estimate the impact of trust on output within the same region, across cities. This approach alleviates part of the concern that regional-invariant characteristics could determine both today's trust and income per capita. Guiso et al. find striking results. Northern cities that experienced independence and self-government in the Middle Ages now have 17% more non-profit associations than similar northern cities that never shared that experience. This higher level of social capital is associated with higher contemporaneous output: a one standard deviation increase in social capital increases income per capita by around 20%.

Still, as Guiso et al. stressed, their strategy cannot fully alleviate the identification concerns faced by Tabellini. First, the concern about the validity of the exclusion restriction for the instrument used for trust remains. One cannot exclude the possibility, that the historical shocks that affected cities at the turn of the millennium have a direct impact on income today. Having been a free city in the 13th century could have shaped other values or factors that exert long-lasting effects on economic outcomes. For example, free cities might have bred the spirit of entrepreneurship, or enhanced human capital. Second, trust can still pick up the effect of invariant local characteristics. Even if Guiso et al. identify the effect of trust within regions, they cannot control for geographic fixed effect at the city level.

This concern applies generally to all the literature that looks at the historical determinants of trust. As documented in Section 2.4, a burgeoning literature shows that trust is affected in the long-run by climate shocks, natural catastrophes, or history like the slave trade. But using those shocks as an instrument for trust in a growth equation is questionable. In particular, it is likely that climate shock or the slave trade affects growth by other channels than social capital, making the exclusion restriction disputable.

2.5.4 Time Varying Instruments: Inherited Trust and Growth

The historical approach leaves open the question of whether the level of trust does matter per se in explaining economic development, or whether it is not rather picking up the deeper influence of time invariant features such as legal origins, the quality of institutions, initial education, the extent of ethnic segmentation, and geography. What is needed is to find a measure for trust with time variation, allowing the investigator to control for time invariant specific factors. The difficulty in performing such an exercise is that there is no extended-time series on the evolution of trust.

Algan and Cahuc (2010) propose to use this time variation in inherited trust in the growth Equation (2.2). Since it is already well established that the parents' social capital is a good predictor of the social capital of children, they use the trust that US descendants have inherited from their forebears who immigrated from different countries at different dates to detect changes in inherited trust in the countries of origin (see Fernandez for a synthesis on the impact of culture on economic performance by using this epidemiological approach, 2011). For instance, by comparing Americans of Italian and German origin whose forebears migrated between 1950 and 1980, they can detect differences in trust inherited from these two source countries between 1950 and 1980. They can get time varying measures of trust inherited from these two countries by running the same exercise for forebears who immigrated in other periods, for instance between 1920 and 1950. With time varying measures of inherited trust, they can estimate the impact of changes in inherited trust on changes in income per capita in the countries of origin. This method allows us to address the main challenges mentioned above that arise in identifying the effect of trust on economic development. By focusing on the inherited component of trust, the authors avoid reverse causality. By providing a time varying measure of trust over long periods, they can control for both omitted time invariant factors and other observed time varying factors such as changes in the economic, political, cultural, and social environments.

More specifically, Algan and Cahuc re-estimate Equations (2.1) and (2.2) by allowing time variation in trust and economic performance, and including local fixed effects. We can rewrite the system of equations in the following way:

$$Y_{ct} = a_0 + a_1 T_{ct} + a_2 X_{ct} + F_c + F_t + e_{ct}, \tag{2.1$'$}$$

$$T_{ct} = b_0 + b_1 T_{ct-1} + b_2 X_{ct} + G_c + G_t + r_{ct}, \tag{2.2$'$}$$

where t is an index of the time period, and (F_c, G_c) and (F_t, G_t) denote country fixed effect and time effect, respectively. The authors thus estimate the impact of the variation in trust on the variation in income per capita within countries. In the benchmark estimation of the model, data availability led them to consider two periods: 1935–1938 and 2000–2003. More distant periods are also considered, but with fewer observations. The estimates are based on 24 countries from all over the world, including Anglophone countries, Continental European countries, Mediterranean European countries, Nordic countries, Eastern European countries, India, Mexico, and Africa.

To cope with the lack of information on trust of the previous generations in Equation (2.2′), the authors proxy the inherited trust of people living in country c by the trust that the descendants of US immigrants have inherited from their ancestors coming from country c. This yields an estimate of the term $b_1 T_{ct-1}$ in Equation (2.2′), which can be used as a proxy for inherited trust. This strategy leads to estimating a single equation of the form (2.1′), where T_{ct} is replaced by the proxy of inherited attitudes.

This strategy can address part of the identification issues discussed above. First, by using the trust US immigrants inherited from the home country instead of the average trust of the residents, we can exclude reverse causality. While trust in the home country has evolved according to what happened in that country, the inherited trust of US immigrants is only affected by shocks to the US economy. Besides, since we can have a direct measure of inherited trust, we do not have to worry about instruments that are unlikely to satisfy the exclusion restriction. Second, by looking at different waves of immigration, one can get time variation in inherited trust and thus include country fixed effects in Equation (2.1′).

The authors estimate the trust inherited by US immigrants from their home countries by using the General Social Survey. Inherited trust is measured as the country of origin fixed effect on individual regression of the generalized trust question, controlling for individual characteristics. The authors focus on inherited trust in the two periods 1935–1938 and 2000–2003 (1935 and 2000 henceforth) and impose a lag of 25 years between inherited trust and income per capita at time t. Therefore, inherited trust in 1935–1938 is that of second-generation Americans born before 1910 (i.e. whose parents certainly arrived one generation before 1935, a generation being defined as a 25-year period), of third-generation Americans born before 1935, and of fourth-generation Americans born before 1960. In the same way, the level of inherited trust in 2000–2003 corresponds to the trust inherited by: second-generation Americans born between 1910 and 1975; third-generation Americans born after 1935; and fourth-generation Americans born after 1960. This decomposition excludes any overlap in the inherited trust of the two groups.

The authors show that inherited trust for the period 2000 strongly correlates with trust in the home country during the same period, measured from the WVS. Additionally, the authors document substantial variation in inherited trust between 1935 and 2000. Swedish Americans have inherited higher trust in 2000 relative to the period 1935. Inherited trust from continental European countries, and to a lesser extent from the United Kingdom, has deteriorated over the period. Trust inherited in 2000 from French ancestors is 4.7% points lower relative to trust inherited from Sweden in 1935. Inherited trust has decreased even more among the immigrants from Eastern European countries and Mediterranean countries. The authors do not address the explanation for such variations—but there is a rich set of candidates. The ancestors of the current US respondents are likely to have undergone very different national crises. The ancestors who transmitted their trust for the period 1935 mainly migrated before World Wars One and Two. The level of trust of immigrants from countries deeply affected by these crises, like France, Germany, and Eastern European countries, might have deteriorated over the intervening period compared to descendants from Sweden, since this latter country is one of the European countries least affected by these traumatic mid-century events.

Algan and Cahuc (2010) then estimate the impact of change in inherited trust on changes in income per capita within country between 1935 and 2000. The estimates also control for changes in lagged income, political institutions, education, and other values (like work ethic or family values) over the period to isolate the specific effect of trust. The impact of inherited trust is substantial.

Figure 2.9 displays the change in income per capita in period 2000–2003 that countries would have experienced if the level of inherited trust in a given country had been the same as the trust inherited by Swedes. Income per capita in 2000 would have been increased by 546% in Africa (not reported) if the level of inherited trust had been the

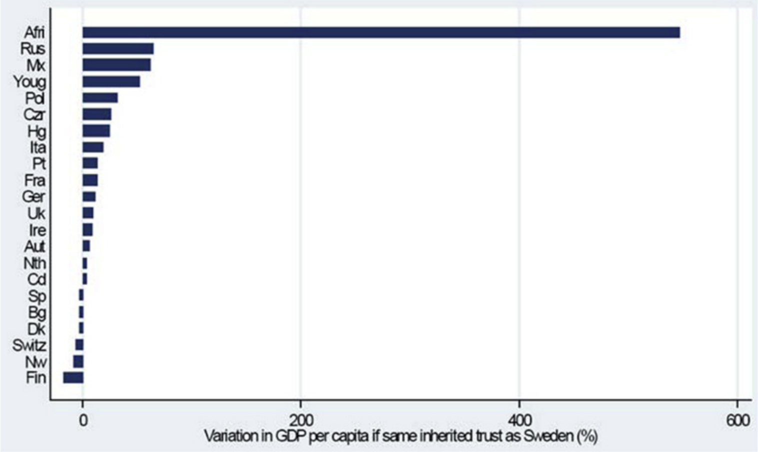


Figure 2.9 Predicted variation in GDP p.c. relative to Sweden. *Interpretation:* The figure shows the predicted variations in GDP per capita over the period 2000–2003 in a given country if it had the same level of inherited social attitudes as Sweden. *Source:* *Algan and Cahuc (2010)*.

same as inherited trust from Sweden. Inherited trust also has a non-negligible impact on GDP per capita in Eastern European countries, and Mexico. Income per capita would have increased by 69% in Russia, 59% in Mexico, 30% in Yugoslavia, 29% in the Czech Republic, and 9% in Hungary, had these countries inherited the same level of trust as Sweden. The effect, if less important, is also sizable in more developed countries. Income per capita would have been up by 17% in Italy, 11% in France, 7% in Germany, and 6% in the United Kingdom, if these countries had the same level of inherited trust as Sweden. The authors also compare the effect exerted by trust to the effect exerted by initial income per capita, or by time invariant factors such as geography, or by time invariant institutions. For poor countries from Africa or Latin America, initial economic development and invariant factors have a larger impact on income per capita. In striking contrast, change in income per capita within developed countries is overwhelmingly explained by inherited trust.

2.5.5 Individual Trust and Individual Economic Performance

Very few studies have explored whether high trusting individuals have higher economic performances in terms of wages or economic prospects. This is because of the difficulty of identifying the causal impact of individual trust on individual economic outcomes. Guiso et al. (2006) show, using the General Social Survey, that high-trusting individuals are more likely to become entrepreneurs in the US. To test for causality, they use inherited trust of US immigrants from their home country as an instrument for individual trust in the destination country. They find a significant, but somewhat too larger effect of inherited trust compared to the OLS estimates. As stressed by the authors, since inherited trust is time invariant, this variable may be picking up other inherited traits from the home country like risk aversion or saving behavior. This would explain the large difference in the OLS and 2SLS estimates. Ljunge (2012) draws on the same methodology by looking at how inherited trust of second-generation US immigrants is correlated with their economic success: second-generation immigrants with higher trusting ancestry earn significantly more than those with lower trust. They also have a higher labor supply, lower unemployment spell, and higher education. The correlation remains significant, even after controlling for additional ancestral influences such as income per capita and institutions. The paper cannot control for country of origin fixed effect though.

In another contribution, Butler et al. (2009) use the European Social Survey to test the relationship between individual trust and individual economic performance. The advantage of the ESS is to provide a question on generalized trust whose answers are scaled from 1 to 10, rather than just binary answers. The authors show that individual income is hump-shaped with the intensity of trust. Individuals whose level of trust is too high in relation to the civic-mindedness of their fellow citizens have levels of income inferior to those of individuals whose level of trust is intermediate. Being more frequently deceived by their fellow citizens hampers them. At the other extremity, individuals with little trust in others miss out on opportunities to make beneficial exchanges. Thus, there exists a "good" intermediate level of trust, the one that matches the level of civic-mindedness of the fellow citizens with whom one deals.

The conclusions drawn in this article might be limited by the quality of the ESS data. In these international values surveys, the measure of income levels is very imprecise and noisy. Nor do the questions about having been the victim of deceit focus on economic exchanges that might have a real impact on income, such as the interactions of professional life. But this article has the great merit of showing that the relationship between trust and economic performance is not necessarily monotonic. Trusting too much can have detrimental consequences. The recent financial crisis is a good illustration. The Icelanders, one of the most trusting peoples in international rankings, must still regret their excessive trust in their banks. Bernard Madoff's victims were likewise overly trusting.

If the analysis of the relationship between trust and economic performance at the individual level is to be advanced, the way ahead would seem to be field experiments, with an experimental measure of trust that measures behaviors precisely in economic exchanges and within firms. At the moment, the literature has done little to develop this approach. The only real study done on the terrain is that of Karlan (2005), who shows that, among Peruvian villagers, those most trusting in experimental games are also those who most often repaid their loans. But this study is not focused on the economic impact of trust on income. Some recent work heads in this direction but on limited samples. Barr and Serneels (2009) use a standard trust game to establish a relationship between experimental measures of reciprocating behavior among Ghanaian colleagues and the observed labor productivity of the firm in which they work. Similarly, Carpenter and Seki (2011) have Japanese fishermen play a repeated public goods game with and without an option for "social disapproval." They show that fishing crews that exhibit higher levels of reciprocity and more disapproval of shirking are more productive.

The way ahead in attempting to pin down the impact of trusting behavior on individual economic performance must be to combine the insights of experimental economics with experimentation—field, natural, and randomized. Doing so is a prerequisite if we are to better understand the channels through which trust affects economic performance and growth.

2.6. CHANNELS OF INFLUENCE OF TRUST ON ECONOMIC OUTCOMES

The empirical work presented in the previous section suggests that trust does indeed have an impact on growth. Macroeconomic in scope, this research is limited to the study of the relations obtaining among variables of a highly aggregated kind. It can therefore shed no more than a feeble light on the mechanisms or channels by which trust may act upon growth. Analyses more microeconomic in scope, focused on the relations obtaining among finance, insurance, the organization of firms, the labor market, public regulation, and trust, meet this need.

2.6.1 Financial Markets

In order to function, financial markets must rely heavily on trust, inasmuch as operations in these markets consist of promises of future payment which carry effect by reason of the fact that debtors are largely trustworthy, for legal protection would necessarily be costly and undependable. Figure 2.10 illustrates this positive relationship between trust and the development of financial markets in 86 countries over the course of the last three decades. As a gauge of the development of financial markets, we use the sum total of the credit granted by banks and financial institutions to private actors, as a percentage of GDP (see Levine, 2004).

Recent contributions to the literature have aimed at going beyond this positive correlation between trust and financial development, and pinpoint more closely the causal impact of trust. Guiso et al. (2004) study the relationship between the development of financial markets and trust in the regions of Italy in the 1980s and 1990s. They observe that households make more frequent use of cheques, keep a smaller portion of their savings in cash and a larger one in the stock market, and resort more frequently to credit-granting institutions, in the northern regions of the peninsula, where there is prevalent trust and high rates of blood donation and political participation. In the southern regions, moreover, borrowers resort more frequently to their families or near circles for loans than they do in the north.

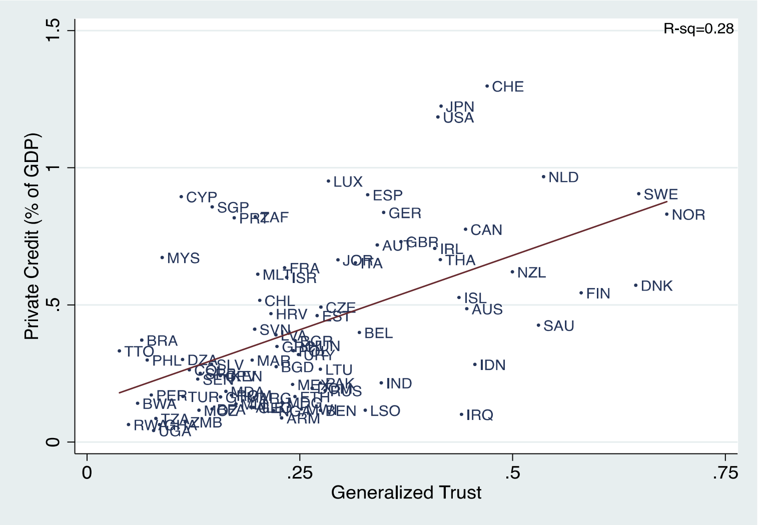


Figure 2.10 Financial development and generalized trust in 88 countries. *Sources: Financial development: Private credit by deposit banks and other financial institutions as a percentage of GDP, obtained from the World Bank Indicators (1980–2010). Generalized Trust is taken from the World Values Survey (1981–2008).*

As well as the composition of assets and volume of credit, trust can influence the propensity of investors to seek the counsel of financial intermediaries and delegate decisions to them. In a setting where financial products are complex, delegation to intermediaries who have a good knowledge of these products can ameliorate the diversification of investments and their rate of return. Guiso and Jappelli (2005) have shown that investors who have more trust in financial intermediaries delegate more decisions to them and thus obtain better-diversified and more efficient portfolios. The part played by trust in the propensity to turn to financial intermediaries capable of supplying products that will ameliorate risk coverage is replayed when it comes to insurance. Cole et al. (2013) have looked at the reasons why insurance contracts covering climate risks to their harvests in two rural regions of India were hesitantly received by locals, even though they bore a low cost. *A priori*, such contracts ought to have been attractive to households where variations in income are largely determined by the vagaries of precipitation during harvest season. Cole et al. show that lack of trust in and comprehension of the contracts explains a significant part of the refusal of households to take up this insurance. A randomized experiment shows that instructors who explain to folk the content of the contracts can have a significant influence on the take-up of this insurance, but only if they come recommended by a microcredit agency with a well-established reputation in the households. If so, the intervention of the instructors increases the uptake of the insurance by 36%.

If the instructor does not have this backing, or if the households are not acquainted with the institution backing him, his intervention has no significant impact.

Trust patently plays a part in situations of financial crisis. The GSS shows that trust in financial institutions declined steeply after the failure of Lehman Brothers in 2008 (Guiso, 2010). Such failures are themselves provoked by drops in confidence. Guiso observes that persons who had the least trust in their banks withdrew their savings earliest in periods of financial distress. And trust during these periods of financial distress is linked to trust prior to their onset. This observation suggests that a structural deficit of trust in financial intermediaries may favor the onset of financial crises.

The interpretation of the correlation between trust and finance is beset with difficulties. First, the correlation may result from other factors, like optimism or risk aversion, potentially linked to trust and exerting influence on the propensity to utilize financial products. Trust, however, is identified in the available research as a quite distinctive characteristic, different from risk aversion or optimism and exerting a specific effect on the utilization of financial products (Guiso et al. 2008a). Second, in the correlation between finance and trust, the causal sequence may run the other way: the quality of finance, itself linked to the quality of institutions, may explain trust. Guiso et al. (2004) show, however, that there does exist an inherited portion of trust, independent of environmental influence on the development of financial markets, and that it does influence the resort to financing. The authors observe that residents of northern Italy who arrived there from regions in the south characterized by weak trust and weak civic spirit view financial products more distrustfully than do those born in the north. On identical observable characteristics, moreover, they get fewer loans from financial institutions. Such influence exerted by region of birth suggests that trust, and civic spirit as well, constitute partly heritable traits that may act as obstacles to the development of finance.

2.6.2 Innovations and Firm Organization

2.6.2.1 Innovations

Trust must play a preponderant role in the sort of economic activities—investment and especially innovation—that are attended by uncertainty on account of moral hazard and the difficulties of contract enforcement. In their path-breaking article on the link between trust and growth, Knack and Keefer (1997) already threw into relief a positive correlation between trust and investment as percentage of GDP. The correlation should be even more significant for research and development, and factor productivity.

Figure 2.11a documents the steady positive correlation between trust and a measure of total factor productivity, taken from Hall and Jones (2009), for a sample of 62 countries. Around one-third of the cross-country variation in TFP is associated to differences in trust across countries. Figure 2.11b illustrates the positive cross-country variation between average trust and innovation in 93 countries, with innovation measured by expenditure on research and development as percentage of GDP. The countries where trust is highest are

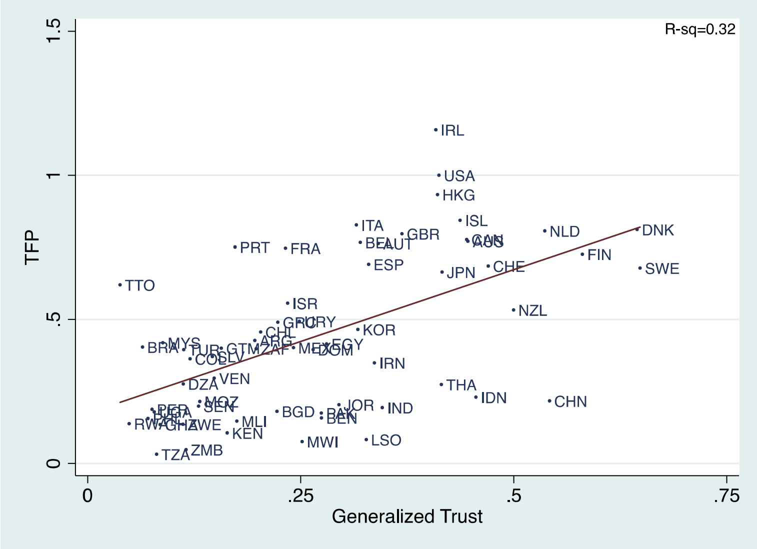


Figure 2.11a Total factor productivity and generalized trust in 62 countries. *Sources: Total Factor Productivity is taken from Hall and Jones (1999). Trust is measured from the World Values Survey (1981–2008).*

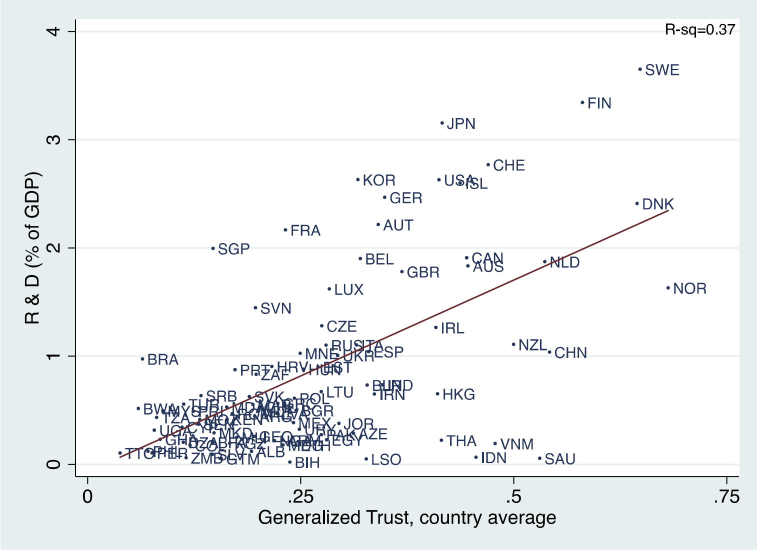


Figure 2.11b R&D expenses and generalized trust. *Sources: R&D expenses as a percentage of GDP over the period 1980–2010 are taken from the World Bank Development Indicators. Trust is measured from the World Values Survey (1981–2008).*

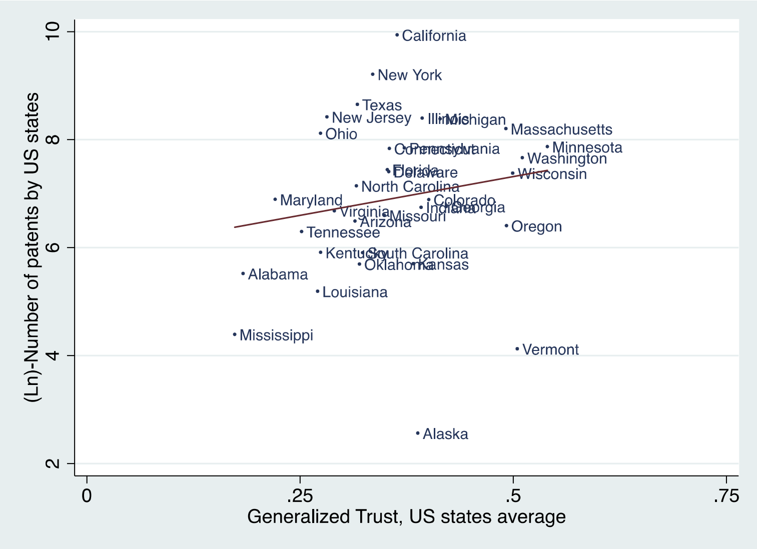


Figure 2.11c Cross US states correlation between R&D ((ln)-number of patents over the period 1980–2010) and generalized trust (1976–2008). *Sources: Income data is taken from the US Census Bureau and averaged for the years 1972–2011. The proportion of people that trust is taken from the General Social Survey (1973–2006). The trust measure is computed as the state average from responses to the question "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" Trust is equal to 1 if the respondent answers "Most people can be trusted" and 0 otherwise.*

the ones with elevated R&D, in point of fact, the Anglophone and Nordic countries. Trust on its own explains more than a third (37%) of the dispersion of rates of expenditure on R&D across countries. This relationship remains statistically significant at the 5% level after controlling for initial income per capita, population density, and education. Figure 2.11c shows that the same correlation between innovation and trust holds across US states, whereby innovation is measured by the (ln)-number of patents per state. Remarkably, we find that this relationship also remains statistically significant at the 1% level after controlling for income per capita, population density, and the share of the population holding a PhD at the state level. The relationship between trust and innovation operates through a specific channel different from education or population density.

While the correlation between innovation and trust appears strong, we have, as yet, few studies that attempt to pin down the direction of the causality. The literature gives much greater prominence to another mechanism influencing innovation—the organization of firms and especially their degree of decentralization.

2.6.2.2 *Firm Organization*

By facilitating cooperation among anonymous persons, trust favors the emergence and growth of private and public organizations (Fukuyama, 1995; La Porta et al. 1997; Bertrand

and Schoar, 2006). Trust favors the decentralization of decisions within organizations, allowing them to adapt better to alterations in the environment.

Figure 2.12 documents this relationship by showing a positive correlation between firm decentralization and generalized trust for 72 countries. Firm decentralization is measured by the following question from the Global Competitiveness Report 2009 (GCR): "In your country, how do you assess the willingness to delegate authority to subordinates? 1 = low: top management controls all important decisions; 7 = high: authority is mostly delegated to business unit heads and other lower-level managers." Generalized trust is measured as the country average from WVS 1981–2009. The positive relationship is substantial: 37% of the cross-country variation in firm decentralization is associated with country differences in trust.

This aspect of trust is illustrated by Cingano and Pinotti (2012) who find that trust is associated with greater decentralization and larger firm size across Italian regions. Exploiting industry variation (and controlling for region- and industry-specific factors) they show that high-trust regions exhibit a larger share of value added and exports in industries characterized by greater need-for-delegation. The effect is driven by a shift of the firm size distribution away from the smallest units toward firms in higher size classes. Their estimated relationships are not only statistically significant but also economically meaningful when compared to such other determinants of industry specialization and

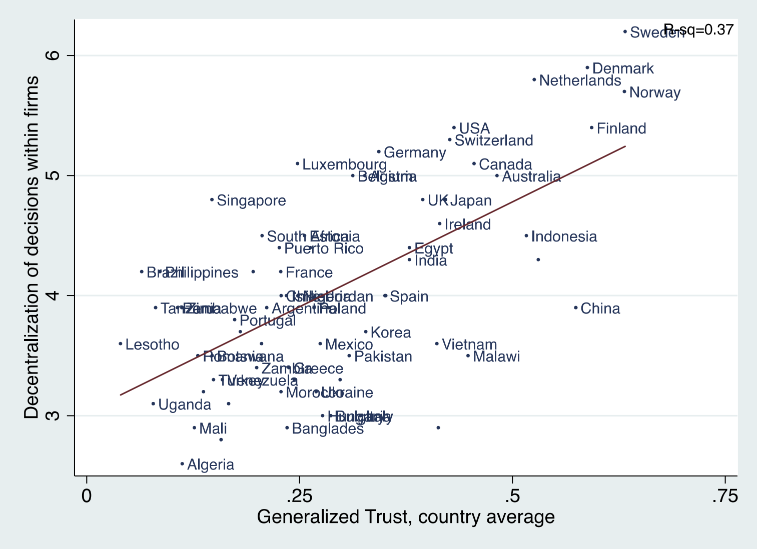


Figure 2.12 Cross-country correlation between decentralization of firms and trust. *Sources: Firm decentralization is measured by the following question from the Global Competitiveness Report 2009 (GCR): "In your country, how do you assess the willingness to delegate authority to subordinates?" Answers range from "1 = low: top management controls all important decisions, " to "7 = high: authority is mostly delegated to business unit heads and other lower-level managers." Generalized trust is measured as the country average from WVS 1981–2009.*

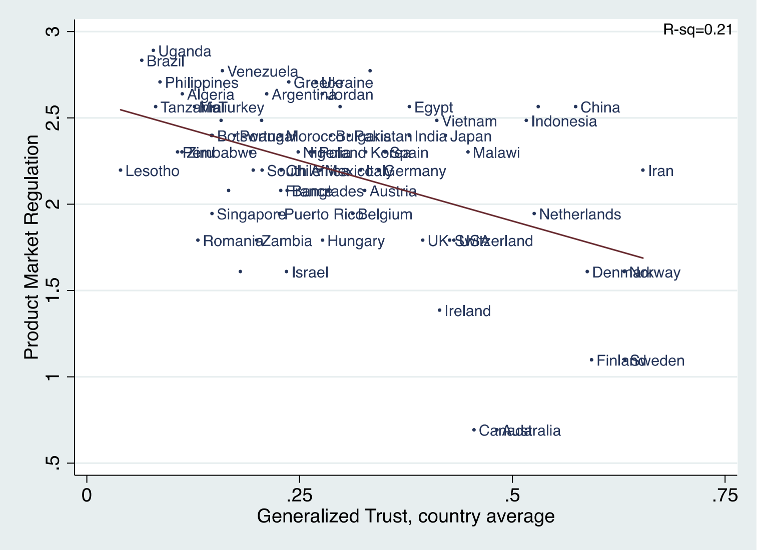


Figure 2.13 Product market regulation and trust in 73 countries. *Sources: Product market regulation is measured as the (ln)-number of steps for opening a business, taken from the World Bank (2009). Generalized trust is measured as the country average from WVS 1981–2009.*

firm organization as human capital, physical capital, or judicial quality. For example, they imply that increasing trust by an amount corresponding to the inter-quartile range of its distribution across Italian regions would raise value added in a delegation-intensive industry (such as manufacture of machinery and equipment) relative to a less intensive industry (such as leather, leather products and footwear) by 24% (or by 19%, when using cross-country data). This amounts to around two-thirds of the implied effect of raising human capital, and is larger than the effect of physical capital or contract enforcement.

In the same vein, Bloom et al. (2012) show that trust can improve aggregate productivity by facilitating firm decentralization. They first provide a model supplying a rational foundation for the correlation between trust and decentralization of firms. Following Aghion and Tirole (1997) in their analysis of the congruence of preferences between CEOs and managers, the authors posit two opposite ways of organizing production. The CEO can either solve production problems directly or delegate these decisions to plant managers. When trust is high, plant managers tend to solve problems in congruence with the CEO's expectations rather than exploiting resources for their own interest. The CEO is thus more likely to delegate. In this perspective, trust affects the economic performance of firms through two channels. First, greater trust within the firm improves performance thanks to decentralized decision-making. A low-trust environment is a hindrance to the growth of the most productive firms. Second, economies characterized by low trust may orient themselves toward sectors in which decentralized decision making is less imperative. Sectors close to the leading edge of technology such as IT have to grant space for individual decision-making in order to innovate and constantly adapt to the environment. Bloom et al. (2012) test these predictions empirically. They collect new data on the decentralization of investment, hiring, production, and sales decisions from corporate headquarters to local plant managers in almost 4000 firms in the United States, Europe, and Asia. They find substantial differences in the cross-country decentralization of firms: those in the United States and northern Europe appear to be the most decentralized and those in southern Europe and Asia the most centralized. The authors match their database on management practices with the level of trust where the headquarters are located, using regional information from the WVS. They find that firms headquartered in high-trust regions are significantly more likely to decentralize. To identify the causal impact of trust on decentralization, they examine multinational firms and show that higher levels of bilateral trust between the multinational's country of origin and subsidiary's country of location increases decentralization. Finally, the authors show that more decentralized firms are also more productive and tend to specialize in innovation and information technology. Trust, indispensable for the decentralization of firms, thus affects innovation and aggregate productivity.

2.6.3 The Labor Market

Trust likewise exerts influence on the functioning of the labor market, through several channels affecting growth.

2.6.3.1 The Quality of Labor Relations

Countries with higher generalized trust also have higher levels of cooperative relations between labor and management and higher levels of unionization. Unions have more members when generalized trust is high. Opportunistic and non-cooperative behavior constitutes a significant barrier to joining a union (Olson, 1965). Mutual trust and cooperation make it possible to lift these barriers. Cross-country analyses also show that relations between employers and employees are more cooperative when unions are more powerful (Aghion et al. 2011). The quality of employer-employee relations is associated to an array of factors that favor growth. The first is low unemployment (Blanchard and Philippon, 2004). Next, firms that have unions representing their employees are better able to adapt to new management methods, have more cooperative labor relations, and better productivity (Black and Lynch, 2001). Unions can ameliorate the quality of labor relations by allowing wage-earners to voice their views rather than be forced to stark either/or alternatives. Conceived this way, the role played by unions recalls Tocqueville's account of associations as little social laboratories where persons might learn cooperation first hand. It has been noted that farmers are more careful to use water sparingly the more they have had a voice in the framing of the irrigation regulations. Communes and cantons where political democracy is most strongly rooted, with high rates of voter turnout, have the lowest levels of tax evasion (Frey, 1998). Laboratory experiments confirm this observation, as shown in the next section. Players who decide on the rules governing their cooperation are more generous and trusting than those upon whom the same rules are imposed by an outsider. In other words, regulation and policy have a better chance of favoring cooperation to the extent they have been decided by a shared resolution and not imposed (Ostrom, 1990).

Hence the reaction of governments when there is a failure of the union-management dialog, the social dialog as it is called in Europe, can make it worse. Aghion et al. (2011) show that state regulation of labor markets is negatively correlated with the quality of labor relations. They argue that these facts reflect different ways of regulating labor markets, either through the state or through the civil society, depending on the degree of cooperation in the economy. They rationalize these facts with a learning model of the quality of labor relations. Distrustful labor relations lead to low unionization and high demand for direct state regulation of wages. In turn, state regulation crowds out the possibility for workers to experiment with negotiation and grasp the possibilities of cooperation in labor relations. This crowding out effect can give rise to multiple equilibria: a "good" equilibrium characterized by cooperative labor relations and high union density, leading to low state regulation, high employment, and production; and a "bad" equilibrium, characterized by distrustful labor relations, low union density, and strong state regulation of the minimum wage.

2.6.3.2 Flexicurity

The countries of southern Europe have chosen to offset the shocks that affect all working lives by prioritizing employment through rigorous employment protection, rather than prioritizing individuals through a generous unemployment benefit and an effective public agency to help in the job search. Conversely, the countries of northern Europe have adopted a "flexicurity" model that combines generous unemployment benefit, effective public job search agencies, and weak employment protection. Flexicurity is associated to better labor market performance, with higher rates of employment and a better reallocation of jobs toward more productive enterprises. On this basis, international institutions like the OECD and the European Commission recommend the adoption of flexicurity. Yet this model has a low rate of take-up outside northern Europe. Algan and Cahuc (2009) show that a trust deficit can create a barrier to the adoption of flexicurity. They provide evidence of cross-country correlations between national civic attitudes and the design of labor market insurance. Countries displaying high trust tend to insure their workers through unemployment benefits instead of using stringent employment protection. Such a relationship is robust to the inclusion of country fixed effects which account for time invariant national features and which could affect the design of unemployment insurance and employment protection. This finding is consistent with the strongly marked contrast between the flexicurity model in Nordic countries such as Denmark, and the continental European and Mediterranean countries. Naturally, the correlation between civic attitudes and the design of labor market institutions does not mean that there is a straight causal relationship going from social attitudes to the unemployment benefits/employment protection trade-off. There is a potential for reverse causality, since labor market institutions are likely to affect civic attitudes. For instance, administrative inefficiencies in the provision of unemployment insurance could influence guilty feelings about cheating on unemployment benefits. To deal with this reverse causality issue, Algan and Cahuc (2009) estimate the inherited part of civic attitudes that are not instantaneously influenced by the economic and institutional environment of the country in which people are living, by estimating the civic attitudes inherited by the American-born from their ancestors' country of origin, using the General Social Survey database. Using this inherited part of civic attitudes by country of origin as an instrument for civic attitudes in the home country, the authors show that there is a significant impact of civic attitudes on unemployment benefits and on employment protection in OECD countries during the period 1980–2003.

2.7. INSTITUTIONS, POLICIES, AND TRUST

2.7.1 Can Trust be Changed? Putnam I versus Putnam II

If trust plays a key role in explaining economic outcomes, it becomes urgent to identify the institutions and public policies for it to develop. Research related to this subject is still in its early stages. As discussed in Section 2.4.3, a large part of the literature considers trust to be a cultural component hardly malleable, whose determinants have to be searched for in the long history of each country, and with little room for immediate action. Yet, recent studies looking at immigrants show that their level of trust converge gradually to the average level of trust in their country of destination.

This ambiguity is well illustrated by the two conflicting views of the evolution of trust given by Putnam in his two books dating from 1993 and 2000. According to Putnam I (see the book from Putnam et al. 1993), social capital is largely determined by history. Elevated levels of social capital in the regions of north Italy compared to those in the south originated in the free-city experience during the medieval era.

Contrarily, according to Putnam II (see Putnam's book *Bowling Alone* in 2000), trust evolves quickly and is strongly influenced by the environment. In his book *Bowling Alone* Putnam shows that the levels of social capital, as measured by associations and club membership, have starkly declined in the United States since World War II. One of his main explications of this decline is the individualization of leisure activities, with an increasing amount of time spent watching television. Olken (2009) also identifies a negative impact of television and radio on association membership and self-reported trust in Indonesia by using variation in Indonesia's mountainous terrain and differential introduction of private television.

Depending on which perspective we take, from Putnam I or Putnam II, the room for policy intervention would be rather small or large. Section 2.4.3 documents that both approaches have an element of truth. Trust is partly inherited from past generations and shaped by historical shocks, because the underlying beliefs regarding the benefits of trust and cooperation are transmitted in communities through families (Bisin and Verdier, 2001; Benabou and Tirole, 2006; Tabellini, 2008b; Guiso et al. 2008a). But another part of trust is shaped by personal experience from the current environment, let it be social, economic, and political. In Bisin and Verdier's terminology, both the vertical channel of transmission from parents and the oblique/horizontal channel from the contemporaneous environment are at play in the fabric of trust.

This debate on the adjustment of trust to its environment also depends on what generalized trust really measures. If trust consists of beliefs about the trustworthiness of others, it is likely that individuals can update upward or downward their beliefs depending on the environment where they live, the civic spirit of their fellow citizens, and the transparency of their institutions. If trust consists of ingrained preferences and moral values, transmitted in early childhood and disconnected from personal experience as suggested by Uslaner (2008) and others, it might take more time to adjust. In the latter case, the action steps necessary to increase trust differ and depend on long-term policy, such as education. In this section, we consider the various policies that can shape both contextual beliefs and deeper preferences.

2.7.2 Institutions and Trust

How can institutions, and which institutions, shape trust? Do formal rules and norms embedded in institutions act as a complement or a substitute for informal values such as trust? These questions are key to identifying how and which specific institution could build up trust.

2.7.2.1 Relation Between Trust and Institutions

Figure 2.14 shows a strong positive correlation between trust and the quality of the legal system for a sample of 100 countries. Figure 2.15 displays a similar correlation between trust and the quality of governance in 163 European regions. These correlations are robust to using different measures of institutional quality commonly used in the economic literature (see Tables 2.6a and 2.6b), such as the rule of law, the strength of property right protection, the enforcement of contracts; as well as government effectivity, accountability, corruption (Rothstein and Uslaner, 2005) and controlling for other influences of institutional quality.

Recent papers try to go beyond this correlation by showing a causal impact of legal enforcement on trust. Tabellini (2008b) provides suggestive evidence that generalized morality is more widespread in European regions that used to be ruled by non-despotic political institutions in the distant past. Using data from the General Social Survey,

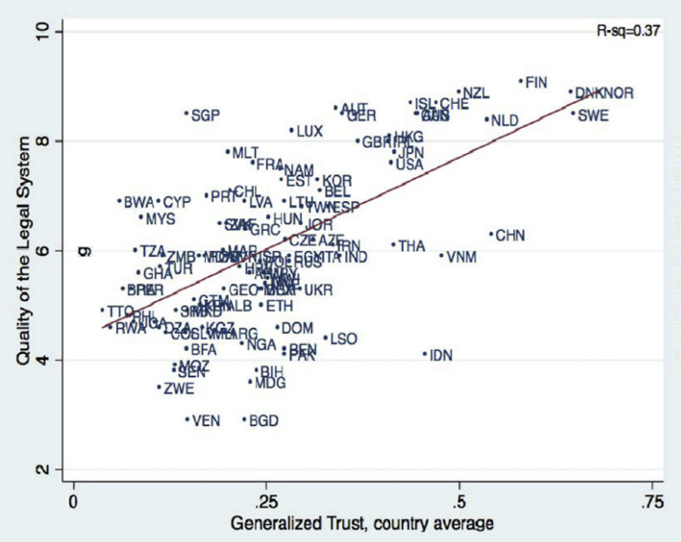


Figure 2.14 Quality of the legal system and trust in 100 countries. *Sources: The Quality of the Legal System is taken from the Economic Freedom of the World Index (2007). Generalized trust is measured as the country average from WVS (1981–2009) and EVS (1981–2008).*

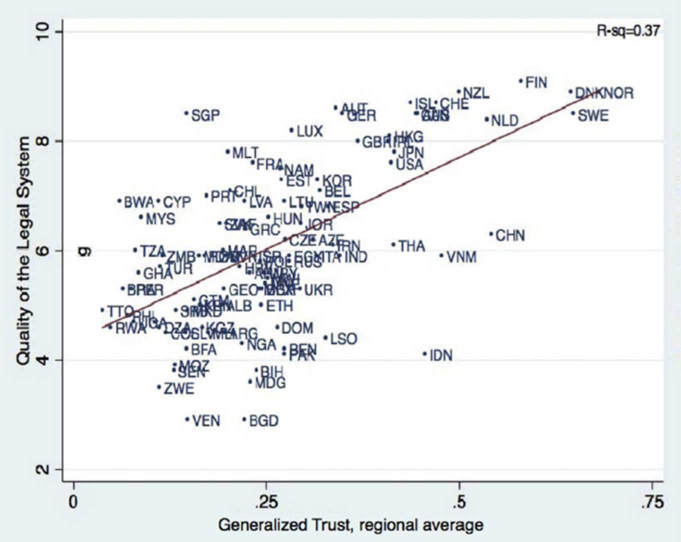


Figure 2.15 Quality of governance and generalized trust in 163 European regions. *Sources: The Quality of Governance is taken from the Quality of Government Index (2010). Generalized trust is measured as the country average from the WVS (1981–2009) and EVS (1981–2008).*

Table 2.6a Trust and institutions

Cross-country correlation

	Quality of legal system (1)	Rule of law (2)	Property rights (3)	Enforcement of contracts (4)
Trust	3.942***	1.271**	1.604***	2.864***
	(0.719)	(0.484)	(0.602)	(0.674)
Income per capita	0.646***	0.420***	0.531***	0.930***
	(0.126)	(0.0891)	(0.101)	(0.250)
Population	−0.167***	−0.109***	−0.195***	−0.0284
	(0.055)	(0.035)	(0.050)	(0.092)
Education	0.0146	0.0558	0.0120	0.178**
	(0.053)	(0.047)	(0.052)	(0.087)
Ethnic segmentation	0.152	−0.242	0.0572	1.614***
	(0.440)	(0.251)	(0.377)	(0.535)
Observations	90	93	91	46
R^2	0.684	0.681	0.589	0.807

Notes: Dependent variables: (1) *Quality of Legal System* measures the overall quality of the legal system, taken from Economic Freedom of the World Index, 2007. (2) *Rule of Law* gives the average rule of law between 1996–2010, taken from Kaufmann et al. (2010). (3) *Property Rights* are a measure of property rights taken from the Heritage Foundation, 2004. (4) *Enforcement* measures enforceability of contracts, taken from Botero et al. (2004).

Trust is measured from the answer to the question *"Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?"* Trust is equal to 1 if the respondent answers *"Most people can be trusted"* and 0 otherwise.

OLS regressions with robust standard errors in parentheses.

Sample (93 countries): Albania, Algeria, Argentina, Australia, Austria, Bangladesh, Belgium, Benin, Botswana, Brazil, Bulgaria, Canada, Chile, China, Colombia, Croatia, Cyprus, Czech Republic, Denmark, Dominican Republic, Egypt, El Salvador, Estonia, Finland, France, Germany, Ghana, Great Britain, Greece, Guatemala, Hong Kong, Iceland, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Japan, Jordan, Kenya, Kyrgyzstan, Latvia, Lesotho, Liberia, Lithuania, Luxembourg, Malawi, Malaysia, Mali, Malta, Mexico, Moldova, Morocco, Mozambique, Namibia, Netherlands, New Zealand, Norway, Pakistan, Peru, Philippines, Poland, Portugal, Romania, Russian Federation, Rwanda, Saudi Arabia, Senegal, Singapore, Slovakia, Slovenia, South Africa, South Korea, Spain, Sweden, Switzerland, Taiwan, Tanzania, Thailand, Trinidad and Tobago, Turkey, Uganda, Ukraine, United States, Uruguay, Venezuela, Vietnam, Zambia, Zimbabwe.

Sources: The trust data comes from the five wave of the World Values Survey (1981–2008), the four waves of the European Values Survey (1981–2008), and the third wave of the Afrobarometer (2005). Additional Controls: Investment Share measures Investment % of GDP 1980–2009, Penn World Tables 7.0. Income per capita measures GDP per capita (ln), const. prices, averaged for the years 1980–2009, taken from the Penn World Tables 7.0. Population measures population (ln), averaged between 1980 and 2009, Penn World Tables 7.0.

*Coefficients are statistically different from 0 at the 10% level.

**Coefficients are statistically different from 0 at the 5% level.

***Coefficients are statistically different from 0 at the 1%, level.

Tabellini regresses individual trust of US immigrants on various indicators of legal enforcement at stake in their ancestor's country at the end of the 19th century. He finds that immigrants from countries with more democratic institutions in the distant past have inherited a higher level of trust, even when controlling for historical economic development and school enrollment in the home country.

Table 2.6b Trust and institutions

Cross-regional correlation in Europe

	Quality of governance (1)	**Quality of governance (2)**	**Rule of law (3)**	**Effectivity (4)**	**Accountability (5)**
Trust	4.376***	1.291**	3.285***	5.423***	2.463*
	(0.924)	(0.559)	(0.736)	(1.356)	(1.222)
Population		−0.263*	0.05	−0.253	−0.160
		(0.147)	(0.120)	(0.270)	(0.103)
Ln GDP p.c.		0.932***	0.487**	0.684	1.039***
		(0.191)	(0.222)	(0.583)	(0.220)
Education		0.03	−0.029**	0.0246	−0.0127
		(0.027)	(0.011)	(0.043)	(0.021)
Autonomous		−0.267	0.275**	0.0685	0.477***
		(0.164)	(0.105)	(0.334)	(0.147)
Bilingual		−0.0513	0.0791	1.207**	−0.32
		(0.198)	(0.199)	(0.556)	(0.184)
Area		0.216**	−0.0351	0.134	0.227
		(0.087)	(0.073)	(0.187)	(0.131)
Observations	163	163	163	163	163
R^2	0.342	0.613	0.499	0.450	0.552

Notes: Dependent variables: Columns (1) and (2): *Quality of Governance* index measures the overall quality of regional institutions, taken from the Quality of Governance Institute, 2010. (3) *Rule of Law* measures the quality of the rule of law, taken from the Quality of Governance Institute, 2010. (4) *Effectivity* measures the governance effectivity, taken from the Quality of Governance Institute, 2010. (4) *Accountability* measures the quality of media and elections, taken from the Quality of Governance Institute, 2010.

Trust is measured from the answer to the question *"Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?"* Trust is equal to 1 if the respondent answers *"Most people can be trusted"* and 0 otherwise.

OLS regressions with robust standard errors, clustered at the country level, in parentheses.

Sample (163 regions): 163 European regions in the following countries: Austria, Belgium, Bulgaria, Czech Republic, Denmark, France, Germany, Greece, Hungary, Italy, Netherlands, Poland, Portugal, Romania, Slovakia, Spain, Sweden, United Kingdom.

Sources: The trust data is taken from the four waves of the European Values Survey (1981–2008. Population measures the log of the average number of inhabitants 2007–2009 per region, taken from Eurostat. GDP p.c. gives the log of the regional average GDP per capita between 2007 and 2009, taken from Eurostat. Education gives the percentage of population with some type of tertiary degree in 2006, taken from Eurostat. Bilingual equals to 1 if more than one official languages exists in the region. Autonomous equals 1 if the region is an autonomous region. Logarea gives the log value of the region's area.

*Coefficients are statistically different from 0 at the 10% level.

**Coefficients are statistically different from 0 at the 5% level.

***Coefficients are statistically different from 0 at the 1% level.

Naturally, this approach does not prove that past democratic institutions have a causal impact on trust. Since those institutions are invariant, they could pick up any other invariant aspect of the home country. Yet, Tabellini's analyses are intriguing since historical political institutions could explain up to 57% of the country of origin fixed effect. This share is much larger than the one explained by income per capita and education in the distant past. Institutions can have long-lasting impact on social and economic outcomes, but the persistence channel goes through their effect on values. This is really different from the traditional explanation of the persistence of institutions through elites capture (Acemoglu et al. 2001). Weak legal enforcement forces citizens to rely on informal and local rules and to develop limited trust as opposed to generalized trust. A good illustration of this diffusion of limited morality in the presence of weak institution is given by the Mafia. Gambetta (1993) documents that feudalism was formally abolished in Sicily much later than in the rest of Europe (in 1812). The State was too weak to enforce the introduction of private property rights of the lands. The Mafia benefited from this institutional vacuum and offered local protection through informal patronage, drawing a clear distinction between those under its protection and the others. In the same vein, Section 2.4.2 above has documented recent studies showing that non-democratic and corrupt institutions in the distant past in Italy or in the Habsburg Empire are related to lower trust nowadays.

Other contributions use natural experiments to show the effect of democratic institutions on cooperative behavior. Bardhan (2000) finds that farmers are less likely to violate irrigation rules when they themselves have set up those rules. Frey (1998) shows that tax evasion in Swiss cantons is lower when democratic participation is greater. All these different works are suggestive of an impact of democracy on cooperation. But even those latter natural experiments cannot rule out the existence of omitted factors determining both the selection of institutions and the response to institutions. Besides, the precise mechanism through which democracy (and more generally, formal rules) shapes cooperative behavior and the identification of its effect still needs more research (see Benabou and Tirole for a theoretical model that rationalizes the interplay between laws and norms, 2011).

2.7.2.2 Experimental Games

An alternative approach for identifying the effect of institutions on cooperation is to mimic formal and legal rules in the context of experimental games. Naturally, formal and legal rules in experimental games differ from real institutions. But this has the advantage of providing a controlled experiment to estimate how people change their level of cooperation and trust depending on exogenous variations in the rules of the games.

Initially, the literature has looked at the interaction between formal and informal institutions, but in the context of cooperation with reputational incentives, such as repeated games (Kranton, 1996). One main conclusion of this approach is that legal enforcement can crowd out reputational incentives and undermine informal institutions. Yet, this prediction seems to be very specific to situations of cooperation with reputational incentives, and do not apply to cooperation embedded in moral values such as generalized trust.

Fehr and Gatcher (2000) analyze cooperation in a public good game. Interestingly, the authors changed the setup of the traditional public good experiment by allowing the cooperators to punish the defectors. They demonstrate that the free riders are heavily penalized even if punishment is costly and does not provide any material benefits to the punisher. The opportunity for costly punishment causes a large increase in cooperation levels because potential free riders face a credible threat. In the presence of a costly punishment opportunity, almost complete cooperation can be achieved and maintained during the games. The main conclusion is that human beings are conditional cooperators, they cooperate providing that others do. The introduction of formal rule is key to enforcing this conditional cooperation.

Herrmann et al. (2008) have used this setup to measure conditional cooperation in 16 different cities across the world. They find that cooperation for the funding of the public good is the highest in Boston or Melbourne and the lowest in Athens and Muscat. This ordering is highly correlated with the rule of law and the transparency of institutions in the corresponding country. More strikingly, Herrmann et al. find that participants in some cities, like Athens, display anti-social punishment behavior: that is, they punish the high contributor instead of the low contributor. The weakness of the rule of law is a strong predictor of this anti-social behavior. Similarly, Rothstein (2011) used various experiments with students in Sweden and Romania to show that their generalized trust and trust in civil servants declined substantially after witnessing a police officer accepting a bribe. His interpretation is that the absence of transparency of institutions and civic spirit of public officials can have very large damaging effects on generalized trust. If public officials, who are expected to represent the law, are corrupt, people infer that most other people cannot be trusted neither.

Other promising research looks at the impact of democracy on cooperation in an experimental setting. Contrary to natural experiments, it is possible to control in the laboratory how cooperation changes when a policy is imposed endogenously through a democratic process or imposed exogenously. This is the design used by Dal Bo et al. (2010). Subjects participate in several prisoners' dilemma games and may choose, by simple majority, to establish a policy that could encourage cooperation by imposing fines on non-cooperators. In some cases, the experimental software randomly overrides the votes of the subjects and randomly imposes, or not, the policy. Before proceeding to play again with either the original or the modified payoffs, the subjects are informed of whether payoffs are modified and whether it was decided by their vote or by the computer. The authors show that the effect of the policy on the percentage of cooperative actions is 40% greater when it is democratically chosen by the subjects than when it is imposed by the computer.

All in all, these studies show that formal rules and conditional cooperation might work as a complement in sustaining cooperative behavior. This is the case when the content of the rules, as in Dal Bo et al. (2010), creates focal points or provides signals about the group members' willingness to cooperate. In other cases, the sudden introduction of formal rules or tougher incentives to cooperate might signal instead that principals do not trust agents or that non-cooperative behavior is diffused in the society. For example, Falk and Kosfeld (2006) study the behavior of experimental subjects in the role of agent, choosing a level of production that was costly to them and beneficial to the principal (the authority). Before the agent's decision, the principal could decide to leave the choice of the level of production completely to the agent's discretion or impose a lower bound on the agent's production. In postplay interviews, most agents agreed with the statement that the imposition of the lower bound was a signal of distrust. In another study, Galbiati and Vertova (2008) investigate a similar effect in the context of cooperation in a minimum effort game. In this case, the authors find that, when principals opt to introduce a formal cooperation rule after having observed agents' effort levels in the first experimental round, most cooperative individuals might reduce their effort level. Eliciting individuals' expectations about others' efforts, the authors find that if principals opt to introduce a formal sanction for those that do not cooperate, most cooperative individuals prefer to live in a society where non-cooperation is widespread.

2.7.2.3 Co-Evolution of Trust and Institutions

Rather than stressing the causal impact of institutions, recent contributions look at the co-evolution of trust and institutions, leading to multiple equilibria. The diffusion of limited morality can reinforce the weakness of institutions because a society with limited morality can be more tolerant of weaker compliance with legal enforcement. The society might thus be trapped in a bad equilibrium where mistrust and weak institutions reinforce each other. In this context, promoting better enforcement might not have any support and effect since limited morality makes the trade opportunities too negligible anyway. Several contributions have documented more precisely the type of institutions that could co-evolve with trust. In particular, recent contributions show the interplay between trust and regulation (Aghion et al. 2010; Pinotti, 2012; Carlin et al. 2009; Francois and Van Ypersele, 2009).

Figure 2.13 shows that there exists a negative correlation between generalized trust and the extent of market regulation, measured by the number of steps required to open a business. Aghion et al. (2010) document that this correlation works for a range of measures of trust, from trust in others to trust in firms and political institutions, as well as for a range of regulatory measures from product markets to labor markets.

Explanations of this negative correlation between trust and regulatory intervention by the public authorities are grounded in the assumption that the state must step in to regulate the relations among individuals when they are incapable of cooperating spontaneously. In this perspective, Aghion et al. (2010) present a simple model explaining this correlation. In their setup, individuals make two decisions: whether or not to become civic, and whether to become entrepreneurs or choose routine (perhaps state) production. Those who become uncivic impose a negative externality on others when they become entrepreneurs (e.g. pollute), whereas those who become civic do not. The community (through voting or some other political mechanism) regulates entry into entrepreneurial activity when the expected negative externalities are large. Regulation narrows choices and hence negative externalities. But regulation itself is implemented by government officials, who demand bribes when they are not civic-minded. In this model, when people expect to live in a civic-spirited community, they expect low levels of regulation and corruption, and so become civic. Their beliefs have a self-justifying property, as their choices lead to civic-mindedness, low regulation, and high levels of entrepreneurial activity. When, in contrast, people expect to live in an uncivic-minded community, they expect high levels of regulation and corruption, and do not become civic. Again, their beliefs are justified, as their choices lead to uncivic-mindedness, high regulation, high corruption, and low levels of entrepreneurial activity. The model has two equilibria: a good one with a large share of civic individuals and no regulation; and a bad one where a large share of uncivic individuals support heavy regulation. Production and welfare are higher in the good equilibrium.

The model explains the correlation between regulation and distrust, and has a number of further implications which are empirically documented using international surveys. The model predicts, most immediately, that distrust influences not just regulation itself, but also the demand for regulation. Distrust generates demand for regulation even when people realize that the government is corrupt and ineffective; they prefer state control to unbridled activity by uncivic entrepreneurs.

The most fundamental implication of the model, however, is that beliefs (as measured by distrust) and institutions (as measured by regulation) co-evolve. Beliefs shape institutions, and institutions shape beliefs. The interactions between institutions and beliefs comprise complementarities that induce multiple equilibria, as in Aghion et al. (2011).

Beyond regulation, trust and social capital are likely to affect the overall quality of institutions and government through political accountability. This is the point made by Nannicini et al. (2010). In a political agency model, the authors show that civic agents are more likely to hold politicians accountable for the aggregate social welfare of the community. They tend to punish politicians who pursue vested interests and grab rents for some specific groups. In contrast, uncivic agents' votes are based on their own or group-specific interest and are more tolerant with amoral politicians. Nannicini et al. (2010) convincingly test the prediction of their model by using cross-district variation in the criminal prosecution of members of the Parliament in Italy. They find that the electoral punishment of political misbehavior, corresponding to receiving a request of criminal prosecution or shirking in parliamentary activity, is considerably larger in electoral districts with high social capital.

2.7.3 Community Characteristics

Distinguished from formal institutions, a large body of the research stresses the role of community characteristics in building trust. One of the most prominent factors identified in this realm is the extent of inequality and ethnic fractionalization.

2.7.3.1 Inequality

The focus on inequality is fueled by the strong negative correlation between trust and Gini indexes across countries and US states in Figures 2.16 and 2.17. High-trusting societies are also more equal, measured by low Gini coefficients, while low-trusting societies show typically higher levels of income inequality, as given by high Gini coefficients. Cross-country and cross-US state regressions controlling for income, population, education, and ethnic fractionalization confirm this correlation (see Table 2.7). Alesina and La Ferrara (2000) show that this negative relationship between trust and income inequality also holds at a more local level within US localities and municipalities. Rothstein and Uslaner (2005) document a within-US-states correlation between the rise in equalities and the decline of trust over the last decades.

A pending issue is that of causality. Inequality might correlate negatively with trust for several reasons. First, as suggested by Rothstein and Uslaner, high levels of trust and cooperation might go along with high preferences for redistribution and can so contribute to lower inequality. On the reverse, high inequality could make individuals perceive themselves unfairly treated by people belonging to social classes different from their own, such that they restrict cooperative action and trust to members from their own class (Rothstein and Uslaner, 2005). Future research is still needed to nail down the causal effect of inequality on trust.

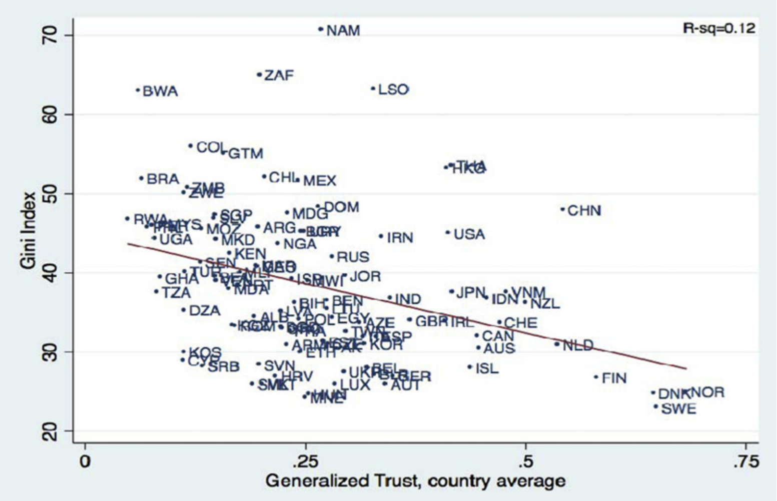


Figure 2.16 Inequality and generalized trust in 101 countries. *Sources: Inequality is measured by average of the Gini Index between 2005 and 2012 (World Bank). Generalized trust is measured as the country average from WVS (1981–2009) and EVS (1981–2008).*

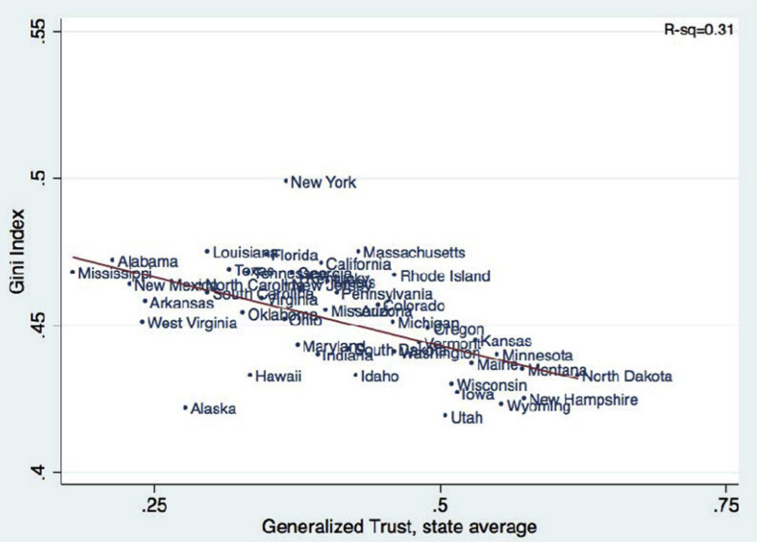


Figure 2.17 Inequality and generalized trust in 46 US states. *Sources: Inequality is measured by the Gini Index in 2010 (US Census Bureau). Generalized trust is taken from the General Social Survey (1973–2006).*

2.7.3.2 Ethnic Fractionalization and Segmentation

The second community characteristic that has attracted attention is ethnic fractionalization or segregation. In a highly debated contribution, Putnam (2007) argues that ethnic diversity drives down trust. Using cross-cities evidence, the author shows that in ethnically diverse neighborhoods, residents' trust is lower; altruism and community cooperation rarer; and friends fewer. Alesina and La Ferrara (2000, 2002) find similar evidence across US states. The explanation for this result is that individuals have natural in-group preferences and have a tendency to trust less those people that are different from them. In the same vein, higher ethnic diversity is associated with lower cooperation as measured by the level of funding and the quality of public goods (Alesina et al. 1999; Miguel and Gugerty, 2005). The main explanations of why ethnic diversity affects those outcomes are the heterogeneity of preferences, and the free-rider problem which undermines collective action. Uslaner (2012) challenges Putnam's thesis and argues that residential segregation, rather than ethnic diversity per se, drives down trust. Using cross-US states evidence, Uslaner shows that both integrated and diverse neighborhoods are associated with higher levels of trust only when people have diverse social networks. Conversely, in areas with a lot of segregation and where individuals from different ethnic backgrounds cannot meet each other, distrust is higher. One conclusion is that immigration and urbanization policy should avoid ethnic ghettos to maintain trust.

Yet, the literature on the relationship between cooperation and diversity raises an important identification issue. Due to endogenous residential sorting of individuals on

Table 2.7 Trust and inequality

	Inequality			
	Cross country		US states	
	(1)	(2)	(3)	(4)
Trust	−24.96***	−12.63*	−0.093***	−0.064***
	(5.600)	(7.451)	(0.017)	(0.016)
Income per capita		0.0954		−0.01
		(1.240)		(0.022)
Population		0.324		0.007***
		(0.791)		(0.002)
Education		−1.116**		0.002
		(0.542)		(0.001)
Ethnic segmentation		7.385		
		(5.003)		
Latitude				−0.0004*
				(0.0002)
Longitude				0.0002**
				(0.0001)
Observations	101	89	46	46
R^2	0.122	0.276	0.314	0.680

Notes: The dependent variable *Inequality* measures income inequality as given by the Gini Index. *Trust* is measured from the answer to the question*"Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?"* Trust is equal to 1 if the respondent answers *"Most people can be trusted"* and 0 otherwise.

OLS regressions with robust standard errors in parentheses.

Sample (101 countries): Albania, Algeria, Argentina, Armenia, Australia, Austria, Azerbaijan, Bangladesh, Belarus, Belgium, Benin, Bosnia and Herzegovina, Botswana, Brazil, Bulgaria, Burkina Faso, Canada, Chile, China, Colombia, Croatia, Cyprus, Czech Republic, Denmark, Dominican Republic, Egypt, El Salvador, Estonia, Ethiopia, Finland, France, Georgia, Germany, Ghana, Great Britain, Greece, Guatemala, Hong Kong, Hungary, Iceland, India, Indonesia, Iran, Ireland, Israel, Italy, Japan, Jordan, Kenya, Korea, South, Kosovo, Kyrgyzstan, Latvia, Lesotho, Lithuania, Luxembourg, Macedonia, Madagascar, Malawi, Malaysia, Mali, Malta, Mexico, Moldova, Montenegro, Morocco, Mozambique, Namibia, Netherlands, New Zealand, Nigeria, Norway, Pakistan, Peru, Philippines, Poland, Portugal, Romania, Russian Federation, Rwanda, Senegal, Serbia, Singapore, Slovakia, Slovenia, South Africa, Spain, Sweden, Switzerland, Taiwan, Tanzania, Thailand, Turkey, Uganda, Ukraine, United States, Uruguay, Venezuela, Vietnam, Zambia, Zimbabwe.

Sources: Trust data used in regressions in columns (1) and (2) comes from the five waves of the World Values Survey (1981–2008), and the four waves of the European Values Survey (1981–2008), for regressions in columns (3) and (4) from the US GSS (1973–2006). Income per capita measures the regions average log income per capita. Population gives the log of the total population living in the region. Education in column (2) measures average years of schooling between 1950 and 2010 and is taken from Barro and Lee (2010), in column (4) the fraction of population having an advanced degree. Ethnic fractionalization measures the degree of ethnic fractionalization and is taken from Alesina et al. (2003). Latitude and longitude refer to the region's geographic position.

*Coefficients are statistically different from 0 at the 10% level.

**Coefficients are statistically different from 0 at the 5% level.

***Coefficients are statistically different from 0 at the 1% level.

ethnic grounds, the estimates are likely to be biased. The attempts to establish causality rely mainly on instrumental variables. However convincing the instruments might be, this strategy cannot overcome the concern as to whether the instruments fulfill the exclusion restriction and do not have a direct effect on public goods. For instance, Miguel and

Gugerty (2005) use the pre-colonial patterns of settlement as instruments, assuming that these variables have no direct impact on present-day ethnic relations. But, since past settlement patterns are likely to have at least some direct impact on the present-day level of cooperation, the exclusion restriction might still be violated. Algan et al. (2012b) address this issue by using a natural experiment in which households in France are allocated to public housing blocks without taking their ethnic origin or their preference for diversity into account. Due to a strongly republican ideology, the French public housing system allocates state-planned, moderate-cost, rental apartments to natives and immigrants without concern for their cultural and ethnic background, mixing people indiscriminately. Using data from housing blocks made up of 20 adjacent households, the authors show that higher ethnic diversity is associated with social anomia rather than distrustful relationships. Yet, more research has to be done before drawing policy conclusions. One of the most promising agendas would be to use a randomized housing mobility program, in the same vein of Moving to Opportunity (see Katz et al. 2013), to investigate how the changes in the ethnic composition of the neighbors modify cooperation and trust.

2.7.4 Education and Trust

A large literature argues that a central component trust derives from moral values deeply ingrained in personality traits, and does not just boil down to context-dependent beliefs about others', trustworthiness. A trusting person that accidentally meets an non-trustworthy person will not change his moral values right away. Moral values of cooperation have a rather stable component because they have been shaped in the early ages by parents or at school. In this section, we review the evidence on the relationship between education and trust.

There is some evidence that a greater quantity of schooling is associated with higher social capital (Helliwell and Putnam, 2007). Yet, variation in the average years of education of the population across developed countries is too small to explain the observed cross-country differences in trust.

Algan et al. (2013a) propose a complementary explanation by looking at the relationship between how students are taught, and students' beliefs in cooperation. They show that methods of teaching differ greatly across countries, between schools, and within schools within a country. Some schools and teachers emphasize vertical teaching practices, whereby teachers primarily lecture, students take notes or read textbooks, and teachers ask students' questions. The central relationship in the classroom is between the teacher and the student. Other schools and teachers emphasize horizontal teaching practices, whereby students work in groups, do projects together, and ask teachers' questions. The central relationship in the classroom is among students. Consistent with the idea that beliefs underlying social capital are acquired through the practice of cooperation, and that social skills are acquired in early childhood, Algan et al. (2013a) test whether horizontal teaching practices can develop social capital. They use various international surveys, like the Civic

Education Study (CES), the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS), covering around 60 countries. They emphasize the distinction between "teacher lectures" and "students work in groups" as measures of vertical and horizontal teaching practices, respectively.

Figure 2.18 shows that teaching practices vary systematically across countries. The *x*-axis represents the average gap between vertical teaching (teacher lectures) and horizontal teaching (students work in groups) in a typical hour of class. The higher the indicator, the more the country is tilted toward vertical teachings. Students work in groups more in Nordic countries (Denmark, Norway, Sweden) and Anglophone countries (Australia, United States, and to a lesser extent, Great Britain). This teaching practice is less common in east European countries and in the Mediterranean (Greece, Cyprus, Portugal and, to a lesser extent, Italy). In these countries, teachers spend more time lecturing. Education in some countries, like France, is almost entirely based on vertical teaching. Figure 2.18 also shows that vertical teaching is highly negatively correlated with generalized trust across countries. This result still holds when per capita income, education expenditures, and average years of education are controlled for.

The authors then investigate within-school and within-classroom variation in teaching practices to identify the causal impact of these practices on students' beliefs. By looking at teaching practices and student beliefs across classrooms within a school, the authors can alleviate concerns regarding omitted variables that might drive the self-sorting of

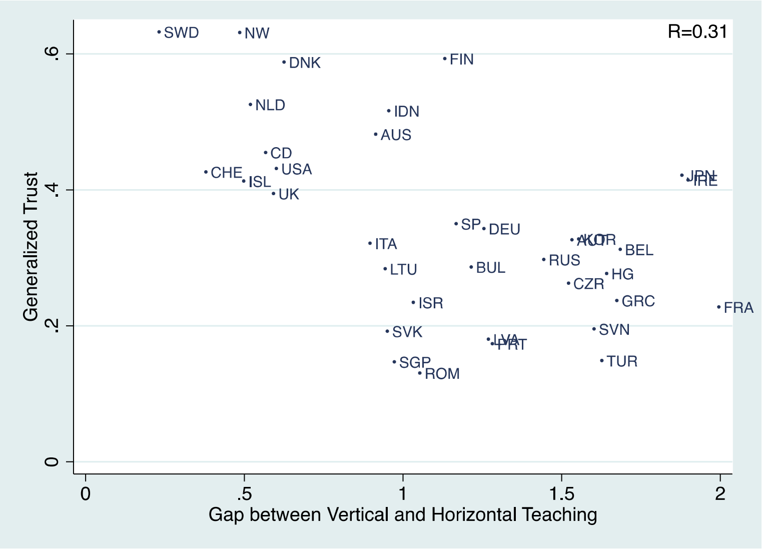


Figure 2.18 Trust and the gap between vertical and horizontal teaching. *Sources: TIMSS, WVS.*

parents, students, and teachers into schools. They also use within-classroom variation in teaching practices and student beliefs. This strategy eliminates concerns about omitted variables linked to selection into classrooms. It also provides an alternative strategy for excluding reverse causality by comparing teaching practices of different teachers faced with exactly the same group of students. The authors show that horizontal teaching practices have a substantial positive impact on students' social capital (trust in teachers, in other students, association membership…), while vertical teaching practices crowd out beliefs in cooperation. The relationship between working in groups and students' social capital is robust whatever the specification: across schools, within schools and within classrooms. The within school (and within classroom) estimates allow the authors to address self-selection and reverse causality. But another concern is that horizontal teaching practices just proxy for a teacher being good or nice. This is a traditional issue raised by cross-section analysis since it is impossible to control for teacher-fixed effect in this setting. The authors show that teaching practices are not a proxy for "good" or "nice" teachers based on observable teacher characteristics. But the teaching practice can still be driven by an unobserved teacher (or student) characteristic.

A promising avenue of research would consist in providing randomized evaluations of early childhood intervention aimed at developing children's social skills, e.g. their aptitude to cooperate with others. This investigation is timely and important given that recent longitudinal studies suggest that much of the impact of programs that improve adult achievement (such as the Perry Preschool program or project STAR) flows through some sort of non-cognitive channel, and thus raise the question of what those non-cognitive skills are, and how much of the impact comes through social skills (see Heckman et al. for a recent synthesis, 2010). In the literature, non-cognitive skills embrace all personality traits that are non-related to cognitive skills (e.g. IQ and grades), such as self-esteem and emotional well-being measured on psychological scales. This is thus a rather vague notion and it is still unclear how non-cognitive skills relate to social skills. Besides, there is little evidence on whether and how intervention can improve those skills, in particular among children the most at risk of becoming anti-social adults.

Algan et al. (2013b) provide a first attempt to estimate the long-term effects of an early intervention that is specifically dedicated to social skills development. The authors use data from a large and detailed longitudinal study following the social, cognitive, and emotional development of 895 men who were kindergarteners in neighborhoods of low socioeconomic status in Montreal in 1984. The study incorporates a randomized evaluation of an intensive two-year social skills training program at the beginning of elementary school for the most disruptive subjects ($n = 250$). The training program involves the subjects themselves, parents, and peers. These detailed data are matched with self-reported outcomes and administrative records. As adults, the subjects in the treated group have significantly better labor market performance than the non-treated group, with an increase in the likelihood of employment at age 26 of 10% points. Individuals who belong to the treated group have significantly more favorable social outcomes, measured by lower criminality rates and higher social capital. By distinguishing the different cognitive and non-cognitive channels through which this intervention operates, the authors find that the only significant channel for economic outcomes is social skills. The overall rate of return of this program in terms of expected lifecycle income ranges from 282% to 452%, implying that every $1 invested yields $2.8 to $4.5 in benefits. This result provides room for policy intervention to develop social skills in early childhood. They call for future experiments to assess the deep personality traits that explain social skills and how they relate to non-cognitive skills.

2.8. FUTURE AVENUES: TRUST AND WELL-BEING

This survey documents two main findings. First, trust has a causal impact on economic development, through its channels of influence on the financial, product, and labor markets, and with a direct effect on total factor productivity and organization of firms. Second, trust and institutions strongly interact, with causality running in both directions. These findings set new avenues of research to identify the policies that could promote social capital and cooperation, from rule of law and democracy to education policies.

This survey has mainly focused on economic and institutional issues related to trust. Yet there is a growing consensus that economic development is poorly measured by income per capita alone, and should include measures of well-being. One reason for that is the well-known Easterlin paradox, stressing that the increase in income per capita within countries has not been associated with an increase in happiness. To explain this result, recent contributions suggest that well-being depend essentially on the quality of social relationship, instead of individual income. From this perspective, we should expect a strong correlation between trust and well-being.

Figure 2.19 illustrates this relationship by using measures of life satisfaction from the World Values Survey question: "All things considered together, how satisfied are with your life as a whole these days." Life satisfaction ranges from 1 to 10, a higher score indicating a higher life satisfaction. The correlation between life satisfaction and generalized trust is positive: 17% of the variance in life satisfaction is associated with cross-country differences in generalized trust, with a few outliers like Portugal. The same positive correlation holds if we consider the question on happiness: "Taking all things together, would you say that you are: very happy, happy, quite happy, not happy, not at all happy?"

Helliwell and Wang (2010) provide cross-country micro evidence on the positive relationship between trust and well-being. From the 2006 wave of the Gallup World Poll, they use the wallet trust question for 86 countries. Individuals are asked what is the hypothetical likelihood of the respondent's lost wallet (with clear identification and $200 cash) being returned if found by a neighbor, a police officer, or a stranger. Helliwell and Wang estimate that an increase in income by two-thirds is necessary to compensate

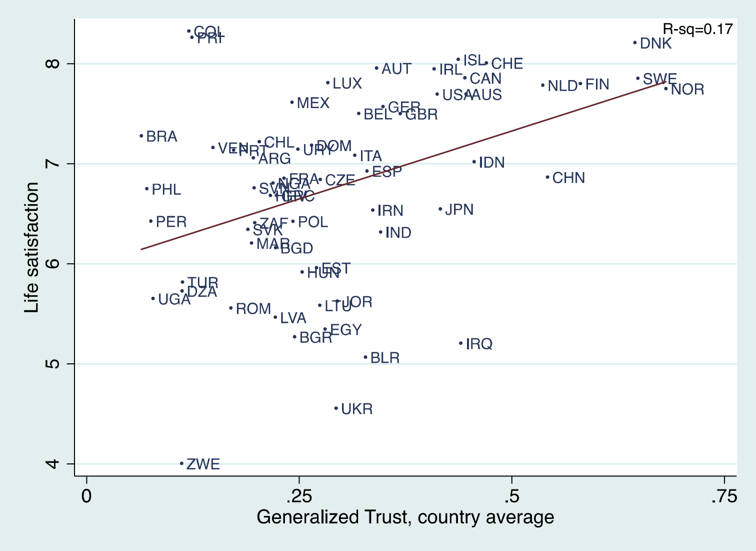


Figure 2.19 Trust and life satisfaction. *Sources: Life satisfaction (1–10) and generalized trust are taken from the World Values Survey (2008).*

the welfare loss associated with thinking that no one will return your wallet and your documents. For example, to live in a country like Norway (highest mean expected wallet return of 80%) rather than in Tanzania (lowest mean expected wallet return of 27%) is equivalent to an increase by 40% of household income. Helliwell and Huang (2010) showed that the same result holds in the workplace. Using micro data from Canada (2003 wave Equality, Security, and Community Survey) and US (2000 wave of the Social Capital Benchmark Survey), the authors find that the climate of trust in workplace, in particular trust in managers, is strongly related to subjective well-being. On a 1–10 scale, an increase by one point of trust in managers has the same effect on life satisfaction as an increase in household income by 30%.

Examining our psychological reactions allows us to better understand the importance of these relations. Imagine that you participate in the trust game, but that one measures now the level of oxytocin in your blood. As mentioned above, oxytocin is a neurotransmitter released by our lymbic system, the part of our brain which is responsible for pleasure or fright. Zak et al. (2004) have tried to find out if trust and reciprocity are equally linked to that love hormone. For that, they have applied the trust game during which levels of oxytocin are measured in the blood of the receiver, once he finds out whether the sender has trusted him by sending a non-negligible amount. The results indicate that trust produces happiness: the more the signaled level of trust is increased (meaning, the more the amount transferred is increased) the more the level of oxytocin increases in the blood of the receiver. Zak et al. (2004) also conducted an experiment using a particularly instructive variant, in which the receiver receives a monetary transfer not from a real person, but from a lottery. In this variant, the level of oxytocin does not rise with the money received. This result well illustrates that it is trust that is associated with sentiments of happiness, and not the mere fact of receiving money.

These results have been confirmed by brain images made by Sanfey et al. (2002). As soon as the participants of the trust game note that the others do not cooperate, the insular part of the cortex in their brain illuminates. This brain part is known for being active in states of pain and disgust. The main conclusion of this line of research is that the non-monetary dimension of having cooperative social relationship with others affects more happiness than the monetary gains derived from cooperation. All in all, those results suggest that trust affects many dimensions of economic development, including both income and happiness, and is a key component in human development at large.

ACKNOWLEDGMENTS

We thank Elizabeth Beasley, Samuel Bowles, Ernst Fehr, Oded Galor, Paola Guiliano, John Helliwell, Karla Hoff, Andrei Shleifer, and Mathias Thoenig for their helpful comments. We also thank Johannes Buggle for his outstanding work as a research assistant for this article. Yann Algan acknowledges financial support from the ERC Starting Grant 240923 under the European Community Seventh Framework Programme (FP7/2007–2013).

REFERENCES

Acemoglu, D., Robinson, J., Johnson, S., 2001. The colonial origins of comparative development: an empirical investigation. American Economic Review 91, 1369–1401.

Aghion, P., Algan, Y., Cahuc, P., 2011. Can policy affect culture? minimum wage and the quality of labor relations. Journal of the European Economic Association 9 (1), 3–42.

Aghion, P., Algan, Y., Cahuc, P., Shleifer, A., 2010. Regulation and distrust. Quarterly Journal of Economics 125 (3), 1015–1049.

Aghion, P., Tirole, J., 1997. Formal and real authority in organizations. The Journal of Political Economy 105 (1), 1–29.

Alesina, A., Baqir, R., Easterley, W., 1999. Public goods and ethnic divisions. Quarterly Journal of Economics 114 (4), 1243–1284.

Alesina, A., La Ferrara, E., 2000. Participation in heterogeneous communities. Quarterly Journal of Economics 115 (3), 847–904.

Alesina, A., La Ferrara, E., 2002. Who trusts others? Journal of Public Economics 85 (2), 207–234.

Alesina, A., Devleeschauwer, A., Wacziarg, R., Kurlat, Sergio, Easterly, W., 2003. Fractionalization. Journal of Economic Growth 8 (2), 155–194. <http://ideas.repec.org/a/kap/jecgro/v8y2003i2p155-94.html>. <http://ideas.repec.org/s/kap/jecgro.html>.

Algan, Y., Cahuc, P., 2009. Civic virtue and labor market institutions. American Economic Journal: Macroeconomics 1 (1), 111–145.

Algan, Y., Cahuc, P., 2010. Inherited trust and growth. American Economic Review 100, 2060–2092.

Algan, Y., Cahuc, P., Sangnier, M., 2011. Efficient and Inefficient Welfare States, Institute for the Study of Labor, DP 5445.

Algan, Y., Benkler, Y., Fuster Morell, M., Hergueux, J., 2012a. "Cooperation in a Peer Production Economy: Experimental Evidence from Wikipedia", Working Paper Sciences Po.

Algan, Y., Hémet, C., Laitin, D., 2012b. The Social Effect of Ethnic Diversity at a Local Level: A Natural Experiment with Exogeneous Residential Allocation, Working Paper Sciences Po.

Algan, Y., Cahuc, P., Shleifer, A., 2013a. Teaching practices and social capital. American Economic Journal: Applied Economics, 5(3), 189–210.

Algan, Y., Beasley, E., Tremblay, R., Vitaro, F., 2013b. The Long Term Impact of Social Skills Training at School Entry: A Randomized Controlled Trial. Working Paper.

Almond, G., Verba, S., 1963. The Civic Culture: Political Attitudes and Democracy in Five Nations. Sage Publications, London (first ed. 1989).

Arrow, K., 1972. Gifts and exchanges. Philosophy and Public Affairs 1, 343–362.

Ashraf, Q., Galor, O., 2013. The "Out-of-Africa" hypothesis, human genetic diversity, and comparative economic development. American Economic Review, 103(1), 1-46.

Banfield, E., 1958. The Moral Basis of a Backward Society. Free Press, New York.

Barro, R., Lee, J.W., 2010. A new data set of educational attainment in the world, 1950–2010. NBER Working Paper No. 15902.

Barr, A., Serneels, P., 2009. Reciprocity in the workplace. Experimental Economics 12 (1), 99–112.

Becker, S., Boeckh, K., Hainz, C., Woessmann, L., 2011. The Empire Is Dead, Long Live the Empire! Long-Run Persistence of Trust and Corruption in the Bureaucracy. IZA, Discussion Paper No. 5584, Mars 2011.

Benabou, R., Tirole, J., 2011. Laws and Norms. NBER Working Paper no 17579.

Benabou, R., Tirole, J., 2006. Incentives and prosocial behavior. American Economic Review 96 (5), 1652–1678.

Benz, M., Meier, S., 2008. Do people behave in experiments as in the field?—evidence from donations. Experimental Economics, 11(3), 268–281.

Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity and social history. Games and Economic Behavior 10, 122–142.

Bertrand, M., Schoar, A., 2006. The role of family in family firms. The Journal of Economic Perspectives 20 (2), 73–96.

Bidner, C., Francois, P., 2011. Cultivating trust: norms, institutions and the implications of scale. Economic Journal 121 (5), 1097–1129.

Bisin, A., Verdier, T., 2001. The economics of cultural transmission and the dynamics of preferences. Journal of Economic Theory 97, 298–319.

Blanchard, O., Philippon, T., 2004. The Quality of Labor Relations and Unemployment. MIT Department of Economics Working Paper No. 04–25.

Bardhan, P., 2000. Irrigation and cooperation: an empirical analysisof 48 irrigation communities in South India. Economic Development and Cultural Change 48 (4), 847–865.

Black, S., Lynch, L., 2001. How to compete: the impact of workplace practices and information technology on productivity. The Review of Economics and Statistics 83 (3), 434–445.

Bloom, N., Sadun, R., Van Reenen, J., 2012. The organization of firms across countries. Quarterly Journal of Economics, 1663–1705.

Bohnet, I., Zeckhauser, R., 2004. Trust, risk and betrayal. Journal of Economic Behavior & Organization 55, 467–484.

Botero, J., Djankov, S., La Porta, R., Lopez-De-Silanes, F., Shleifer, A., 2004. The regulation of labor. Quarterly Journal of Economics 119 (4), 1339–1382. <http://ideas.repec.org/a/tpr/qjecon/v119y2004i4p1339-1382.html>. <http://ideas.repec.org/s/tpr/qjecon.html>.

Bowles, S., Polania-Reyes, S., 2012. Economic incentives and social preferences: substitutes or complements? Journal of Economic Literature 50 (2), 368–425.

Butler, J., Paola, G., Guiso, L., 2009. The Right Amount of Trust. NBER Working Paper 15344.

Capra, M., Lanier, K., Meer, S., 2008. Attitudinal and Behavioral Measures of Trust: A New Comparison. Working Paper, Emory University, Department of Economics.

Carlin, B.I., Dorobantu, F., Viswanathan, S., 2009. Public trust, the law, and financial investment. Journal of Financial Economics 92 (3), 321–341.

Carpenter, J., Seki, E., 2011. Do social preferences increase productivity? Field experimental evidence from fishermen in Toyama Bay. Economic Inquiry, 49(2), 612–630.

Castillo, M., Carter, M., 2011. Behavioral, Responses to Natural Disasters. Working Paper, University of Wisconsin-Madison.

Cavalli-Sforza, L.L., Feldman, M., 1981. Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press, Princeton.

Cingano, F., Pinotti, P., 2012. Trust, Firm Organization, and the Structure of Production. Working Paper.

Cole, S., Gine, X., Tobacman, J., Townsend, R., Vickery, J., 2013. Barriers to household risk management: evidence from India. American Economic Journal: Applied Economics 5(1), 104–135.

Coleman, J., 1990. Foundations of Social Theory. Harvard University Press.

Cox, J., 2004. How to identify trust and reciprocity. Games and Economic Behavior 46, 260–281.

Dal Bo, P., Forster, A., Putterman, L., 2010. Institutions and behavior, experimental evidence on the effects of democracy. American Economic Reveiw 100 (5), 2205–2229.

Dinesen, P.T., 2012. Parental transmission of trust or perceptions of institutional fairness? Explaining generalized trust of young non-Western immigrants in a high-trust society. Comparative Politics 44(3), 273–289.

Dinesen, P.T., Hooghe, M., 2010. When in Rome, do as the Romans do: the acculturation of generalized trust among immigrants in western Europe. International Migration Review 44 (3), 697–727.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., 2012. The intergenerational transmission of risk and trust attitudes. Review of Economic Studies 79 (2), 645–677.

Durante, R., 2010. Risk Cooperation and the Economic Origin of Social Trust: An Empirical Investigation. Working Paper, Economic Department, Sciences-Po.

Durlauf, S., 2002. On the empirics of social capital. Economic Journal 112 (438), 459–479.

Durlauf, S., Fafchamps, M., 2005. Social capital, handbook of economic growth. In: Aghion, Philippe, Durlauf, Steven (Eds.), Handbook of Economic Growth, vol. 1. North Holland, pp. 1639–1699 (Chapter 26).

Ellison, G., 1994. Cooperation in the Prisoner's dilemma with anonymous random matching. The Review of Economic Studies 61 (3), 567–588.

Ermisch, J., Gambetta, D., 2010. Do strong family ties inhibit trust? Journal of Economic Behavior and Organisations 75(3), 365–376.

Ermisch, J., Gambetta, D., Heather, L., Siedler, T., Uhrig, N., 2009. Measuring people's trust. Journal of the Royal Statistical Society Series A 172 (4), 749–769.

Falk, A., Kosfeld, M., 2006. The hidden costs of control. American Economic Review 96 (5), 1611–1630.

Fehr, E., Schimdt, K., 1999. A theory of fairness, competition and cooperation. Quarterly Journal of Economics 114 (3), 817–868.

Fehr, E., Gaetcher, S., 2000. Cooperation and punishment in public goods games. American Economic Review 4, 980–994.

Fehr, E., Fischbacher, U., Schupp, B., Von Rosenbladt, J., Wagner, G., 2002. A Nation Wide Laboratory. Examining Trust and Trustworthiness by Integrating Behavioral Experiments into Representative Surveys. CESifo Working Paper.

Fehr, E., 2009. On the economics and biology of trust, presidential address at the 2008 meeting of the european economic association. Journal of the European Economic Association 7 (2–3), 235–266 (04–05).

Fernandez, R., 2011. Does culture matter? In: Benhabib, J., Bisin, A., Jackson, M.O. (Eds.), Handbook of Social Economics. North Holland.

Fisman, R., Miguel, E., 2007. Culture of corruption: evidence from diplomatic parking ticket. Journal of Political Economy 115 (6), 1020–1048 (2007).

Francois, P., Zabojnik, J., 2005. Trust, social capital, and economic development. Journal of the European Economic Association, MIT Press, 3 (1), 51–94 (03).

Francois, P., van Ypersele, T., 2009. Doux Commerce: Does Market Competition Cause Trust? Working Paper.

Frey, B., 1998. Institutions and morale: the crowding-out effect. In: Ben-Ner, Avner, Putterman, Louis (Eds.), Economics, Values, and Organization, Cambridge University Press, New York, pp. 437–460.

Fukuyama, F., 1995. Trust: The Social Virtues and the Creation of Prosperity. Free Press, New York.

Galbiati, R., Vertova, P. 2008. Obligation and cooperative behavior in public good games. Games and Economic Behavior 64 (1), 146–170.

Gambetta, D., 1993. The Sicilian Mafia. The Business of Private Protection, Harvard University Press.

Gennaioli, N., La Porta, R., Lopez-de-Silanes, F., Shleifer, A., 2013. Human capital and regional development. Quarterly Journal of Economics 128 (1), 105–164.

Gintis, H., Bowles, S., Boyd, R., Fehr, E., 2005. Moral sentiments and material interests: origins, evidence, and consequences. In: Moral Sentiments and Material Interests. MIT Press (Chapter 1).

Glaeser, E., Laibson, D., Scheinkman, J., Soutter, C., 2000. Measuring trust. Quarterly Journal of Economics 115, 811–846.

Glaeser, E., La Porta, R., Lopez-de-Silanes, F., Shleifer, A., 2004. Do Institutions cause growth? Journal of Economic Growth 9, 271–303.

Greif, A., 1993. Contract enforceability and economic institutions in early trade: the Maghribi traders' coalition. American Economic Review 83, 525–548.

Greif, A., 1994. Cultural beliefs and the organization of society: a historical and theoretical reflection on collectivist and individualist societies. Journal of Political Economy 102, 912–950.

Greif, A., Tabellini, G., 2010. Cultural and institutional bifurcation: China and Europe compared. American Economic Review Papers and Proceedings 100 (2), 1–10.

Guiso, L., Jappelli, T., 2005. Awarness and stock market participation. Review of Finance 9 (4), 537–567.

Guiso, L., 2010. A Trust-Driven Financial Crisis, Implications for the Future of Financial Markets. EIEF Working Paper.

Guiso, L., Sapienza, P., Zingales, L., 2004. The role of social capital in financial development. American Economic Review 94 (3), 526–556.

Guiso, L., Sapienza, P., Zingales, L., 2006. Does culture affect economic outcomes? Journal of Economic Perspectives 20 (2), 23–48.

Guiso L., Sapienza, P., Zingales, L., 2008a. Long Term Persistence. Working Paper 14278, National Bureau of Economic Research August 2008.

Guiso, L., Sapienza, P., Zingales, L., 2008b. Alfred Marshall lecture: social capital as good culture. Journal of the European Economic Association 6 (2–3), 295–320.

Guiso L., Sapienza, P., Zingales, L., 2011. Civic capital as the missing link. In: Benhabib, Jess, Bisin, Alberto, Jackson, Matthew O. (Eds.), Handbook of Social Economics, vol. 1A. North Holland.

Hall, R., Jones, C., 1999. Why do some countries produce so much more output per worker than others. Quarterly Journal of Economics 114 (1), 83–116.

Hauk, E., Saez-Marti, M., 2002. On the cultural transmission of corruption. Journal of Economic Theory 107 (2), 311–335.

Heckman, J.J., Malofeeva, L., Pinto, R., Savelyev, P., 2010. Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. American Economic Review.

Helliwell, J., Putnam, R., 2007. Education and social capital. Eastern Economics Journal 33 (1), 1–19.

Helliwell, J., and Huang, H. 2010. How's the Job? Well-being and social capital in the workplace. Industrial and Labor Relations Review 63 (2), 205–227.

Helliwell, J., Wang, S., 2010. Trust and well-being. International Journal of Well-Being 1 (2), 42–78.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., 2001. In search of Homo Economicus: Behavioral experiments in 15 small-scale societies. American Economic Review Papers and Proceedings, 73–79.

Herrmann, B., Thöni, C., Gachter, S., 2008. Antisocial punishment across societies. Science 319, 2008.

Hoff, K., Kshetramade, M., Fehr, E., 2011. Caste and punishment: the legacy of caste culture in norm enforcement. Economic Journal 121, 449–475.

Holm, H., Danielson, A., 2005. Tropic trust versus nordic trust: experimental evidence from Tanzania and Sweden. The Economic Journal 115, 505–532.

Ichino, A., Maggi, G., 2000. Work environment and individual background: explaining regional shirking differentials in a large Italian firm. Quarterly Journal of Economics 115 (3), 1057–1090.

Jacob, M., Tyrell., M., 2010. The Legacy of Surveillance: An Explanation for Social Capital Erosion and the Persistence of Economic Disparity Between East and West Germany. European Business School Oestrich-Winkel, mimeo.

Kahneman, D., Tversky, A., 2000. Choices, Values and Frames. Cambridge University Press.

Kandori, M., 1992. Social norms and community enforcement. Review of Economic Studies 59, 63–80.

Karlan, D., 2005. Using experimental economics to measure social capital and predict financial decisions. American Economic Review 95 (5), 1688–1699.

Katz, L., Ludwig, J., Duncan, G., Gennetian, L., Kessler, R., Kling, J., Sanbonmatsu, L., 2013. Long-term neighborhood effects on low-income families: Evidence from moving to opportunity, American Economic Review, 103, 226–231.

Kaufmann, D., Kraay, A., Mastruzzi, M., 2010. The worldwide governance indicators: methodology and analytical issues. World Bank Policy Research Working Paper No. 5430.

Kosfeld, M., Heinrichs, M., Zak, P., Fischbacher, U., Fehr, E., 2005. Oxytocin increases trust in humans. Nature 435, 673–676.

Knack, S., Keefer, P., 1997. Does social capital have an economic payoff? a cross-country investigation. Quarterly Journal of Economics 112 (4), 1252–1288.

Knack, S., Zak, P., 1999. Trust and growth. Economic Journal 111 (470), 295–321.

Kranton, R. 1996. Reciprocal exchange, a self-sustaining system. American Economic Review 86 (4), 830–851.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R., 1997. Trust in large organizations. American Economic Review Papers and Proceedings 87 (2), 333–338.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., 2008. The economics consequences of legal origins. Journal of Economic Literature 46 (2), 285–332.

Laury, S.K., Taylor, L.O., 2008. Altruism spillovers: are behaviors in context-free experiments predictive of altruism toward a naturally occurring public good? Journal of Economic Behavior & Organization 65 (1), 9–29.

Lazzarini, S., Artes, R., Madalozzo, R., Siqueira, J., 2005. Measuring trust: an experiment in Brazil. Brazilian Journal of Applied Economics 9 (2), 153–169.

Levine, R., 2004. Finance and growth: theory, evidence and mechanisms. In: Aghion, P., Durlauf, S. (Eds.), Handbook of Economic Growth, North-Holland, Amsterdam, Netherlands.

Ljunge, M., 2012. Inherited Trust and Economic Success of Second Generation Immigrants. IFN Working Paper.

Michalopoulos, S., Papaioannou, E., 2013. National Institutions and Subnational Development in Africa. Working Paper.

Miguel, E., Gugerty, M.K., 2005. Ethnic diversity, social sanctions, and public goods in Kenya. Journal of Public Economics 89 (11), 2325–2368.

Miguel, E., Saiegh, S., Satyanath, S., 2011. Civil war exposure and violence. Economics and Politics 23, 59–73.

Mill, J.S., 1848. Principles of Political Economy. John W. Parker, London.

Nannicini, T., Stella, A., Tabellini, G., Troiano, U., 2010. Social capital and political accountability, economic policy. American Economic Journal.

Nunn, N., Wantchekon, L., 2011. The slave trade and the origins of mistrust in Africa. American Economic Review 101 (7), 3221–3252.

Nunn, N., 2009. The importance of history for economic development. Annual Review of Economics 1, 65–92.

Oliveira, A.D., Croson, R.T.A., Eckel, C.C., 2009. Are Preferences Stable Across Domains? An Experimental Investigation of Social Preferences in the Field. Working Paper.

Olken, B., 2009. Do TV and radio destroy social capital? Evidence from Indonesian villages. American Economic Journal: Applied Economics 1 (4), 1–33.

Olson, M., 1971 [1965]. The Logic of Collective Action: Public Goods and the Theory of Groups (Revised ed.). Harvard University Press.

Ostrom, E., 1990. Governing the commons: the evolution of institutions for collective action. Cambridge University Press, Cambridge.

Putnam, R., Leonardi, R., Nanetti, R.Y., 1993. Making Democracy Work. Princeton University Press, Princeton, NJ.

Putnam, R., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon and Schuster, New York.

Rodrik, D., 1999. Where did all the growth go? External shocks, social conflict, and growth collapses. Journal of Economic Growth 4 (4), 385–412.

Rohner, D., Thoenig, M., Zilibotti, F., 2013. War signals: a theory of trade, trust and conflict. Review of Economic Studies, 80 (3), 1114–1147.

Rothstein, B., Uslaner, E.M., 2005. All for One: Equality, Corruption, and Social Trust. World Politics 58 (1), 41–72.

Rothstein, B., 2011. The Quality of Government, Social Trust and Inequality in International Perspective. University of Chicago Press.

Sachs, J, D., 2003. Institutions Don't Rule: Direct Effects of Geography on Per Capita Income. Working Paper 9490, National Bureau of Economic Research.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., Cohen, J.D., 2002. The neural basis of economic decision-making in the ultimatum game. Science 300, 1755–1758.

Smith, A. 1997 [1766]. Lecture on the influence of commerce on manners, reprinted. In: Klein, D.B. (Ed.), Reputation: Studies in the Voluntary Elicitation of Good Conduct. University of Michigan Press.

Spolaore, E., Wacziarg. R., 2013. How deep are the roots of economic development. Journal of Economic Literature 51 (2), 325–369.

Tabellini, G., 2008a. The scope of cooperation: values and incentives. The Quarterly Journal of Economics 123 (3), 905–950.

Tabellini, G., 2008b. Institutions and culture. Journal of the European Economic Association, Papers and Proceedings 6 (2–3).

Tabellini, G., 2010. Culture and institutions: economic development in the regions of Europe. Journal of the European Economic Association 8 (4), 677–716.

Uslaner, E.M., 2008. Corruption, inequality and trust. In: Svendsen, Gert T., Svendsen, Gunnar, L. (Eds.), Handbook on Social Capital, Edward Elgar.

Uslaner, 2012. Segregatin and Mistrust: Diversity, Isolation and Social Cohesion. Cambridge University Press.

Valent, P., 2000. Disaster syndrome. In: Fink, George (Ed.), Encyclopedia of Stress. Academic Press, New York.

Wallace, A., 1956. Tornado in Worcester: An Exploratory Study of Individual and Community Behavior in an Extreme Situation. Publication 392, National Academy of Sciences-National Research Council, Washington, D.C.

Zak, P., Kursban R., Matzner, W., 2004. The neurobiology of trust. Annals of the New York Academy of Sciences 224–227.

Zylberberg, Y., 2011. Do Tropical Typhoons Smash Community Ties? Working Paper, Paris School of Economics.

CHAPTER THREE

Long-Term Barriers to Economic Development

Enrico Spolaore[*] **and Romain Wacziarg**[†]

[*]Department of Economics, Tufts University, NBER, CESIfo and CAGE, Medford, MA, 02155-6722, USA
[†]UCLA Anderson School of Management, NBER and CEPR, 110 Westwood Plaza, Los Angeles, CA, 90024, USA

Abstract

What obstacles prevent the most productive technologies from spreading to less developed economies from the world's technological frontier? In this paper, we seek to shed light on this question by quantifying the geographic and human barriers to the transmission of technologies. We argue that the intergenerational transmission of human traits, particularly culturally transmitted traits, has led to divergence between populations over the course of history. In turn, this divergence has introduced barriers to the diffusion of technologies across societies. We provide measures of historical and genealogical distances between populations, and document how such distances, relative to the world's technological frontier, act as barriers to the diffusion of development and of specific innovations. We provide an interpretation of these results in the context of an emerging literature seeking to understand variation in economic development as the result of factors rooted deep in history.

Keywords

Long-run growth, Genetic distance, Intergenerational transmission, Diffusion of innovations

JEL Classification Codes

O11, O33, O40, O57

3.1. INTRODUCTION

Technological differences lie at the heart of differences in economic performance across countries. A large and growing literature on development accounting demonstrates that total factor productivity accounts for a sizeable fraction of cross-country differences in per capita income (Hall and Jones, 1999; Caselli, 2005; Hsieh and Klenow, 2010, among many others). The problem of low technological advancement in poor countries is not primarily one of lack of innovation, because technologies that could make these countries vastly richer exist and are used elsewhere in the world. A major problem, instead, is one of delayed technological adoption. That many countries are subject to large technological usage gaps is a well-documented phenomenon. However, the factors explaining delayed technological adoption are not well understood. What prevents the most productive technologies, broadly understood, from spreading to less developed economies from the

Handbook of Economic Growth, Volume 2A
ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53538-2.00003-4

world's technological frontier? In this chapter, we seek to shed light on this question, by quantifying the geographic and human barriers to the transmission of technologies.

We adopt a long-term perspective. The fortunes of nations are notoriously persistent through time, and much of the variation in economic performance is deep rooted in history. For instance, an important literature has explored the prehistoric origins of comparative development (Diamond, 1997; Olsson and Hibbs, 2005; Ashraf and Galor, 2011, 2013a). While there have been reversals of fortune at the level of countries, these reversals are much less prevalent when looking at the fortunes of populations rather than those of geographic locations.[1] Indeed, contributions by Putterman and Weil (2010), Comin et al. (2010), and Spolaore and Wacziarg (2009, 2012a, 2013) argue that the past history of populations is a much stronger predictor of current economic outcomes than the past history of given geographical locations. Thus, any explanation for the slow and unequal diffusion of frontier technologies must be able to account for the persistence of economic fortunes over the long run. In this chapter, we argue that the intergenerational transmission of human traits, particularly culturally transmitted traits, has led to divergence between populations over the course of history. In turn, this divergence has introduced barriers to the diffusion of technologies across societies. These barriers impede the flow of technologies in proportion to how genealogically distant populations are from each other.

Our starting point is to develop a theoretical model capturing these ideas. This model proceeds in three phases. Firstly, we argue that genealogical separation across populations leads, on average, to differentiation along a wide range of traits transmitted from parents to children either biologically or culturally. Populations that are genealogically distant should therefore also be distant in terms of languages, norms, values, preferences, etc.—a set of traits we refer to as vertically transmitted traits or more simply as vertical traits. Secondly, we consider the onset of a major innovation, which could be interpreted as the Industrial Revolution, and argue that differences in vertical traits introduce barriers to the diffusion of this major innovation across societies and populations. Thus, cross-country differences in aggregate TFP or per capita income should be correlated with their genealogical distance. Finally, we extend the model to allow for innovations taking place over time, and innovation and imitation occurring endogenously. In this more general framework, usage lags in the adoption of specific technologies and consequently, aggregate differences in economic development are correlated with average differences in vertical traits, and thus with genealogical distance.

We next turn to empirical evidence on these ideas. To measure the degree of relatedness between populations, we use genetic distance. Data on genetic distance was gathered by population geneticists specifically for the purpose of tracing genealogical linkages between

[1] See Acemoglu et al. (2002) for the reversal of fortune at the level of geographic locations (for former colonies), and papers by Spolaore and Wacziarg (2013) and Chanda et al. (2013) showing that the reversal of fortune disappears when correcting for ancestry and expanding the sample beyond former colonies.

world populations (Cavalli-Sforza et al. 1994). By sampling large numbers of individuals from different populations, these researchers obtained vectors of allele frequencies over a large set of genes, or loci. Measures of average differences between these vectors across any two populations provide a measure of genetic distance. The measure we rely on, known as F_{ST} genetic distance, is the most widely used measure in the population genetics literature because it has properties that make it well suited to study separation times between populations—precisely the concept we wish to capture. F_{ST} genetic distance has been shown to correlate with other measures of cultural differences such as linguistic distance and differences in answers to questions from the World Values Survey (Spolaore and Wacziarg, 2009; Desmet et al. 2011).

Emphatically, the purpose of our study is *not* to study any genetic characteristics that may confer any advantage in development. The genes used in our measures of genealogical distance purposely do not capture any such traits. It is important to note that the genes chosen to compare populations and retrace their genealogies are neutral (Kimura, 1968). That is, their spread results from random factors and not from natural selection. For instance, neutral genes include those coding for different blood types, characteristics that are known not to have conferred a particular advantage or disadvantage to individuals carrying them during human evolutionary history. The mutations that give rise to specific alleles of these genes arise and spread randomly. The neutral genes on which genetic distance is based thus do not capture traits that are important for fitness and survival. As a result, measures based on neutral genes are like a molecular clock: on average, they provide an indication of separation times between populations. Therefore, genetic distance can be used as a summary statistics for all divergence in traits that are transmitted with variation from one generation to the next over the long run, including divergence in cultural traits. Our hypothesis is that, at a later stage, when such populations enter into contact with each other, differences in those traits create barriers to exchange, communication, and imitation. These differences could indeed reflect traits that are mostly transmitted culturally and not biologically—such as languages, norms of behavior, values, and preferences. In a nutshell, we hypothesize that genetic distance measured from neutral genes captures divergence in intergenerationally transmitted traits—including cultural traits—between populations. This divergence in turn impedes the flow of innovations.

We use these measures of genetic distance to test our model of technological diffusion. Our barriers model implies that the genetic distance measured relative to the world technological frontier should trump absolute genetic distance as an explanation for bilateral income differences. We find this to be the case empirically. Our model also implies that genetic distance relative to the frontier should have predictive power for income differences across time even in periods when the world distribution of income was quite different from today's. We show indeed that the effect of genetic distance remains strong in historical data on population density and per capita income. Our model implies that after a major innovation, such as the Industrial Revolution, the effect of genealogical distance should be pronounced, but that it should decline as more and more societies adopt the frontier's innovation. This too is true empirically. Finally, our model implies that genetic distance should have predictive power at the level of disaggregated technologies, and find this to be the case both historically (when measuring technological usage on the extensive margin) and for more recent technological developments (measuring technological usage along the intensive margin). In sum, we find considerable evidence that barriers introduced by historical separation between populations are central to account for the world distribution of income.

In the final section of this chapter, we broaden our focus and place these hypotheses and findings in the context of the wider emerging literature on the deep historical roots of economic development. Our discussion starts from a taxonomy, based on Spolaore and Wacziarg (2013), describing how historically transmitted traits could conceivably affect socio-economic outcomes. The taxonomy distinguishes between the mode of transmission of vertical traits, and the mode of operation of these traits. In principle, intergenerationally transmitted traits could be transmitted either biologically or culturally. However, the recent development of the research on epigenetics and on gene-culture interactions has made this distinction based on the mode of transmission much less clear-cut empirically and conceptually. A more fruitful discussion, we argue, is to try to better distinguish between the modes of operation of vertical traits. These traits, in principle, could bear direct effects on economic outcomes, or operate as barriers to economic interactions between populations. We discuss existing contributions in light of this distinction, and discuss directions for future research in the emerging new field concerned with the deep historical roots of economic development.

This chapter is organized as follows. Section 3.2 presents a stylized model of the diffusion of technologies as function of differences in vertically transmitted traits across human populations, and ultimately as a function of the degree of genealogical relatedness between them. Section 3.3 presents our empirical methodology and data. Section 3.4 discusses a wide range of empirical results pertaining to contemporaneous and historical measures of economic development and specific technology use measures. Section 3.5 discusses the interpretation of these results in the context of the broader literature on the deep roots of economic development. Section 3.6 concludes.

3.2. A THEORY OF RELATEDNESS AND GROWTH

In this section we present a basic theoretical framework to capture the links among genetic distance, intergenerationally transmitted traits, and barriers to the diffusion of economic development across different societies.[2] The model illustrates two key ideas.

The first idea is that genetic distance between populations captures the degree of genealogical relatedness between populations over time, and can therefore be interpreted

[2] The model builds on Spolaore and Wacziarg (2009, 2012a).

as a general metric for average differences in traits transmitted with variation across generations. Genetic distance measures the difference in gene distributions between two populations, where the genes under consideration are neutral. By definition, neutral genetic change tends to occur randomly, independently of selection pressure, and regularly over time, as in a molecular clock (Kimura, 1968). This divergence provides information about lines of descent: populations that are closer in terms of genetic distance have shared a common "ancestor population" more recently. The concept is analogous to relatedness between individuals: two siblings are more closely related than two cousins because they share more recent common ancestors: their parents rather than their grandparents. Since a very large number of traits—not only biological but also cultural—are transmitted from one generation to the next over the long run, genetic distance provides a comprehensive measure for average differences in traits transmitted across generations. We call vertically transmitted traits (or vertical traits, for short) the set of characteristics passed on across generations within a population over the very long run—that is, over the time horizon along which populations have diverged (thousands of years).[3] Vertical transmission takes place across generations within a given population, and, in our definition, includes not only direct parent-to-child transmission of biological and cultural traits, but also, more broadly, "oblique" transmission of cultural traits from the older to the younger within a genetically related group. In contrast, we define "horizontal transmission" as learning and imitation across different populations at a point in time.

The second idea is that differences in vertically transmitted traits act as barriers to horizontal learning and imitation, and therefore hamper the diffusion of innovations and economic development across societies.[4] We argue that populations that share a more recent common history, and are therefore closer in terms of vertical traits, face lower costs and obstacles to adopting each other's innovations. This view, that differences in persistent societal characteristics may act as barriers, is consistent with a large literature on the diffusion of innovations, starting with the classic work by Rogers (1962). Empirically, we are interested primarily in the diffusion of modern economic development in historical times, and especially after the Industrial Revolution, so our stylized model is designed with that objective in mind.

3.2.1 Genetic Distance and Vertically Transmitted Traits

We model all vertical traits of a population as a point on the real line: each population i has vertical traits v_i, where v_i is a real number. At time o ("origin"), there exists only one population (population 0), with traits normalized to zero: $v_0 = 0$. At time $p > o$

[3] This terminology is borrowed from the evolutionary literature on cultural transmission (for example, see Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1985; Richerson and Boyd, 2005).

[4] Policy-induced barriers to the diffusion of technology are analyzed by Parente and Prescott (1994, 2002). In our framework we interpret barriers more broadly to include all long-term societal differences that are obstacles to the diffusion of development.

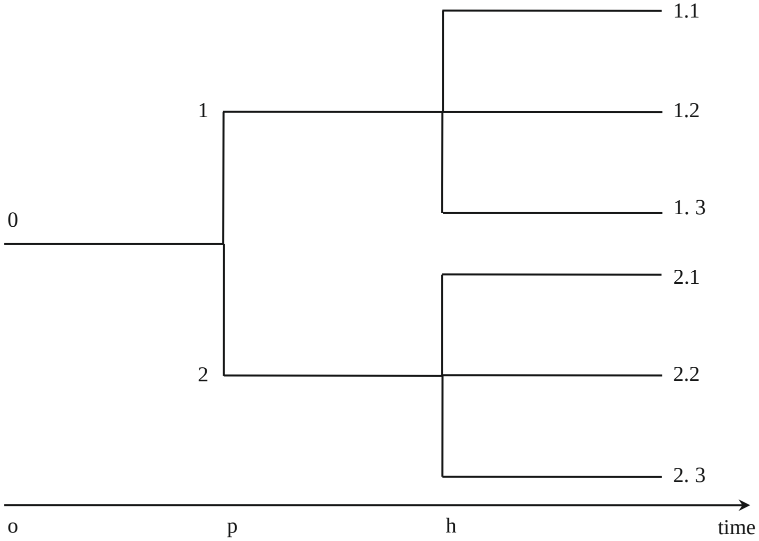


Figure 3.1 Population tree.

("prehistory"), the original population splits into two populations (1 and 2). At time $h > p$ ("history"), each of the two populations splits into three separate populations: population 1 into populations 1.1, 1.2, 1.3; and population 2 into populations 2.1, 2.2, and 2.3.[5] The genealogical tree is displayed in Figure 3.1. By analogy, with the genealogy of individuals, we say that populations such as 1.1 and 1.2 are "sibling" populations, because their last common ancestors (their "parent" population) can be found at the more recent split (time p), while population pairs such as 1.2 and 2.1 are "cousin" populations, because their last common ancestors (their "grandparent" population) must be traced back to a more remote time $o < p$. $G(i, j)$ denotes the genetic distance between population i and population j.[6] The genetic distance between two sibling populations is $g_s > 0$, while the genetic distance between two cousin populations is $g_c > g_s$. Formally,

$$G(1.m, 1.n) = G(2.m, 2.n) = g_s \quad \text{where} \quad m = 1, 2, 3;\ n = 1, 2, 3 \text{ and } 1.m \neq 1.n;\ 2.m \neq 2.n\,, \tag{3.1}$$

and

$$G(1.m, 2.n) = g_c \quad \text{where } m = 1, 2, 3 \text{ and } n = 1, 2, 3. \tag{3.2}$$

[5] In Spolaore and Wacziarg (2009), we presented a similar model with only four populations at time h (1.1, 1.2, 2.1, and 2.2). Here we extend the framework to allow for a more general analysis, in which we also have pairs of populations that, while they are not at the frontier themselves, are both siblings with the frontier population.

[6] By definition, $G(i, i) = 0$.

Each population inherits vertical traits from its ancestor population with variation. In general, vertical traits v_d of population d (the "descendent"), descending from population a (the "ancestor"), are given by:

$$v_d = v_a + \varepsilon_d, \tag{3.3}$$

where ε_d is a shock. In particular, we model the process of variation as a random walk. This simplification is consistent with the molecular-clock interpretation of genetic distance. While more complex processes could be considered, this formalization has two advantages: it is economical and illustrates how random changes are sufficient to generate our theoretical predictions. Formally, we assume that ε_d takes value $\varepsilon > 0$ with probability $1/2$ and $-\varepsilon$ with probability $1/2$. We denote with $V(i,j)$ the distance in vertically transmitted traits (vertical distance, for short) between populations i and j:

$$V(i,j) \equiv |v_j - v_i|. \tag{3.4}$$

We are now ready to summarize our first idea as:

Proposition 1. *The distance in vertical traits $V(i,j)$ between two populations i and j, is, on average, increasing in their genetic distance $G(i,j)$.*

Derivation of Proposition 1:

The expected distance in vertical traits between sibling populations is:

$$E\{V(i,j)|G(i,j) = g_s\} = \varepsilon, \tag{3.5}$$

because their vertical distance is equal to 2ε with probability $1/2$, when one population experiences a positive shock ε and the other a negative shock $-\varepsilon$, and equal to 0 with probability $1/2$, when both populations experience the same shock (either ε with probability $1/4$ or $-\varepsilon$ with probability $1/4$). In contrast, the expected distance in vertical traits between cousin populations is:

$$E\{V(i,j)|G(i,j) = g_c\} = \frac{3\varepsilon}{2}, \tag{3.6}$$

because their vertical distance is 0 with probability $3/8$, 2ε with probability $1/2$, and 4ε with probability $1/8$.[7] Therefore, the expected distance in vertical traits is increasing in

[7] The details of the calculation are as follows. With probability $1/4$, the two populations experienced identical shocks at time h, and their respective ancestor populations experienced identical shocks at time p, implying $V(i,j) = 0$. With probability $1/8$, one population lineage experienced a positive shock ε at time p and a negative shock $-\varepsilon$ at time h while the other population lineage experienced $-\varepsilon$ and ε, implying again $V(i,j) = 0$. With probability $1/4$, the two populations' ancestors experienced identical shocks at time p, but the two populations experienced different shocks at time h, implying $V(i,j) = 2\varepsilon$. With probability $1/4$, the shocks were the same at time h but different at time p, also implying $V(i,j) = 2\varepsilon$. Finally, with probability $1/8$, one population lineage experienced two positive shocks ($\varepsilon + \varepsilon = 2\varepsilon$) and the other two negative shocks ($-\varepsilon - \varepsilon = -2\varepsilon$), therefore leading to a vertical distance equal to 4ε. In sum, their expected vertical distance is given by $E\{V(i,j)|G(i,j) = g_c\} = \frac{3}{8}0 + \frac{1}{2}2\varepsilon + \frac{1}{8}4\varepsilon = \dfrac{3\varepsilon}{2}$.

genetic distance:

$$E\{V(i,j)|G(i,j)=g_c\}-E\{V(i,j)|G(i,j)=g_s\}=\frac{\varepsilon}{2}>0. \tag{3.7}$$

It is important to notice that the relation between distance in vertical traits and genetic distance is not deterministic, but works on average. Some pairs of populations, while genealogically more distant, may end up with more similar vertical traits than two more closely related populations. However, that outcome is less likely to be observed than the opposite. On average, genetic distance and vertical distance go hand in hand.

3.2.2 Barriers to the Diffusion of Economic Development

Our second idea is that differences in vertical traits constitute barriers to the spread of innovations across populations. A stylized illustration of this idea is provided below.

At time p all populations produce output using the basic technology $Y_i = AL_i$, so that all populations have the same income per capita $y = A$. In period h a population happens to find a more productive technology $A' = A + \Delta$ where $\Delta > 0$. We abstract from the possibility that the likelihood of finding the innovation is itself a function of a society's vertical traits. Such direct effects of vertical traits could strengthen the links between genetic distance and economic outcomes, but are not necessary for our results.

We denote the innovating population as f (for technological frontier). To fix ideas and without loss of generality, in the rest of the analysis we assume that population 1.1 is the frontier population ($f = 1.1$). Populations farther from population f in terms of vertical traits face higher barriers to adopt the new technology. Formally, we assume that a society i at a vertical distance from the frontier equal to $V(i,f)$ can improve its technology only by:

$$\Delta_i = [1-\beta V(i,f)]\Delta, \tag{3.8}$$

where the parameter $\beta > 0$ captures the barriers to the horizontal diffusion of innovations due to distance in vertical traits. To ensure non-negativity, we assume that $\beta \leq \frac{1}{\max V(i,f)} = \frac{1}{4\varepsilon}$.[8] Therefore, income per capita in society i will be given by:

$$y_i = A + \Delta_i = A + [1-\beta V(i,f)]\Delta. \tag{3.9}$$

This immediately implies:

Proposition 2. *The difference in income per capita $|y_i - y_j|$ between society i and society j is a function of their relative vertical distance from the frontier $|V(i,f) - V(j,f)|$:*

$$|y_j - y_i| = \beta\Delta|V(i,f) - V(j,f)|. \tag{3.10}$$

[8] Alternatively, the formula could be re-written as $\Delta_i = \max\{[1-\beta V(i,f)]\Delta, 0\}$.

3.2.3 Genetic Distance and Income Differences

Since income differences are associated with differences in vertical traits across populations (Proposition 2), and differences in vertical traits, on average, go hand in hand with genetic distance (Proposition 1), we can now establish a link between expected income differences and genetic distance. These links are formally derived as Propositions 3 and 4 below.

Proposition 3. *The expected income difference $E\{|y_j - y_i|\}$ between societies i and j is increasing in their genetic distance $G(i,j)$.*

Derivation of Proposition 3:

First, we must calculate the expected income of all pairs of populations at genetic distance g_s (sibling populations). $V(i,j)$ between two sibling populations is 0 with probability 1/2 and 2ε with probability 1/2. When the two populations have identical traits, they have identical incomes. When they are at a distance 2ε from each other, one of them must be closer to the frontier's traits by a distance equal to 2ε, no matter where the frontier's traits are located (at $0, 2\varepsilon$, or -2ε), or whether one of the two sibling populations *is* the frontier. Thus, when $V(i,j) = 2\varepsilon$, the income difference between the two populations is $\beta\Delta 2\varepsilon$. In sum, for all pairs of sibling populations is $|y_{k.m} - y_{k.n}| = 0$ with probability 1/2, and $|y_{k.m} - y_{k.n}| = \beta\Delta 2\varepsilon$ with probability 1/2, implying $E\{|y_{k.m} - y_{k.n}|\} = \beta\Delta\varepsilon$ where $k = 1, 2$; $m = 1, 2, 3$; $n = 1, 2, 3$; and $m \neq n$. Consequently, the expected income difference between sibling populations is:

$$E\{|y_j - y_i| \mid\mid G(i,j) = g_s\} = \beta\Delta\varepsilon. \tag{3.11}$$

Now we must calculate the expected income difference between cousin populations. $V(i,j)$ between two cousin populations is 0 with probability 3/8, 2ε with probability 1/2, and 4ε with probability 1/8. The calculation is slightly more complicated, because we must distinguish between pairs that include the frontier and pairs that do not include the frontier $f = 1.1$. First, consider pairs that include the frontier. With probability 3/8 a population $2.n$ shares the same traits (and hence income) with the frontier, with probability 1/2, population $2.n$ has income lower than the frontier's by $\beta\Delta 2\varepsilon$, and with probability 1/8 population $2.n$'s income is lower by $\beta\Delta 4\varepsilon$. Thus, we have:

$$E\{|y_f - y_{2.n}|\} = \frac{\beta\Delta 2\varepsilon}{2} + \frac{\beta\Delta 4\varepsilon}{8} = \frac{3\beta\Delta\varepsilon}{2} \quad \text{where } n = 1, 2, 3. \tag{3.12}$$

Now, consider pairs of cousin populations that do not include the frontier population—that is, pairs $1.m$ and $2.n$, with $m = 2, 3$, and $n = 1, 2, 3$. Again, the income difference between each pair of cousin populations is equal to zero when both populations share the same traits (which happens with probability 3/8), and is equal to $\beta\Delta 2\varepsilon$ when their traits are at a distance 2ε from each other (which happens with probability 1/2), no matter where the frontier is located. However, when the two cousin populations are at a distance 4ε from each other (which happens with probability 1/8), their income distance depends on the location of the traits of the frontier. If the frontier is at an extreme (either 2ε or -$2\varepsilon-$ an event with probability 1/2), the 4ε vertical distance between $1.m$ and $2.n$ implies that their income distance is equal to $\beta\Delta 4\varepsilon$. In contrast, if the frontier's traits are at 0 (also an event with probability 1/2), $1.m$ and $2.n$ are equally distant from the frontier (each at a distance 2ε), and therefore have identical incomes per capita. In sum, we have:

$$E\{|y_{1.m} - y_{2.n}|\} = \frac{\beta\Delta 2\varepsilon}{2} + \frac{1}{2}\frac{\beta\Delta 4\varepsilon}{8} = \frac{5\beta\Delta\varepsilon}{4} \text{ where } m = 2,3;\; n = 1,2,3. \tag{3.13}$$

Consequently, expected income difference between pairs of cousin populations is:

$$E\{|y_j - y_i| \mid\mid G(i,j) = g_c\} = \frac{1}{9}\sum_{m=1}^{3}\sum_{n=1}^{3} E\{|y_{1.m} - y_{2.n}|\} = \frac{1}{9}\left[3\frac{3\beta\Delta\varepsilon}{2} + 6\frac{5\beta\Delta\varepsilon}{4}\right] = \frac{4\beta\Delta\varepsilon}{3}. \tag{3.14}$$

Therefore, the expected income difference between cousin populations is higher than the one between sibling populations: higher genetic distance is associated, on average, with higher income differences, as stated in Proposition 3. Formally:

$$E\{|y_j - y_i| \mid\mid G(i,j) = g_c\} - E\{|y_j - y_i| \mid\mid G(i,j) = g_s\} = \frac{\beta\Delta\varepsilon}{3} > 0. \tag{3.15}$$

Why do populations which are genetically more distant from each other tend to differ more in income per capita, on average? The reason is that populations which are distant from each other genetically are also more likely to find themselves at more different distances from the frontier. Relative distance from the frontier, rather than genetic distance between populations per se, is the key determinant of expected income differences. Therefore, we can find an even stronger relation between income differences and genetic distance if we consider not the absolute genetic distance between two populations $G(i,j)$, but their relative genetic distance from the technological frontier, defined as follows:

$$R(i,j) \equiv |G(i,f) - G(j,f)|. \tag{3.16}$$

Our model predicts that the effect of relative genetic distance on income differences is not only positive, but also larger than the effect of absolute genetic distance:

Proposition 4. *The expected income difference $E\{|y_j - y_i|\}$ between societies i and j is increasing in the two populations' relative genetic distance from the frontier $R(i,j)$. The effect of relative genetic distance $R(i,j)$ on income differences is larger than the effect of absolute genetic distance $G(i,j)$.*

Derivation of Proposition 4:

The expected income difference between pairs of populations at relative genetic distance $R(i,j) = g_s$ is[9]:

$$E\{|y_j - y_i| \mid R(i,j) = g_s\}| = E\{|y_f - y_{1.2}|\} + E\{|y_f - y_{1.3}|\} = \beta\Delta\varepsilon, \tag{3.17}$$

[9] We use the result, derived above, that all expected income differences between siblings are equal to $\beta\Delta\varepsilon$.

while the expected income difference between pairs of populations at relative genetic distance $R(i,j) = g_c$ is[10]:

$$E\{|y_j - y_i| \mid R(i,j) = g_c\}| = \frac{1}{3}\sum_{n=1}^{3} E\{|y_f - y_{2.n}|\} = \frac{3\beta\Delta\varepsilon}{2}. \tag{3.18}$$

Therefore, the effect of an increase of relative genetic distance from g_s to g_c is

$$E\{|y_j - y_i| \mid R(i,j) = g_c\} - E\{|y_j - y_i| \mid R(i,j) = g_s\} = \frac{\beta\Delta\varepsilon}{2} > \frac{\beta\Delta\varepsilon}{3} > 0. \tag{3.19}$$

The effect is positive ($\frac{\beta\Delta\varepsilon}{2} > 0$), and larger than the analogous effect of absolute genetic distance ($\frac{\beta\Delta\varepsilon}{3}$), derived above.

By the same token, the effect of relative genetic distance on expected income differences is also positive when moving from $R(i,j) = g_c - g_s$ to $R(i,j) = g_c$:

$$E\{|y_j - y_i| \,||\, R(i,j) = g_c\} - E\{|y_j - y_i| \,||\, R(i,j) = g_c - g_s\} = \frac{3\beta\Delta\varepsilon}{2} - \frac{5\beta\Delta\varepsilon}{4} = \frac{\beta\Delta\varepsilon}{4} > 0. \tag{3.20}$$

The results above are intuitive. As we increase relative genetic distance from the frontier, the expected income gap increases. The size of the effect is a positive function of the extent of divergence in vertically transmitted traits (ε), the extent to which this divergence constitutes a barrier to the horizontal diffusion of innovations (β), and the size of the improvement in productivity at the frontier (Δ).

In summary, our model has the following testable implications, which are brought to the data in the empirical analysis carried in the rest of this chapter:

1. *Relative genetic distance from the frontier population is positively correlated with differences in income per capita.*
2. *The effect on income differences associated with relative genetic distance from the frontier population is larger than the effect associated with absolute genetic distance.*

3.2.4 A Dynamic Extension

In the stylized model above, for simplicity we assumed that only one big innovation took place at time h. We now present a dynamic example, where innovations take place over time, and innovation and imitation are modeled endogenously.[11] The key insights and results carry over to this extension.

[10] We use the result, derived above, that the expected income difference between the frontier and each of its cousin populations is $\frac{3\beta\Delta\varepsilon}{2}$.

[11] The model builds heavily on Barro and Sala-i-Martin (1997, 2003) and Spolaore and Wacziarg (2012a).

In this dynamic example, we assume for simplicity, that populations do not change in modern times and have fixed size (normalized to one). More importantly, we assume that their inherited vertical traits do not change over the relevant time horizon. This is a reasonable simplification, because changes in vertical traits tend to take place much more slowly and at a longer horizon than the spread of technological innovations, especially if we focus on modern economic growth. Adding small random shocks to vertical traits after time h would significantly complicate the algebra, but would not affect the basic results.

Consider our six populations ($i = 1.1, 1.2, 1.3, 2.1, 2.2, 2.3$), with vertical traits inherited from their ancestral populations as described above, and unchanged in modern times (i.e. for $t \geq h$). Time is continuous. Consumers in economy i at time t maximize:

$$U_i(t) = \int_s^{\infty} \ln c_i(s) e^{-\rho(t-s)} ds, \tag{3.21}$$

under a standard budget constraint, where $c_i(t)$ is consumption, and $\rho > 0$ is the subjective discount rate. We assume that the six economies are not financially integrated, and that each economy i has its own real interest rate, denoted by $r_i(t)$. Hence, the optimal growth rate of consumption in society i is:

$$\frac{dc_i}{dt}\frac{1}{c_i(t)} = r_i(t) - \rho. \tag{3.22}$$

The production function for final output $y_i(t)$ is:

$$y_i(t) = \int_0^{A_i(t)} [x_{zi}(t)]^{\alpha} dz, \quad 0 < \alpha < 1, \tag{3.23}$$

where $x_{zi}(t)$ is the quantity of intermediate good of type z employed at time t in economy i, and the interval $[0, A_i(t)]$ measures the continuum of intermediate goods available in economy i at time t. Each intermediate good is produced by a local monopolist.

As before, without loss of generality we assume that society 1.1 is the technological frontier ($f = 1.1$). In this setting, this means that $A_f(h) > A_i(h)$ for all $i \neq f$. However, unlike in the previous analysis, innovation at the frontier economy now takes place endogenously. Following Barro and Sala-i-Martin (1997, 2003, Chapters 6 and 8), we assume that the inventor of input z retains perpetual monopoly power over its production within the frontier economy. The inventor sells the intermediate good at price $P_z = 1/\alpha$, earning the profit flow $\pi = (1-\alpha)\alpha^{(1+\alpha)/(1-\alpha)}$ at each time t.

The cost of inventing a new intermediate good at the frontier is η units of final output. Free entry into the innovation sector implies that the real interest rate $r_f(t)$ must be equal to π/η, which is assumed to be larger than ρ, therefore implying that consumption grows at the constant rate:

$$\gamma \equiv \frac{\pi}{\eta} - \rho > 0. \tag{3.24}$$

Output $y_f(t)$ and the frontier level of intermediate goods $A_f(t)$ will also grow at the rate γ.

The other populations cannot use the intermediate goods invented in economy f directly, but, as in Barro and Sala-i-Martin (1997), must pay an imitation cost μ_i in order to adapt those intermediate goods to local conditions. Our key assumption is that the imitation costs are increasing in the distance in vertical traits between the imitator and the frontier. Specifically, we assume that society i's imitation cost is:

$$\mu_i(t) = \lambda e^{\theta V(i,f)} \left(\frac{A_i(t)}{A_f(t)}\right)^{\xi}. \tag{3.25}$$

This is an instance of our general idea: a higher $V(i,f)$ is associated with higher imitation costs, because differences in vertical traits between the imitator and the inventor act as barriers to adoption and imitation. The parameter θ captures the extent to which dissimilarity in vertical traits between imitator and inventor increases imitation costs. For a given vertical distance, an imitator in society i faces lower imitation costs when there is a larger set of intermediate goods available for imitation—that is, when $A_i(t)/A_f(t)$ is low. The rationale for this assumption is the usual one: intermediate goods that are easier to imitate are copied first. Hence, the parameter $\xi > 0$ captures this advantage from technological backwardness. Our perspective may indeed shed some light on whether backward economies face higher or lower imitation costs overall, an issue debated in the literature (for instance, see Fagerberg, 2004). As we will see, our model predicts that, in steady state, societies that are farther technologically, and should therefore face lower imitation costs for this reason (captured by the parameter ξ), are also farther in terms of vertical distance from the frontier, and hence should face higher imitation costs through this channel (captured by the parameter θ), with conflicting effects on overall imitation costs.

Again, we assume that an imitator who pays cost $\mu_i(t)$ to imitate good k has perpetual monopoly power over the production of that input in economy i, and charges $P_k = 1/\alpha$, earning the profit flow $\pi = (1-\alpha)\alpha^{(1+\alpha)/(1-\alpha)}$, while output is proportional to available intermediate goods $A_i(t)$ in equilibrium: $y_i(t) = \alpha^{2\alpha/(1-\alpha)} A_i(t)$. With free entry into the imitation sector, economy i's real interest rate in equilibrium is[12]:

$$r_i(t) = \frac{\pi}{\mu_i(t)} + \frac{d\mu_i}{dt}\frac{1}{\mu_i(t)}. \tag{3.26}$$

In steady state, the level of imitation costs μ_i^* is constant. The number of intermediate goods, output, and consumption in all economies grow at the same rate γ as at the frontier. Therefore, in steady state the real interest rates in all economies are identical and equal to $\frac{\pi}{\eta}$, and imitation costs are identical for all imitators, which implies:

[12] See Barro and Sala-i-Martin (1997, 2003) for the details of the derivation.

Proposition 2bis. *The difference in log of income per capita in steady state* $|\ln y_i^* - \ln y_j^*|$ *between society i and society j is a function of their relative vertical distance from the frontier* $|V(i,f) - V(j,f)|$[13]:

$$|\ln y_i^* - \ln y_j^*| = \frac{\theta}{\xi}|V(i,f) - V(j,f)|. \tag{3.27}$$

The intuition of the above equation is straightforward: long-term differences in total factor productivity and output between societies are an increasing function of their relative cost to imitate, which depends on their relative vertical distance from the frontier. Therefore, societies that are more distant from the frontier in terms of vertically transmitted traits will have lower incomes per capita in steady state.

This dynamic model confirms the key implications of the simplified model that we had presented before. In particular, the equivalents of Propositions 3 and 4 hold in this setting as well, as long as one substitutes income differences $|y_j - y_i|$ with differences in log of income per capita in steady state $|\ln y_i^* - \ln y_j^*|$, and $\beta\Delta$ with $\frac{\theta}{\xi}$. We can then re-interpret those results as implying that societies at different relative genetic distance from the technological frontier will have different levels of income per capita in steady state. The effect of relative genetic distance on the income gap is larger when differences in vertical traits are associated with higher imitation costs (higher θ). Interestingly, we also have that the effect of relative genetic distance on income differences is lower when there are larger benefits from technological backwardness (higher ξ). In sum, the effects of relative genetic distance on economic development extend to this dynamic setting.

3.3. EMPIRICAL METHODOLOGY AND DATA

3.3.1 Specification and Estimation

The starting points for our empirical investigation into the long-term barriers to economic development are Propositions 3 and 4. These theoretical results show that if differences in vertical traits act as barriers to the diffusion of technologies, then differences in measures of development or technological sophistication across pairs of countries should (1) be correlated with the absolute genetic distance between these countries, (2) be correlated more strongly with their genetic distance relative to the technological frontier, and (3) genetic distance relative to the frontier should trump simple genetic distance between two countries. Whether these patterns hold true constitutes an empirical test of the barriers model. Denote by D_i a measure of development or technological sophistication in country i. We will consider alternatively per capita income (for the modern period), population density (for the pre-industrial period), and direct measures of technology use,

[13] Of course, we also have $|\ln A_i^*(t) - \ln A_j^*(t)| = |\ln y_i^*(t) - \ln y_j^*(t)|$

to be further detailed below. Denote by FST_{ij}^{W} the absolute genetic distance between countries i and j. Analogous to the theoretical definition, genetic distance relative to the frontier country is defined as: $FST_{ij}^{R} = |FST_{if}^{W} - FST_{jf}^{W}|$ where f denotes the frontier country.

Then the empirical predictions of Propositions 3 and 4 lead to the following empirical specifications:

$$\left|D_i - D_j\right| = \alpha_0 + \alpha_1 FST_{ij}^{R} + \alpha_2' X_{ij} + \varepsilon_{ij}^{\alpha}, \tag{3.28}$$

$$\left|D_i - D_j\right| = \beta_0 + \beta_1 FST_{ij}^{W} + \beta_2' X_{ij} + \varepsilon_{ij}^{\beta}, \tag{3.29}$$

$$\left|D_i - D_j\right| = \gamma_0 + \gamma_1 FST_{ij}^{R} + \gamma_2 FST_{ij}^{W} + \gamma_3' X_{ij} + \varepsilon_{ij}^{\gamma}, \tag{3.30}$$

where X_{ij} is a vector of control variables, primarily composed of alternative sources of barriers to diffusion, primarily geographic barriers. The predictions of our model are that $\alpha_1 > \beta_1, \gamma_1 > 0$, and $\gamma_2 = 0$.

Equations (3.28)–(3.30) are estimated using least squares. However, an econometric concern arises from the construction of the left-hand side variable as the difference in development or technological sophistication across country pairs. Indeed, consider pairs (i, j) and (i, k). By construction, the log per capita income of country i appears in the difference in log per capita incomes of both pairs, introducing some spatial correlation in the error term. To deal with this issue, we correct the standard errors using two-way clustering, developed by Cameron et al. (2006). Specifically, standard errors are clustered at the level of country 1 and country 2. This results in larger standard errors compared to no clustering.[14]

We complement these tests with additional empirical results that can shed light on our barriers interpretation of the effect of genetic distance. In particular, we examine the evolution of the effect of genetic distance through time. If genetic distance continues to have an effect on differences in economic performance in periods where the world distribution of income was very different, it should put to rest the idea that vertically transmitted traits bear direct, unchanged effects on productivity. We therefore examine the effects of genetic distance on population density in the pre-industrial era, going as far back as year 1. In Malthusian times, population density is the proper measure of overall technological sophistication, since per capita income gains resulting from innovation are only transitory, and soon dissipated by an increase in fertility (Ashraf and Galor, 2011 provide empirical evidence on this point). We also study the time path of the effect of genetic distance around the Industrial Revolution. Our model predicts that this effect should peak during the initial phases of the diffusion of the Industrial Revolution, as

[14] In past work, we employed various methods to deal with the spatial correlation that arises as a by-product of the construction of the left-hand side variable, such as including a set of common country dummies. The results were not sensitive to the method used to control for spatial correlation. See Spolaore and Wacziarg (2009) for further details.

only places close to its birthplace have adopted the new innovation. The model predicts that the effect should decline thereafter, as more and more societies adopt industrial and post-industrial modes of production.

3.3.2 Data

3.3.2.1 Genetic Distance Data

Our source for genetic distance data is Cavalli-Sforza et al. (1994). The main dataset covers 42 ethnolinguistic groups samples across the globe.[15] The genetic data concerns 120 gene locus, for which allele frequencies were obtained by population. The gene locus were chosen to represent neutral genes, i.e. genes that did not spread through natural selection but through random drift, as determined by geneticists. Thus, when aggregated over many genes, measures of genetic distance obtained from neutral genes capture separation times between populations, precisely the analog of genealogical distance employed in our theoretical model.

The specific measure of genetic distance we use is known as F_{ST} genetic distance, also known as Wright's fixation index.[16] To illustrate the index, we derive it for the specific case of two populations, one locus and two alleles. The number of individuals in population i is n_i. Total population is $n = \sum_{i=1}^{2} n_i$. The share of population i is $w_i = n_i/n$. Consider one locus with two possible alleles: either Q or q. Let $0 \leq p_i \leq 1$ be the frequency of individuals in population i with allele Q. Let p be this frequency in the whole population $\left(p = \sum_{i=1}^{2} w_i p_i\right)$. The degree of heterozygosity (i.e. the probability that two randomly selected alleles within a population are different) within population i is $H_i = 2p_i(1 - p_i)$, and average heterozygosity across populations is $H_S = \sum_{i=1}^{2} w_i H_i$. Heterozygosity for the whole population is $H_T = 2p(1 - p)$. Then Wright's fixation index, F_{ST}, is defined as:

$$F_{ST} = 1 - \frac{H_S}{H_T} = 1 - \frac{n_1 p_1(1 - p_1) + n_2 p_2(1 - p_2)}{np(1 - p)}. \tag{3.31}$$

This is one minus the ratio of group level average heterozygosity to total heterozygosity. If both populations have the same allele frequencies ($p_1 = p_2$), then $H_i = H_S = H_T$, and $F_{ST} = 0$. In the polar opposite case, individuals within each population all have the same alleles, and these alleles differ completely across groups ($p_1 = 1 - p_2$). Then $F_{ST} = 1$ (total fixation). In general, the higher the differences in allele frequencies across populations, the higher is F_{ST}. The formula can easily be extended to account for more than two alleles. F_{ST} can be averaged in a variety of ways across loci, so that the resulting F_{ST} distance is a summary measure of relatedness between the two populations. Moreover, bootstrapping techniques can be used to obtain standard errors on estimates of F_{ST}. Details of these

[15] We will also make use of a more detailed dataset covering 26 European populations. Since populations were sampled at the country level rather than at the ethnic group level for the European dataset, matching populations to countries was an easier task.

[16] In past work, we also used the Nei index. Results did not hinge on the use of either index.

extensions are provided in Cavalli-Sforza et al. (1994, pp. 26–27). We rely on the genetic distance data that they provide, i.e. we rely on population geneticists' best judgment as to the proper choice of alleles, the proper sampling methods, and the proper way to aggregate heterozygosity across alleles.

The genealogical tree of human populations is displayed in Figure 3.2, where the genetic distance data was used to construct a tree showing the successive splits between human populations over the course of the last 70,000 years or so. In this figure, recent splits indicate a low genetic distance between the corresponding populations. In the source data pertaining to 42 world populations, the largest F_{ST} genetic distance between any two populations is between the Mbuti Pygmies and the Papua New Guineans ($F_{ST} = 0.4573$). The smallest is between the Danish and the English ($F_{ST} = 0.0021$).

Genetic distance is obtained at the level of populations but it was necessary to construct measures pertaining to countries. We matched ethnolinguistic groups in Cavalli-Sforza et al. (1994) to ethnic groups for each country using the ethnic group data from Alesina et al. (2003), and then constructed the expected distance between two individuals, each drawn randomly from each of the two countries in a pair. Thus, our baseline measure of genetic distance between countries 1 and 2 is:

$$FST_{12}^{W} = \sum_{i=1}^{I} \sum_{j=1}^{J} (s_{1i} \times s_{2j} \times FST_{ij}), \tag{3.32}$$

where s_{1i} is the share of population i in country 1, s_{2j} is the share of population j in country 2, and FST_{ij} is genetic distance between population i and j. This index is also known as the Greenberg index (after Greenberg, 1956), and is increasingly used in economics as a measure of ethnolinguistic heterogeneity (see for instance Bossert et al. 2011).[17]

The measure derived above, FST_{12}^{W}, is the absolute measure of expected distance between any two countries 1 and 2. In keeping with the theoretical definition, we can also define a measure of these countries' relative distance to the technological frontier f:

$$FST_{12}^{R} = \left| FST_{1f}^{W} - FST_{2f}^{W} \right|. \tag{3.33}$$

Finally, the procedure above matches populations to ethnolinguistic groups as they occur in the contemporary period. It is, however, also possible to calculate genetic distance as of the year 1500 AD, by matching populations to the plurality group in each country given their composition in 1500. Thus, for instance, in the 1500 match, Australia is matched to the Aborigenes population (while for the contemporary period Australia is matched to a combination of English and Aborigenes—predominantly the former).

[17] In past work we also used the genetic distance between the largest populations (i.e. genetic groups) in countries 1 and 2. The correlation between expected (weighted) genetic distance and this alternative index is very high, and it does not matter which one we use in our empirical work.

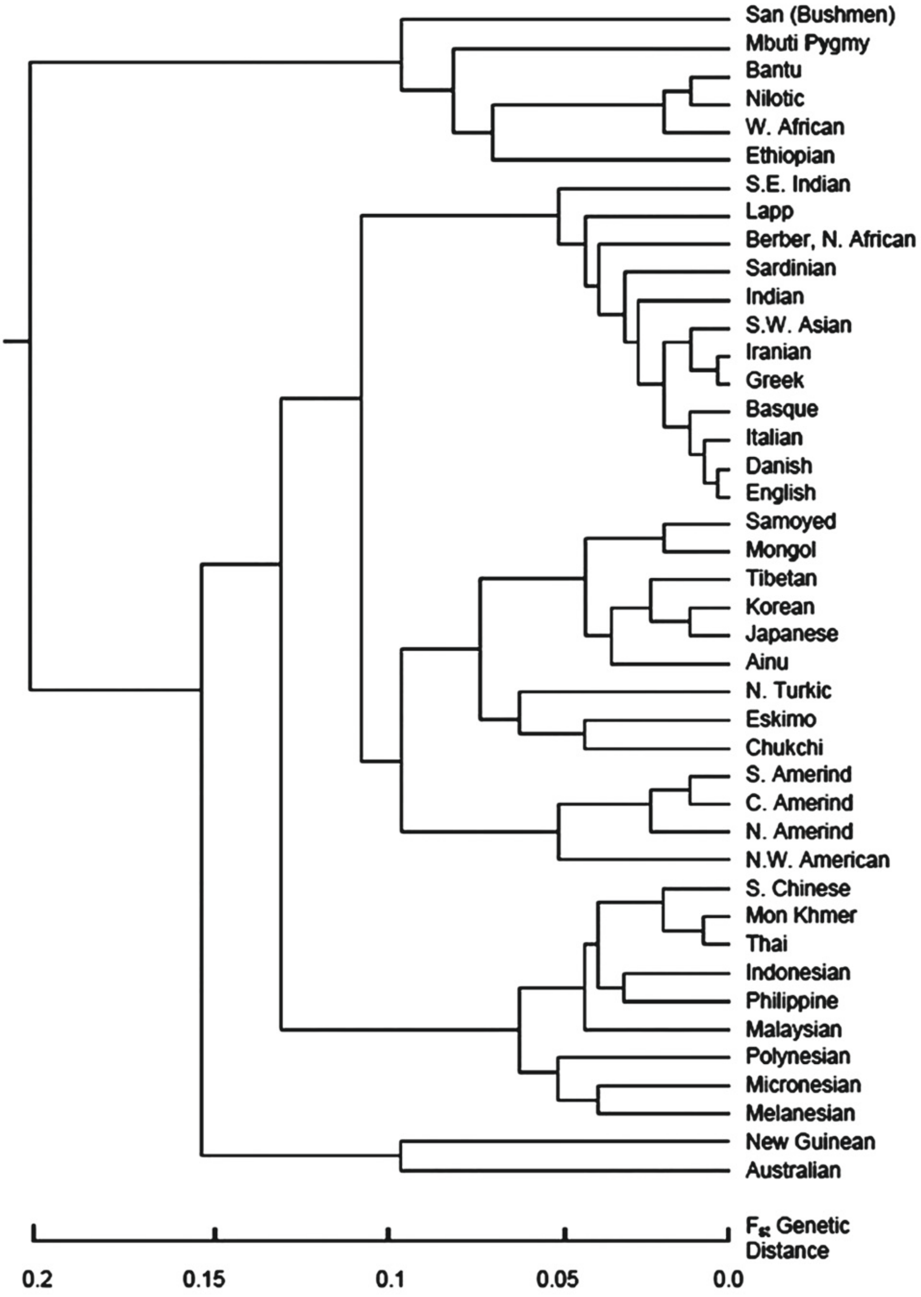


Figure 3.2 Genetic distance among 42 populations. *Source: Cavalli-Sforza et al. 1994*.

We make use of the 1500 match in some historical regressions, or as an instrument for contemporary genetic distance. Again, measures of absolute and relative genetic distance are computed using the 1500 match of populations to countries.

3.3.2.2 Measures of Development and Technological Sophistication

We use a variety of measures of differences in economic development and technological sophistication. The first set of measures is defined at an aggregate level. The primary measure for the contemporary period is the absolute difference in log per capita income in 2005 (from the Penn World Tables version 6.3). For the pre-industrial periods, we consider the absolute difference in population density. The population density data pertains to the year 1500, and the source is McEvedy and Jones (1978). Despite more limited geographic coverage, we also use data on per capita income going back to 1820, from Maddison (2003), in order to examine the time path of the effect of genetic distance around the time of the Industrial Revolution.

The second set of measures includes disaggregated measures of technology usage, either along the extensive margin (for the historical period) or along the intensive margin (for the contemporary period).[18] We rely mostly on data from Comin et al. (2010, henceforth CEG). CEG gathered data on the degree of technological sophistication for the years 1000 BC, 1 AD, 1500 AD, and the contemporary period (1970–2000 AD). We make use of the data for 1500 AD and the contemporary period, since this corresponds most closely to the available genetic distance data. The data for 1500 pertain to the extensive margin of adoption of 24 separate technologies, grouped into 5 categories: military, agricultural, transportation, communication, and industry. For each technology in each category, a country is given a score of 1 if the technology was in use in 1500, 0 otherwise. The scores are summed within categories, and rescaled to vary between 0 and 1. An overall index of technological sophistication is also obtained by taking the simple average of the technological index for each of the 5 categories.

For the 1970–2000 AD data, technology usage is measured along the intensive margin. The basic data covers the per capita usage intensity of nine technologies, obtained from the database of Comin et al. (2008). For each technology, a country's usage is characterized as the number of years since the technological frontier (the United States) had the same level of per capita usage. The index is then rescaled to vary from 0 to 1, where 1 denotes usage at the same level as the frontier. Technologies are aggregated into 4 of the 5 aforementioned categories (all except the military category), and a simple average of the four measures is also available.

Finally, we attempted to measure technological sophistication at a more disaggregated level. This allows for a more refined analysis based on individual technologies that were not aggregated into broader categories, as is the case in the CEG dataset. For this, we relied

[18] These technologies are listed in Appendix 1.

on the CHAT dataset (Comin and Hobijn, 2009), which contains data on usage of 100 technologies. We restricted attention to technologies for which data is available for at least 50 countries over the 1990–1999 period. This led to a restricted set of 33 technologies, covering a wide range of sectors—medical, transportation, communications, industrial, and agricultural technologies. For each of the underlying 33 technologies, we calculated usage per capita, in order to maintain a consistent definition of the intensity of use.[19] For instance, for the technology "personal computers," the dependent variable is the absolute difference, between country i and country j, in the number of computers per capita. For all technologies, the technological leader was assumed to be the United States, an assumption confirmed in virtually all cases when examining the actual intensity of use.

All of these measures of technological sophistication were available at the country level, so we computed the absolute difference in technology measures across all available pairs of countries for the purpose of empirical analysis.

3.3.2.3 *Measures of Geographic Barriers*

Measures of genetic distance are correlated with geographic distance. Indeed, homo sapiens is estimated to have migrated out of East Africa around 70,000 years ago, and from there spread first to Asia, and then later fanned out to Europe, Oceania, and the Americas. As early humans split into subgroups, the molecular clock of genetic drift started operating, and populations became more genetically distant. It is not surprising that the farther in space, the more genetically distant populations are expected to be. It is therefore important to control for geographic distance when estimating the human barriers to the diffusion of innovations. At the same time, as we describe below, the correlation between geographic distance and genetic distance is not as large as one might expect. This is the case for two major reasons: First, genetic drift occurred along rather specific geographic axes. For instance, a major dimension along which populations array themselves in proportion to their genetic distance is a rough straight line between Addis Ababa and Beijing. There need not be a strict correspondence, then, between genetic distance and common measures of geographic distance relevant as geographic barriers to the spread of innovations, such as the greater circle distance or latitudinal distance. Second, more recent population movements have served to break the initial links between geographic distance and genetic distance. Two highly relevant population movements were the conquests of parts of the New World by Europeans, and the slave trades occurring thereafter. We obtain some (but not all) of our identifying variation off of these post-1500 population movements.

To capture geographic distance we use a large array of controls, capturing both simple geodesic distance, distance along the longitudinal and latitudinal dimensions, and binary indicators of micro-geography such as whether the countries in a pair are contiguous, are

[19] One exception was for the share of cropland area planted with modern varieties, for which it would make little sense to divide by population. All other technologies were entered in per capita terms.

islands, are landlocked, or share a common sea or ocean. This set of controls was included in every regression, and was supplemented in robustness tests by additional geographic controls such as climatic differences, continent effects, and freight costs.

3.3.2.4 Summary Statistics and Data Patterns

Figure 3.3 presents a simple plot of weighted genetic distance to the USA against per capita income, and Figure 3.4 does the same after partialling out the effect of geodesic distance (a similar figure is obtained after partialling out the effect of a longer list of geographic distance metrics). Both figures reveal a negative association between per capita income and genetic distance to the USA. Table 3.1 presents summary statistics to help in the interpretation of regression estimates. Panel B displays correlations based on 10,440 country pairs, based on 145 countries. These correlations are informative: the absolute genetic distance between pairs bears a correlation of 19.5% with the absolute difference in log per capita income. Genetic distance relative to the USA, however, bears a much larger correlation of 32.26%, a pattern consistent with the predictions of the barriers model, implying a larger effect of relative genetic distance compared to absolute genetic distance. Finally, as mentioned above, the correlation between genetic distance (either relative to the frontier or not) with geodesic distance, is positive but moderate in magnitude, offering hope that the effect of genealogical barriers can be estimated separately from that of geographic barriers.

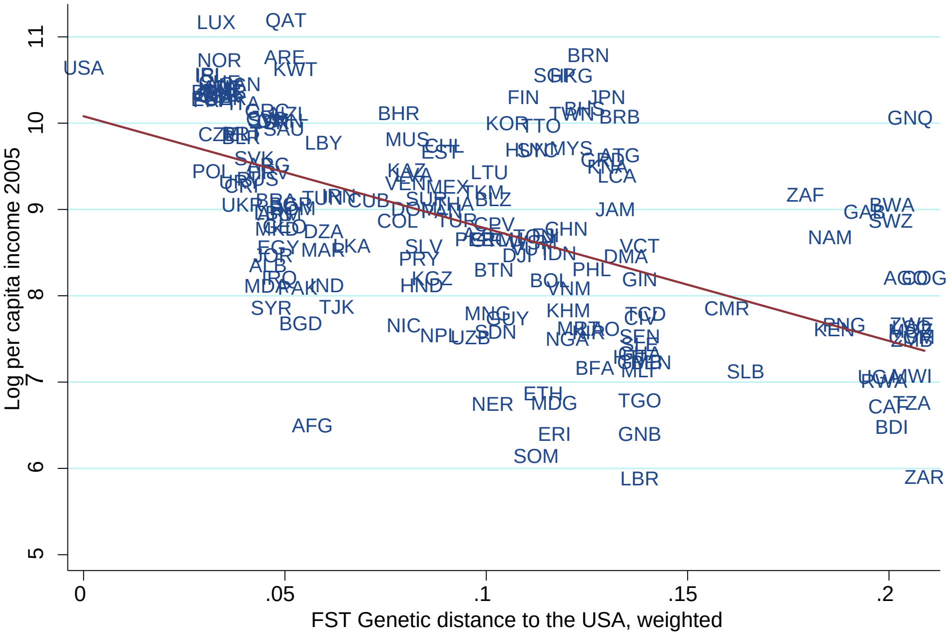


Figure 3.3 Log income in 2005 and genetic distance to the USA.

Table 3.1 Summary statistics for the main variables of interest

Variable	Mean	Standard deviation	Minimum	Maximum
Panel A—mean and variation				
Difference in log income per capita 2005	1.3844	0.9894	0.0000241	4.8775
FST genetic distance relative to the English, 1500	0.0710	0.0555	0	0.2288
Weighted FST genetic distance relative to the USA	0.0612	0.0475	0	0.2127
Weighted FST genetic distance between pairs	0.1124	0.0818	0	0.3364
Geodesic distance (thousands of km)	7.1349	4.1330	0.0105	19.9512

10,440 observations.

	Difference in log income per capita 2005	FST genetic distance relative to the English, 1500	Weighted FST gen. dist. relative to the USA	Weighted FST genetic distance between pairs
Panel B—correlations				
FST genetic distance relative to the English, 1500	0.2745*	1		
Weighted FST genetic distance relative to the USA	0.3226*	0.6105*	1	
Weighted FST genetic distance between pairs	0.1950*	0.2408*	0.5876*	1
Geodesic distance (thousands of km)	0.0126	0.0644*	0.0899*	0.3317*

* Significant at the 5% level. 10,440 observations.

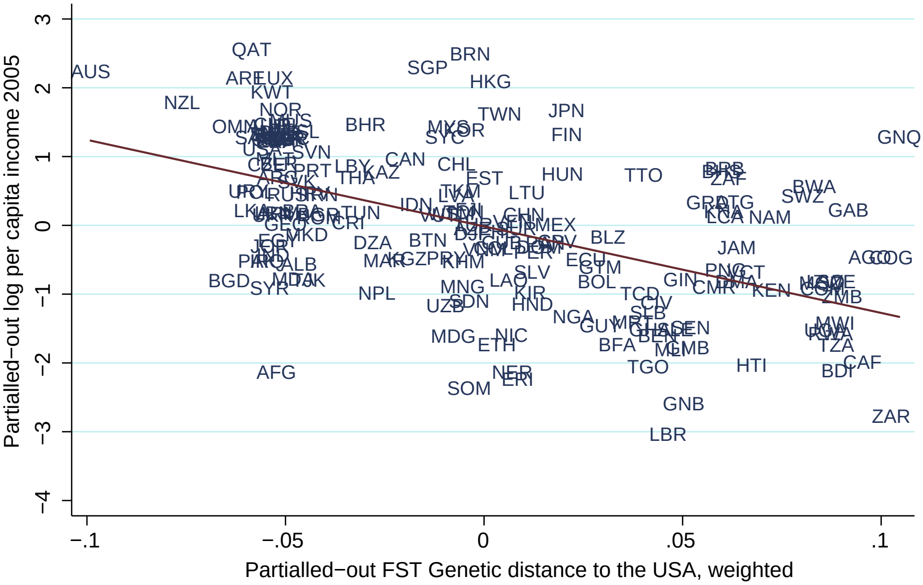


Figure 3.4 Log income in 2005 and genetic distance to the USA, partialling out geodesic distance to the USA.

3.4. BARRIERS TO DEVELOPMENT: EMPIRICAL RESULTS

3.4.1 Results for Aggregate Measures of Economic Development

3.4.1.1 Baseline Estimates

Baseline estimates of Equations (3.28)–(3.30) are presented in Table 3.2. The predictions of the barriers model are borne out: after controlling for various measures of geographic distance, differences in per capita income are significantly correlated with both absolute and relative genetic distance (columns 1 and 2).[20] However, the magnitude of the effect of genetic distance relative to the technological frontier (column 1) is about three times as large as the effect of absolute genetic distance (column 2). This is true when comparing both the estimated coefficient and a standardized measure of magnitude (the standardized beta, reported in the next to last row of Table 3.2). When including both measures in the regression (column 3), genetic distance relative to the frontier remains significant while absolute genetic distance becomes insignificantly different from zero. In terms of magnitudes, a one standard deviation increase in F_{ST} genetic distance relative to the USA

[20] A myriad additional controls were included as robustness tests in analogous regressions presented in Spolaore and Wacziarg (2009). These included climatic differences, freight costs, etc. Results were robust to the inclusion of these additional control variables.

Table 3.2 Income difference regressions (dependent variable: difference in log per capita income, 2005)

	(1) OLS with relative GD	(2) OLS with simple GD	(3) Horserace between simple and relative GD	(4) 2SLS with 1500 GD
FST gen. dist. relative to the USA, weighted	6.290		6.029	9.720
	(1.175)***		(1.239)***	(1.974)***
FST genetic distance		2.164	0.275	
		(0.596)***	(0.541)	
Absolute difference in latitudes	0.232	0.559	0.255	0.152
	(0.245)	(0.279)**	(0.248)	(0.300)
Absolute difference in longitudes	−0.025	−0.196	−0.007	0.238
	(0.220)	(0.240)	(0.213)	(0.247)
Geodesic distance	−0.012	−0.008	−0.016	−0.042
	(0.026)	(0.027)	(0.025)	(0.028)
=1 for contiguity	−0.418	−0.495	−0.414	−0.326
	(0.060)***	(0.060)***	(0.061)***	(0.069)***
=1 if either country is an island	0.174	0.143	0.174	0.211
	(0.083)**	(0.083)*	(0.083)**	(0.084)***
=1 if either country is landlocked	0.008	0.024	0.005	−0.029
	(0.085)	(0.090)	(0.087)	(0.085)
=1 if pair shares at least one sea or ocean	−0.001	0.028	−0.000	−0.024
	(0.067)	(0.077)	(0.067)	(0.078)
Constant	1.022	1.143	1.017	0.891
	(0.089)***	(0.086)***	(0.090)***	(0.099)***
Standardized Beta (%)	30.18	10.39	28.93	46.49
R-Squared	0.11	0.07	0.11	0.09

Two-way clustered standard errors in parentheses.
All regressions are based on 10,440 observations.
* Significant at 10%.
** Significant at 5%.
* * * Significant at 1%.

is associated with an increase in the absolute difference in log income per capita of almost 29% of that variable's standard deviation.

Column 4 of Table 3.2 reports results of IV estimation, using relative genetic distance to the English population in 1500 as an instrument for current genetic distance to the USA. This is meant to address two specific concerns: First, matching the 42 populations for which genetic distance data is available to contemporaneous ethnolinguistic groups may introduce measurement error. The main difficulties in the match arise for the New World where it is sometimes difficult to assess which European population to match with the descendents of past European settlers; which African populations to match with former slaves; and what shares to ascribe to these various populations in the total population, given that many of them mixed over time, resulting in significant shares of populations with mixed ancestry (the latter issue arises mainly in Latin America). In contrast, the 1500 match of genetic groups (populations) to the plurality ethnic group is much more straightforward, since the Cavalli-Sforza et al. (1994) data was gathered precisely to represent the makeup of countries as they stood in 1492, prior to the population movements associated with the conquest of the New World. The second concern is endogeneity: genetic distance between countries changed in the post-1492 era due to the aforementioned conquest of the New World and the slave trades. It is possible that areas well suited to high incomes in the industrial era, perhaps due to geographic factors such as a temperate climate, happened to attract certain populations (for instance Europeans) as settlers. In this case, it would be the potential for differential incomes that would causally affect genetic distance rather than the opposite. Using genetic distance lagged by 500 years as an instrument addresses this particular endogeneity concern. The results presented in column 4, show that, if anything, OLS understated the effect of relative genetic distance: its standardized effect rises under IV to 46.49%. Since the IV estimates are larger than the OLS estimates, to remain conservative we rely in the rest of this chapter on OLS estimates.

3.4.1.2 Regional Controls and Analysis

In Table 3.3, we run a variety of regressions accounting for regional effects. In column 1, we include a full set of continental dummy variables capturing both whether the countries in a pair are both located on the same specific continent (an effect presumed to go in the direction of reducing the difference in economic performance between these countries); and whether they are located on different ones (as further defined in the footnote to Table 3.3). The idea behind this test is to further control for geographic factors not already captured by the included geographic distance variables. However, this is a demanding test, since continent effects could capture geographic barriers but also part of the effect of human barriers that could be mismeasured when using genetic distance. Nonetheless the effect of genetic distance remains robust to controlling for a full set of

Table 3.3 Income difference regressions, regional controls, and sample splits (dependent variable: difference in log per capital income in 2005, 1870 for column 3)

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	Continent dummies	**Europe 2005 income**	**Europe with 1870 income**	**Excluding Europe**	**Control for Europeans**	**Excluding SS Africa**
Fst gen. dist. relative to the USA, weighted	3.403			5.183	5.624	4.851
	(1.284)**			(1.232)***	(1.143)***	(1.443)***
Genetic distance, relative to the English		25.920	27.054			
		(11.724)**	(6.557)***			
Abs. difference in the shares of people of European descent					0.626	
					(0.125)***	
Constant	1.541	0.345	0.495	1.006	0.864	0.853
	(0.315)**	(0.201)*	(0.154)***	(0.123)***	(0.097)***	(0.071)***
Observations	10,440	253	136	6,328	10,153	5,253
Standardized Beta (%)	16.27	31.28	43.62	24.99	27.15	17.12
R-Squared	0.20	0.24	0.24	0.08	0.17	0.06

Two-way clustered standard errors in parentheses.

In all regressions, controls are included for: Absolute difference in latitudes, absolute difference in longitudes, geodesic distance, dummy for contiguity, dummy if either country is an island, dummy if either country is landlocked, dummy if pair shares at least one sea or ocean.

Column 1 includes continental dummies defined as follows: both in Asia dummy, both in Africa dummy, both in Europe dummy, both in North America dummy, both in Latin America/Caribbean dummy, both in Oceania dummy, dummy if one and only one country is in Asia, dummy if one and only one country is in Africa, dummy if one and only one country is in Europe, dummy if one and only one country is in North America, dummy if one and only one country is in South America.

* Significant at 10%.

** Significant at 5%.

*** Significant at 1%.

12 same- and different-continent dummies. While the effect of genetic distance falls in magnitude, it remains large and highly significant statistically.

Columns 2 and 3 make use of the separate genetic distance dataset we have for 26 countries in Europe. Here, the relevant measure of genetic distance is F_{ST} distance to the English (England being the birthplace of the Industrial Revolution), though the results do not change if we use distance to the Germans instead. We find that within Europe, genetic distance is again a strong predictor of absolute differences in log per capita income. The standardized beta on genetic distance relative to the English is of the same order of magnitude as that found in the world sample, and it is highly significant. There are two major genetic clines in Europe: one separating the north and the south, another one separating the east and the west. These correspond to north-south and east-west income differences. Since the east-west cline overlaps to a large degree with regions that were on either side of the Iron Curtain during the Cold War, to assess whether this historical feature explains all of the effect of genetic distance on economic performance, we repeat our regression using income in 1870 (from Maddison), well prior to the rise of the Eastern bloc. We find that the effect of genetic distance is in fact larger in magnitude in the immediate aftermath of the Industrial Revolution, with the standardized beta rising to almost 44%. This result assuages concerns that the contemporary results were a result of the fact that the Iron Curtain as a first approximation, separated Slavic from non-Slavic Europeans. It is also highly consistent with the barriers story since, as we further explore below, the effect of genetic distance should be larger around the time of a large innovation, in the midst of the process whereby countries other than the frontier are busy adopting the frontier technology in proportion to how genetically far they are from the frontier. In sum, our effects hold within Europe, where genetic distance is better measured.

Since the basic result of this chapter holds so strongly for Europe, might Europe drive the World results? To test this, in column 4 we exclude any pairs of countries containing at least one European country. Compared to the baseline results, the standardized effect of genetic distance relative to the USA declines from 30% to 25%, but remains large and statistically significant—highlighting that the results are not due to Europe alone. To drive home the point, in column 5 we control for the absolute difference in the share of the population of European descent, using data from the Putterman and Weil (2010) migration matrix. The regression now controls more broadly for the effect of European-ness, and while the effect of the absolute difference in the share of Europeans is a positive and statistically significant determinant of differences in per capita income, its inclusion in the regression only moderately reduces the standardized effect of relative genetic distance (to 27%). We conclude that our results are not driven by the inclusion of European countries in the sample, nor are they driven by the genetic difference between Europeans and the rest.

The final geographic concern that we explore is whether Sub-Saharan Africa drives our results. As Figure 3.2 illustrates, Sub-Saharan African populations are genetically distant from the rest of the world: the out-of-Africa migrations occurring about 70,000 years ago were the first foray of modern humans out of Africa, and consequently Africans and other world populations have had the longest time to drift apart genetically from each other. Sub-Saharan populations also have some of the lowest per capita GDPs recorded in the world. While it is part of our story to ascribe some of the poverty of Africa to the barriers to technological transmission brought about by its high degree of genealogical distance from the rest of the world, it would be concerning if our results were entirely driven by Sub-Saharan Africa. To address this concern, in column (6) of Table 3.3 we exclude any pair that involves at least one Sub-Saharan country from our sample. We find that the effect of genetic distance falls a little, but remains positive, statistically significant, and large in magnitude with a standardized beta equal to 17%. Together with the strong results within Europe, this should lay to rest any notion that our results are driven solely by Sub-Saharan Africa.

3.4.1.3 Historical Analysis

We now turn to a historical analysis of the determinants of aggregate measures of economic performance, seeking to achieve two main goals. The first is to assess the robustness of the effect of genetic distance through time. The second goal is to describe the time path of the standardized effect of genetic distance around the time of the Industrial Revolution. In our barriers model, a major innovation such as the Industrial Revolution should lead to a specific pattern in the evolution of the effect of relative genetic distance on differences in economic development. Specifically, the effect of genetic distance should be large in the aftermath of the onset of the Industrial Revolution in the frontier country. As more and more societies adopt the Industrial Revolution, the effect should gradually decline. We now redefine the frontier country as the United Kingdom (i.e. the English population) since it is a more appropriate choice for the period concerned.[21]

Table 3.4 displays pairwise correlations between historical measures of differences in economic development and genetic distance. For the 1500 period, we consider the correlation between relative genetic distance to the English using the 1500 match, and population density. For periods from 1820 to today, it is best to rely on the correlation between contemporaneous weighted genetic distance relative to the UK, and the absolute difference in log per capita income at various dates.[22] A few remarks are in order: First,

[21] This choice is not very material. In fact, relative genetic distance to the English and relative genetic distance to the United States are very highly correlated, because the United States are primarily composed of populations from Western Europe—either the English or populations genetically very close to the English. In fact, by world standards, genetic distances between Western European populations are so small that it matters little empirically which Western European population is chosen as the frontier. For instance, for 1500 we experimented with using Italy as the frontier country; results were unchanged.

[22] We lack genetic distance data suitable for the millenia prior to 1500, despite the existence of some population density data for early dates. At any rate it is not clear that our barriers story would apply with

Table 3.4 Pairwise correlations between historical measures of economic development

	Relative genetic distance to the English, 1500	**Relative genetic distance to the UK, (contemporary)**	**Abs. difference in population density, 1500**	**Abs. difference in log income, 1820**	**Abs. difference in log income, 1870**	**Abs. difference in log income, 1913**	**Abs. difference in log income, 1960**
Relative genetic distance to the UK, weighted (contemporary)	0.6205* (10,585)	1 (10,585)					
Abs. difference in population density, 1500	0.1594* (10,153)	0.0461* (10,153)	1 (10,153)				
Abs. difference in log income, 1820	0.1763* (1,035)	0.1327* (1,035)	0.1701* (990)	1 (1,035)			
Abs. difference in log income, 1870	0.1360* (1,485)	0.1811* (1,485)	0.1125* (1,378)	0.6117* (1,035)	1 (1,485)		
Abs. difference in log income, 1913	0.0840* (1,653)	0.1839* (1,653)	0.0739* (1,540)	0.5411* (1,035)	0.8996* (1,485)	1 (1,653)	
Abs. difference in log income, 1960	0.2347* (4,753)	0.3229* (4,753)	0.1242* (4,560)	0.4018* (820)	0.6154* (1,035)	0.7201* (1,176)	1 (4,753)
Abs. difference in log income, 2005	0.2745* (10,440)	0.3228* (10,440)	0.1173* (10,011)	0.3297* (990)	0.4722* (1,431)	0.4844* (1,596)	0.6199* (4,753)

* Significant at the 5% level; # of obs. in parentheses.

this data reveals some persistence in economic fortunes. In spite of being different measures, even the absolute difference in population density in 1500 and the absolute difference in log per capita income in 2005 bear a correlation of about 12% with each other. Correlations between income-based measures are much higher (for instance, the correlation of income differences in 1820 and 2005 is 33%). Second, genetic distance is positively and significantly correlated with these measures of differences in economic performance at all dates. For instance, the correlation between the absolute difference in population density in 1500 and relative genetic distance to the English in 1500 is about 16%. This rises to 32% in 2005 (comparisons of magnitudes should be made cautiously from this table as the underlying samples differ by date—but in the case of 1500 and 2005 the samples are very similar—more on this point below). In general, simple correlations reveal that despite some changes in the relative fortunes of nations over the last 500 years, the correlation between genetic distance and development seems to exist at all dates.

Table 3.5 turns to regression analysis. Across all columns, corresponding to different dates, genetic distance relative to the UK comes out with a positive, significant coefficient. Thus, the effect of genetic distance is robust to considering different dates and a different measure of economic development for the Malthusian period. The penultimate row of Table 3.5 shows the evolution of the standardized effect of genetic distance over time for a common sample of 820 country pairs (41 countries), for which income data is available at all dates. The magnitudes here are somewhat smaller than when using unrestricted samples across periods, in part because the 41 countries only include one Sub-Saharan African country (and that country is South Africa, which is relatively rich). However, restricting the sample to pairs available at all dates allows for a comparison of magnitudes across time. To facilitate interpretation, the standardized effects from the common sample are displayed in Figure 3.5.

This figure lends further credence to the barriers model. Indeed, just as predicted above, the effect of genetic distance, which is initially modest in 1820, rises by around 75% to reach a peak in 1913, and thereafter declines. Thus, in the few decades following the adoption of the Industrial Revolution by countries in the (genetic) periphery of England, the effect of genetic distance was maximal. Thereafter, as more and more societies industrialized, the effect fell steadily.

3.4.2 Results for Specific Innovations

The analysis above concerns determinants of differences in aggregate productivity. This is useful to analyze very broad trends like the diffusion of the Industrial Revolution. Yet our model also applies to the diffusion of more specific technologies. Indeed, if our empirical results applied to aggregate measures of development or technological sophistication

as much force in periods where geographic barriers to the diffusion of innovation were so overwhelming, except perhaps in a regionally narrow context.

Table 3.5 Regressions using historical income data

	(1) Year 1500 population density	(2) Income 1820	(3) Income 1870	(4) Income 1913	(5) Income 1960	(6) Income 2005
Relative Fst genetic distance to the UK, 1500 match	29.751 (7.168)***					
Relative Fst genetic distance to the UK, weighted		0.671 (0.344)*	1.691 (0.836)**	1.984 (0.907)**	3.472 (0.783)**	5.075 (0.941)**
Constant	6.693 (0.981)***	0.313 (0.063)**	0.365 (0.076)**	0.421 (0.064)**	0.478 (0.077)**	1.017 (0.088)**
Observations	10,153	1,035	1,485	1,653	4,753	10,440
Standardized beta (%), maximal sample	17.77	8.75	15.02	15.02	28.82	30.58
Standardized beta, restricted sample*	-	9.89	16.30	17.36	11.15	7.49
R-Squared	0.07	0.22	0.16	0.17	0.17	0.11

Two-way clustered standard errors in parentheses.

In all regressions, controls are included for: Absolute difference in latitudes, absolute difference in longitudes, geodesic distance, dummy for contiguity, dummy if either country is an island, dummy if either country is landlocked, dummy if pair shares at least one sea or ocean.

Population density data for 1500 are from McEvedy and Jones (1978). Income data for 1820, 1870, and 1913 are from Maddison (2003). Income data for 1960 and 2005 are from the Penn World Tables.

* The restricted sample for columns (2)–(6) consists of 820 country pairs constructed from 41 countries (Algeria, Australia, Austria, Belgium, Brazil, Canada, China, Denmark, Egypt, Finland, France, Greece, India, Indonesia, Iran, Ireland, Italy, Jamaica, Japan, Jordan, Korea, Malaysia, Mexico, Morocco, Nepal, Netherlands, New Zealand, Norway, Philippines, Portugal, South Africa, Spain, Sri Lanka, Sweden, Switzerland, Syria, Taiwan, Thailand, Turkey, USA, United Kingdom).

* Significant at 10%.

** Significant at 5%.

*** Significant at 1%.

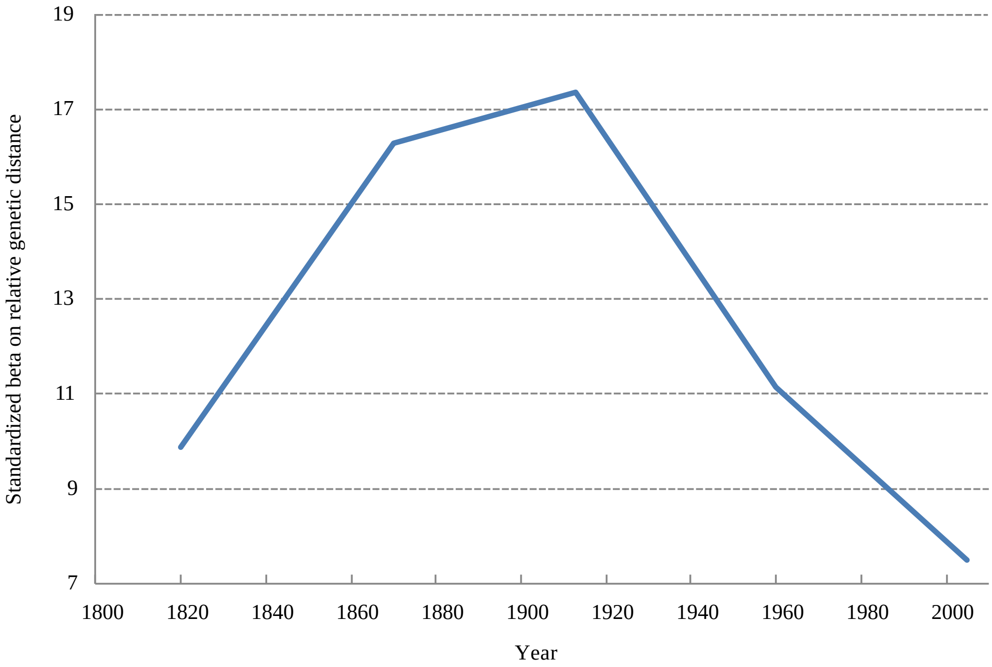


Figure 3.5 Standardized effect of genetic distance over time, 1820–2005.

only, but did not extend to more disaggregated technologies, it would cast doubt on the hypothesis that the main effect of genetic distance is to hinder the transmission of technologies across societies with very different cultures and histories. In this subsection, we use data directly at the technology usage level to address this issue.

Table 3.6 starts with some summary statistics from the CEG dataset, pertaining to the contemporary period. Panel A is mainly meant to assist in the interpretation of the regressions that come next, but Panel B already contains interesting information. The first observation is that differences in the intensity of technology usage in 1970–2000 across various technological categories are correlated, but imperfectly. Second, differences in technology usage intensity are positively correlated with per capita income, but the correlations are in the 0.4–0.7 range depending on the technological category, so these variables do not all measure the same thing. In other words, our measures of differences in technology usage are not simply indicators of differences in overall economic performance. Third, differences in technology usage are correlated more strongly with genetic distance relative to the frontier than with genetic distance per se. In fact, correlations with the latter are often close to zero while correlations with the former are always positive and significant.

Table 3.6 Summary Statistics for Genetic Distance and Technological Adoption Levels (from the Comin et al., 2010 Data)

Variable	# of Obs.	Mean	Standard deviation	Minimum	Maximum
Panel A—mean and variation					
Avg. tech. adoption in agriculture in 1970–2000	6,105	0.1998	0.2327	0	0.8553
Avg. tech. adoption in communications in 1970–2000	7,381	0.2601	0.1894	0	0.7911
Avg. tech. adoption in transportation in 1970–2000	6,441	0.1986	0.1600	0	0.8443
Avg. tech. adoption in industry in 1970–2000	5,565	0.3005	0.2153	0	1.0278
Avg. of the sectoral tech. adoption indexes in 1970–2000	7,503	0.2129	0.1807	0	0.8378
Absolute difference in log income, 2005	10,440	1.3844	0.9894	0	4.8775
FST gen. dist. relative to the USA, weighted	10,585	0.0611	0.0473	0	0.2127
Simple FST genetic distance	10,585	0.1126	0.0816	0	0.3364

(Continued)

Table 3.6 (*Continued*)

	Avg. tech. adoption in agriculture in 1970–2000	Avg. tech. adoption in communi-cations in 1970–2000	Avg. tech. adoption in transportation in 1970–2000	Avg. tech. adoption in industry in 1970–2000	Avg. of the sectoral tech. adoption indexes in 1970–2000	Absolute difference in log income, 2005	FST gen. distance relative to the USA, weighted
Panel B—pairwise correlations							
Avg. tech. adoption in communications in 1970–2000	0.5550* (5,886)	1 (7,381)					
Avg. tech. adoption in transportation in 1970–2000	0.5308* (5,356)	0.4331* (6,216)	1 (6,441)				
Avg. tech. adoption in industry in 1970–2000	0.5396* (5,460)	0.6335* (5,460)	0.5192* (5,050)	1 (5,565)			
Avg. of the sectoral tech. adoption indexes in 1970–2000	0.7615* (5,995)	0.7010* (7,260)	0.7591* (6,441)	0.7735* (5,565)	1 (7,503)		
Absolute difference in log income, 2005	0.4106* (6,105)	0.5619* (7,381)	0.4662* (6,441)	0.7210* (5,565)	0.6521* (7,503)	1 (10,440)	
FST gen. dist. relative to the USA, weighted	0.1301* (6,105)	0.1877* (7,381)	0.1248* (6,441)	0.2958* (5,565)	0.1975* (7,503)	0.3226* (10,440)	1 (10,585)
Simple FST genetic distance	−0.0562* (6,105)	0.0862* (7,381)	−0.0409* (6,441)	0.1407* (5,565)	0.0042 (7,503)	0.1950* (10,440)	0.5859* (10,585)

*Significant at the 5% level; # of obs. in parentheses.

Table 3.7 carries out the regression analysis for the contemporary period, controlling for geographic distance. Genetic distance relative to the frontier comes out positive in all cases, and significant at the 5% level or better for 3 of the 4 technological categories, as well as for the summary index of overall technology usage. The only category for which genetic distance is not significant is agricultural technologies. One possible interpretation is that agricultural technologies for the contemporary period under consideration have already widely diffused around the globe and are already intensively in use in much of the developing world, so that the effect of genetic distance as a barrier to their adoption can no longer be detected. We also carried out the same regression analysis as that in Table 3.8, but adding to the specification the measure of absolute genetic distance between pairs.[23] We found that relative genetic distance always trumped absolute distance, which sometimes carried a negative sign and was statistically insignificant in most cases. Thus, our test of the barriers story (Equation 3.30) also works when considering technology usage intensity rather than aggregate measures of development.

Turning to the historical evidence, Table 3.8 examines the determinants of technology usage differences along the extensive margin in the year 1500. As before, we use the English population as the frontier (as before, it matters little if we use the Italians instead—Italy was arguably the most technologically sophisticated country in the world in 1500). For 1500, we have 5 rather than 4 technological categories, plus the overall index of technological sophistication. We find that in all cases, genetic distance relative to the English is positive and statistically significant at the 10% level. In 5 of the 6 columns, it is significant at the 1% level (as before, the weakest results are for agricultural technologies). This is remarkable given the crudeness of the measure of technological use in 1500, based on counting whether or not each of 24 technologies, grouped in functional categories, were in use at all in a given country at the time. Moreover, as before, we also conducted horseraces between relative genetic distance and absolute genetic distance.[24] For five of the six indicators we again found that relative genetic distance trumps absolute genetic distance, with the latter entering with either the wrong sign, a very small magnitude, or low significance levels. The only exception, once again, was for agriculture.

Finally, we carried out the same analysis with the 33 disaggregated technologies chosen from the CHAT dataset. The results are presented in Table 3.9. For each technology, the table reports the coefficient on relative genetic distance to the USA (from a regression in which the standard set of geographic controls is included), the number of observations and countries, the standardized beta coefficient on genetic distance, and the R^2. The results vary across technologies of course, but interesting observations emerge: (1) In every single case the effect of genetic distance on differences in technology usage intensity is positive. (2) In 22 of the 33 cases, the coefficient on genetic distance is significant at the 10% level,

[23] Results are available upon request.

[24] Results are available upon request.

Table 3.7 Technological distance and genetic distance in the contemporary period (1970–2000) (dependent variables: measures of technological usage from Comin et al. as described in row 2)

	(1) Agricultural technology	(2) Communications technology	(3) Transportation technology	(4) Industrial technology	(5) Overall technology
FST gen. dist. relative to	0.402	0.500	0.608	1.149	0.745
the USA, weighted	(0.268)	(0.212)**	(0.185)***	(0.288)***	(0.216)***
Absolute difference in	0.687	0.274	0.306	0.329	0.361
latitudes	(0.121)***	(0.066)***	(0.057)***	(0.081)***	(0.082)***
Absolute difference in	0.405	0.089	0.305	0.174	0.243
longitudes	(0.129)***	(0.055)	(0.072)***	(0.069)**	(0.088)***
Geodesic distance	−0.050	−0.016	−0.036	−0.024	−0.032
	(0.014)***	(0.006)**	(0.008)***	(0.007)***	(0.010)***
=1 for contiguity	−0.050	−0.077	−0.053	−0.090	−0.071
	(0.014)***	(0.012)***	(0.013)***	(0.018)***	(0.012)***
=1 if either country is an	0.118	0.057	0.093	0.062	0.116
island	(0.077)	(0.027)**	(0.047)**	(0.023)***	(0.048)**
=1 if either country is	−0.007	0.018	−0.008	0.013	−0.016
landlocked	(0.028)	(0.017)	(0.011)	(0.023)	(0.014)
=1 if pair shares at least	0.036	−0.010	0.014	0.001	0.009
one sea or ocean	(0.027)	(0.015)	(0.015)	(0.020)	(0.019)
Constant	0.089	0.199	0.148	0.198	0.147
	(0.029)***	(0.018)***	(0.018)***	(0.023)***	(0.018)***
Observations	6,105	7,381	6,441	5,565	7,503
Standardized beta (%)	8.38	12.73	18.68	25.97	19.81
R-squared	0.25	0.10	0.14	0.16	0.17

Two-way clustered standard errors in parentheses.
* significant at 10%.
** Significant at 5%.
*** Significant at 1%.

Table 3.8 Technological distance and genetic distance in the year 1500 (dependent variables: measures of technological usage From Comin et al. as described in row 2)

	(1)	(2)	(3)	(4)	(5)	(6)
	Agricultural	**Military**	**Communications**	**Transportation**	**Industrial**	**Overall**
	technology	**technology**	**technology**	**technology**	**technology**	**technology**
Relative Fst genetic	0.551	1.752	1.279	1.926	1.673	1.524
distance to the	(0.281)*	(0.326)***	(0.288)***	(0.299)***	(0.271)***	(0.229)***
English, 1500 match						
Absolute difference	0.189	0.383	0.758	0.172	0.138	0.377
in latitudes	(0.096)**	(0.094)***	(0.092)***	(0.064)***	(0.061)**	(0.065)***
Absolute difference	−0.329	−0.018	−0.017	−0.039	0.061	−0.066
in longitudes	(0.082)***	(0.066)	(0.068)	(0.048)	(0.091)	(0.061)
Geodesic distance	0.049	0.009	0.009	0.014	0.048	0.025
	(0.010)***	(0.010)	(0.008)	(0.007)**	(0.010)***	(0.007)***
=1 for contiguity	0.037	−0.025	−0.042	−0.006	0.023	0.014
	(0.026)	(0.019)	(0.024)*	(0.021)	(0.025)	(0.014)
=1 if either country	−0.049	−0.087	−0.095	-0.073	−0.180	−0.092
is an island	(0.058)	(0.029)***	(0.053)*	(0.020)***	(0.031)***	(0.024)***
=1 if either country	0.017	−0.051	−0.020	−0.048	0.006	−0.022
is landlocked	(0.026)	(0.018)***	(0.016)	(0.011)***	(0.023)	(0.011)**
=1 if pair shares at	−0.006	−0.105	−0.018	−0.046	0.050	−0.019
least one sea or ocean	(0.020)	(0.034)***	(0.033)	(0.029)	(0.029)*	(0.025)
Constant	0.082	0.166	0.086	0.069	−0.126	0.016
	(0.034)**	(0.036)***	(0.026)***	(0.024)***	(0.039)***	(0.020)
Observations	5,253	5,886	5,886	5,253	5,253	5,886
Standardized beta (%)	10.41	29.26	19.95	41.81	25.27	31.63
R-squared	0.23	0.27	0.36	0.32	0.46	0.44

Two-way clustered standard errors in parentheses.
* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

Table 3.9 Bilateral Regressions of Technological Distance on Relative Genetic Distance for 33 Technologies (CHAT Dataset Averaged Over 1990–1999)

		Fst gen. dist. relative to the USA, weighted	Observations (countries)	Standardized beta (%)	R-squared
Agricultural technologies					
(1)	Harvest machines	2.044 (1.134)*	3,486 (84)	5.91	0.17
(2)	Tractors used in agriculture	19.615 (8.245)**	5,778 (108)	9.05	0.25
(3)	Metric tons of fertilizer consumed	73.393 (23.062)***	5,778 (108)	11.68	0.23
(4)	Area of irrigated crops	0.453 (0.276)*	5,565 (106)	7.21	0.03
(5)	Share of cropland area planted with modern varieties (% cropland)	0.182 (0.080)**	3,321 (82)	7.20	0.02
(6)	Metric tons of pesticides	0.738 (0.893)	4,465 (95)	2.62	0.12
Transportation technologies					
(7)	Civil aviation passenger km	0.484 (0.254)*	3,828 (88)	11.29	0.21
(8)	Lengths of rail line	0.397 (0.275)	4,656 (97)	5.26	0.28
(9)	Tons of freight carried on railways	2.330 (1.421)	4,005 (90)	10.63	0.16
(10)	Passenger cars in use	0.245 (0.082)***	5,886 (109)	15.88	0.26
(11)	Commercial vehicles in use	0.066 (0.025)***	5,050 (101)	23.50	0.29

(Continued)

Table 3.9 (*Continued*)

	Fst gen. dist. relative to the USA, weighted	Observations (countries)	Standardized beta (%)	R-squared
Medical technologies				
(12) Hospital beds	1.481 (4.319)	5,565 (106)	1.31	0.17
(13) DPT immunization before age 1	0.137 (0.156)	5,778 (108)	3.54	0.01
(14) Measles immunization before age 1	0.141 (0.162)	5,778 (108)	3.71	0.01
Communications technologies				
(15) Cable TV	74.485 (56.305)	4,753 (98)	4.23	0.16
(16) Cell phones	0.109 (0.044)**	5,778 (108)	8.21	0.12
(17) Personal computers	0.247 (0.099)**	4,950 (100)	12.53	0.21
(18) Access to the Internet	0.192 (0.072)***	5,778 (108)	14.25	0.28
(19) Items mailed/received	0.097 (0.074)	2,346 (69)	11.00	0.21
(20) Newspaper circulation	0.245 (0.101)**	5,886 (109)	10.43	0.25
(21) Radios	0.064 (0.139)	5,886 (109)	1.87	0.12
(22) Telegrams sent	0.312 (0.260)	2,211 (67)	5.74	0.07
(23) Mainline telephone lines	0.185 (0.067)***	5,886 (109)	11.54	0.28
(24) Television sets in use	0.492 (0.141)***	5,886 (109)	18.78	0.31

(*Continued*)

Table 3.9 (*Continued*)

	Fst gen. dist. relative to the USA, weighted	Observations (countries)	Standardized beta (%)	R-squared
Industrial technologies and other				
(25) Output of electricity, KwHr	34.477 (13.849)**	5,565 (106)	8.16	0.23
(26) Automatic looms	0.828 (0.304)***	3,570 (85)	11.19	0.06
(27) Total looms	1.200 (0.361)***	3,570 (85)	8.95	0.08
(28) Crude steel production in electric arc furnaces	0.091 (0.031)***	2,278 (68)	8.10	0.08
(29) Weight of artificial (cellulosic) fibers used in spindles	0.425 (0.354)	2,145 (66)	3.89	0.10
(30) Weight of synthetic (non cellulosic) fibers used in spindles	2.045 (0.819)**	2,145 (66)	9.89	0.20
(31) Weight of all types of fibers used in spindles	7.832 (2.759)***	2,850 (76)	12.10	0.07
(32) Visitor beds available in hotels and elsewhere	24.245 (7.518)***	5,565 (106)	9.31	0.10
(33) Visitor rooms available in hotels and elsewhere	13.518 (3.884)***	5,778 (108)	10.50	0.10

Two-way clustered standard errors in parentheses.

Unless specified in parentheses, the dependent variable is the absolute difference in per capita prevalence of the technology between country *i* and country *j*.

All regressions include controls for absolute difference in latitudes, absolute difference in longitudes, geodesic distance, dummy = 1 for contiguity, dummy = 1 if either country is an island, dummy = 1 if either country is landlocked, dummy = 1 if pair shares at least one sea or ocean.

* Significant at 10%.

** Significant at 5%.

*** Significant at 1%.

and in 19 cases at the 5% level. (3) The effect of genetic distance is particularly strong for disaggregated agricultural technologies and industrial technologies, and weakest for transportation and medical technologies. (4) The magnitude of the standardized effects, for those that are statistically significant, varies from 8% to 24%, a bit smaller but roughly in line with what we found using aggregate measured of productivity or the CEG dataset.[25]

A consideration of technologies at a more disaggregated data, rather than measures of overall productivity at the economy-wide level, provides additional evidence that human barriers matter. Not only is genetic distance relative to the frontier a strong predictor of technological usage differences in 1500 and in the contemporary period, we also find that it generally trumps absolute genetic distance. The fact that genetic distance accounts for differences in technological usage indicates that our previous aggregate results might in large part be accounted for by hindrances to the adoption of frontier technologies brought about by historical separation between populations.

3.5. ANCESTRY AND LONG-RUN DEVELOPMENT

In this section, we broaden the discussion of the role of ancestry as a determinant of the comparative wealth of nations, building on the discussion in Spolaore and Wacziarg (2013).[26] Our basic argument is that traits passed on across generations within societies play a fundamental role in accounting for the persistence of economic fortunes. However, the specific way in which these traits operate can take a variety of forms. In the model presented above, we argued that differences in vertically transmitted traits introduced barriers to the diffusion of innovations across nations. We found much evidence that this was the case for aggregate productivity and for specific innovations going back to the year 1500. However, we have not said much about what causes the onset of these innovations. Other authors have pointed to a role for traits to bear a direct effect on the onset of major productivity enhancing innovations, broadly construed. We have also not said much about the nature and specific method of transmission of the traits that are thought to matter for prosperity. These traits could be transmitted culturally, biologically, or through the interaction of culture and biology.

We proceed in several steps. We start by briefly describing the growing literature on long-run persistence in the wealth of nations. We argue that the intergenerational transmission of traits has a lot to do with explaining long-run persistence, because traits

[25] We also conducted horseraces between absolute and relative genetic distance for each of the 33 disaggregated technologies. Relative genetic distance remains positive and significant in 17 of the 22 cases where relative genetic distance is significant at the 10% level when entered on its own. In the vast majority of cases, absolute genetic distance enters insignificantly or with a negative sign.

[26] The discussion of the relation between cultural traits and economic outcomes is also drawn in part from Spolaore (2014).

are much more easily transmitted across generations than across societies. That is, ancestry matters to explain the wealth of nations. Next, we introduce a taxonomy to understand the manner in which ancestry matters. In particular, we introduce a distinction between barrier effects and direct effects of vertical traits. We also distinguish between the mode of transmission of the traits, either cultural, biological, or dual. Finally, we provide several examples from the recent literature illustrating the various ways in which ancestry can matter.

3.5.1 Persistence and Reversals: The Role of Ancestry

Discussions of the long-run roots of comparative development usually start with geographic factors. A large literature has documented strong correlations between economic development and geographic factors, for instance latitude, climate, and the disease environment.[27] The observation that geographic factors are correlated with development was at the root of Diamond's (1997) book on the long-run development advantage enjoyed by Eurasia—particularly Europe. On the surface, geography is a convenient explanation for persistence, because geography does not change very much, so that this immutable factor can be thought of as a prime reason for persistence in the wealth of nations. This view, however, is overly simplistic, for at least two reasons: First, the effect of geography on economic outcome can change depending on the technology of production. Geographic features useful to produce GDP in an agrarian economy may not be as helpful in an industrial society. Second, the manner in which geographic factors affect development today is open to a variety of interpretations. The factors could operate directly (for instance, a high disease burden can reduce productivity) or have an indirect effect through their historical legacy. While both channels could be operative, the literature has increasingly moved in the latter direction.

In fact, Diamond (1997) pointed out early that geographic factors such as the shape of continents and the availability of domesticable plants and animals probably did not have much to do with current development directly. It is because these factors gave people from Eurasia an early advantage in development, and because this advantage has persisted through the generations, that Europeans were able to conquer the New World (and many parts of the old one) and to remain at the top of the world distribution of income for a long time. This point became more widely recognized since a pathbreaking paper by Acemoglu et al. (2002) where these authors pointed out that the reversal of fortune experienced by former colonies between 1500 and today was inconsistent with a simple, direct effect of geography: for the geographic factors that made countries poor 500 years ago should be expected to make them poor today still. And yet fortunes were reversed

[27] See, for instance: on climate and temperature, Myrdal (1968); Kamarck (1976); Masters and McMillan (2001); Sachs (2001). On the disease environment: Bloom and Sachs (1998); Sachs et al. (2001); Sachs and Malaney (2002). On natural resources: Sachs and Warner (2001).

among a significant portion of the world's countries. This paper pointed to an indirect effect of geography, operating through institutions: where Europeans settled, they brought good institutions, and these are the fundamental proximate cause of development. Where Europeans chose to exploit and extract, the institutions they bequeathed had negative effects on development.

Yet that interpretation, too, became the subject of debates. Glaeser et al. (2004), for instance, state: "the Europeans who settled in the New World may have brought with them not so much their institutions, but themselves, that is, their human capital. This theoretical ambiguity is consistent with the empirical evidence." We would go even further: Europeans who settled in the New World brought with them the whole panoply of vertically transmitted traits—institutions, human capital, norms, values, preferences. This vector of vertical traits was by definition easier to transmit to the descendents of Europeans than it was to convey to colonized populations. This interpretation suggests an important role for ancestry, rather than only institutions, as an explanation for the reversal of fortunes. Locations that were colonized by Europeans and were previously characterized by low population density and the prevalence of non-agrarian modes of subsistence became rich. Locations that were inhospitable to Europeans remained poor, and Europeans remained at the top of the world distribution of aggregate productivity throughout.[28] That the wealth of a nation seems so strongly affected by the wealth of the ancestors of those living in that nation suggests a central role for vertically transmitted traits as an explanation for both long-run persistence and the current distribution of income.

This interpretation led various authors to focus explicitly on persistence and ancestry. First came our own work on genetic distance as a barrier to development, already discussed in the previous sections (Spolaore and Wacziarg, 2009). Next came important papers by Putterman and Weil (2010) and Comin and Hobijn (2010). These papers also explore the deep historical roots of current development.

Putterman and Weil (2010) look at two important determinants of the current wealth of nations: experience with agriculture, measured by the time elapsed since the adoption of sedentary agriculture as a primary means of food production; and experience with a centralized state, measured by the number of years a country has experienced centralized governance, discounting years that occurred farther in the past. Both variables are predictors of today's per capita income, but they enter even more strongly when they are adjusted

[28] We greatly expand on this point in Spolaore and Wacziarg (2013). In that paper, we revisit the Acemoglu et al. (2002) evidence on the reversal of fortune. By examining the correlation between population density in 1500 and per capita income today, we confirm their findings for former colonies. Yet we also show that (1) any evidence of a reversal of fortune disappears when European countries are included in the sample; (2) there is evidence of persistence among countries that were not former European colonies; (3) persistence is even stronger when looking at countries that are populated mostly by their indigenous populations. These facts are suggestive of a strong role for ancestry as an explanation for persistence.

for ancestry. To adjust variables for ancestry, Putterman and Weil construct a migration matrix. In this matrix, a row pertains to a country, and columns contain the fraction of that country's population whose ancestors in 1500 lived in each of the world's countries. For the Old World, entries are mostly diagonal: that is, the ancestors of the French mostly lived in France in 1500. For the New World, however, the ancestors of current populations are often in significant numbers from other continents altogether—primarily European countries for European colonizers, and Sub-Saharan African countries for the descendants of former slaves. By premultiplying a variable by the migration matrix, one obtains this variable's ancestry-adjusted counterpart. For instance, for Australia, the history of the location is the history of the Aborigenes, while the history of the current population is mostly the history of the English. Putterman and Weil's major contribution is to show that ancestry-adjusted years of agriculture and ancestry-adjusted state centralization are much stronger predictors of current income than their non-ancestry adjusted counterparts. This suggests an important role, again, for traits that are passed on intergenerationally within populations.

Comin et al. (2010) take a different approach, but reach a similar conclusion: they show that the degree of technological sophistication of countries is highly autocorrelated even at very long horizons: they detect correlations between current technological usage levels (measured along the intensive margin in the current period) and technological usage as far back as the year 1000 BC (measured along the extensive margin for a set of 12 ancient technologies). Current per capita income is also correlated strongly with past technological sophistication in the years 1000 BC, 1 AD, and 1500 AD. In this case, a history of technological advancement predicts current income and technological advancement, an indication of persistence. The crucial point, however, is again that when the historical (lagged) variables are entered in their ancestry-adjusted forms, they are much stronger predictors of current outcomes than variables that capture the history of a location. In this context also, there appears to be a strong role for ancestry and intergenerational transmission as explanations for the persistence in technology and income levels.

Why does ancestry matter? In what follows, we present a taxonomy of the possible effects of vertically transmitted traits on growth and development. This taxonomy is summarized in the following matrix:

Mode of operation ⟶ Mode of transmission ↓	Direct effect	Barrier effect
Biological Transmission (genetic and/or epigenetic)	Quadrant I	Quadrant IV
Cultural Transmission (behavioral and/or symbolic)	Quadrant II	Quadrant V
Dual Transmission (biological-cultural interaction)	Quadrant III	Quadrant VI

3.5.2 Modes of Transmission

The inheritance of traits from one generation to the next in humans takes place through several modes of transmission and along multiple dimensions. Recent scientific advances stress the complexity of different inheritance mechanisms (for example, see Jablonka and Lamb, 2005) which interact with each other as well as with environmental and societal factors. For simplicity, in our taxonomy we focus on three broad categories: biological transmission, cultural transmission, and the interaction of biological and cultural transmission (dual transmission).

Biological transmission includes genetic transmission. Individuals inherit nuclear DNA from their parents. Humans also inherit mitochondrial DNA (mtDNA) only from their mothers mitochondrial DNA codes for the genes of the cell structures which convert food into useable energy, while nuclear DNA codes for the rest of the human genome. The measures of genetic distance used previously in this chapter are based on differences in the distribution of nuclear DNA across populations—that is, on differences in DNA inherited from both parents. As already mentioned, genetic distance is based on neutral genes, which change randomly and are not affected by natural selection. Other parts of the DNA code for genes that are affected by natural selection, such as those affecting eye color or skin color. All these traits are transmitted biologically.

However, genetic transmission is not the only form of biological transmission. In recent years, biologists have also given much attention to epigenetic inheritance systems. Epigenetics refers to the mechanisms through which cells with the same genetic information (i.e. DNA) acquire different phenotypes (i.e. observable characteristics) and transmit them to their daughter cells. Examples of epigenetic markers are methylation patterns: DNA methylation is a biochemical process that stably alters the expression of genes in cells by adding a methyl group to a DNA nucleotide. There is currently a debate in the scientific literature about the extent to which epigenetic changes can be inherited from one generation to the next—for instance, see Chandler and Alleman (2008) and Morgan and Whitelaw (2008). An example of possible intergenerational epigenetic inheritance, mentioned by Morgan and Whitelaw (2008), is the Dutch Famine Birth Cohort Study by Lumey (1992), reporting that children born during famine in World War II were smaller than average and that the effects could last two generations (but see also Stein and Lumey, 2002). In principle, epigenetic mechanisms could explain rapid biological changes in populations that could not be due to genetic selection. Epigenetic mechanisms have recently been emphasized by microeconomists working on human capital formation, such as Cunha and Heckman (2007, p. 32), who wrote: "the nature versus nurture distinction is obsolete. The modern literature on epigenetic expression teaches us that the sharp distinction between acquired skills and ability featured in the early human capital literature is not tenable."

Of course, biological inheritance is not the only mode of intergenerational transmission of traits across human beings. Many traits are transmitted culturally from one generation to the next. An important example is the specific language that each child acquires through learning and imitation, usually (but not necessarily) from parents or other close relatives. Other cultural traits include values, habits, and norms. In general, culture is a broad concept, which encompasses a vast range of traits that are not transmitted biologically across generations. The Webster's Encyclopedic Unabridged Dictionary defines culture as including "the behaviors and beliefs characteristic of a particular social, ethnic or age group" and "the total ways of living built up by a group of human beings and transmitted from one generation to the other." Richerson and Boyd (2005, p. 5), two leading scholars in the field of cultural evolution, define culture as "information capable of affecting individuals' behavior that they acquire from other members of their species through teaching, imitation, and other forms of social transmission."

Following Jablonka and Lamb (2005), we can distinguish between two forms of cultural transmission, both involving social learning: behavioral transmission and symbolic transmission. Behavioral transmission takes place when individuals learn from each other by direct observation and imitation. Symbolic transmission instead is about learning by means of systems of symbols—for example, by reading books. Most scholars of human evolution believe that the bulk of observed human variation in intergenerationally transmitted traits is mainly due to cultural transmission rather than to biological transmission. For instance, prominent anthropologists Henrich and McElreath (2003, p. 123) write: "While a variety of local genetic adaptations exist within our species, it seems certain that the same basic genetic endowment produces arctic foraging, tropical horticulture, and desert pastoralism [...]. The behavioral adaptations that explain the immense success of our species are cultural in the sense that they are transmitted among individuals by social learning and have accumulated over generations. Understanding how and when such culturally evolved adaptations arise requires understanding of both the evolution of the psychological mechanisms that underlie human social learning and the evolutionary (population) dynamics of cultural systems."

In sum, our classification of modes of intergenerational transmission includes two broad categories: biological transmission (both genetic and epigenetic), and cultural transmission (behavioral and symbolic). However, these two forms of transmission should not be viewed as completely distinct and independent. On the contrary, a growing line of research stresses that human evolution often proceeds from the interaction between biological and cultural inheritance systems, where each system is influenced by the other system. According to Richerson and Boyd (2005, p. 194), genes and culture can be seen as "obligate mutualists, like two species that synergistically combine their specialized capacities to do things that neither can do alone. [...] Genes, by themselves can't readily adapt to rapidly changing environments. Cultural variants, by themselves, can't do anything without brains and bodies. Genes and culture are tightly coupled but subject to evolutionary forces that tug behavior in different directions." This approach to evolution is known as dual inheritance theory or gene-culture coevolution (Cavalli-Sforza and Feldman, 1981;

Cavalli-Sforza et al. 1994; Boyd and Richerson, 1985; Richerson and Boyd, 2005). In such a framework, observable human outcomes can be viewed as stemming from the interplay of genetically and culturally transmitted traits. A well-known example of gene-culture coevolution is the spread of the gene controlling lactose absorption in adults in response to cultural innovations, such as domestication and dairying (Simoons, 1969, 1970; Richerson and Boyd, 2005; Chapter 6). The ability to digest milk as an adult (i.e. to be "lactase persistent") is given by a gene that is unequally distributed among different populations: it is prevalent among populations of European descent, but very rare among East Asians and completely absent among Native Americans. It is well understood that such a gene did spread rapidly after the introduction of domestication among populations that kept milk-producing animals, such as cows or goats, reinforcing the advantages from those practices from an evolutionary perspective. In general, dual inheritance—the third "mode of transmission" in our taxonomy—captures such a complex interaction between genetic and cultural factors.

3.5.3 Modes of Operation

Traits can be transmitted from one generation to the next biologically, culturally, or through the interaction of genes and culture (dual transmission). But how do such traits affect economic outcomes? Our taxonomy distinguishes between direct effects and barrier effects.

Direct Effects. Most of the economic literature has focused on direct effects of vertically transmitted traits on income and productivity. Such effects occur when individuals inherit traits that directly impact economic performance, either positively or negatively. For example, most contributions on the relation between cultural values and economic development stress inherited norms and beliefs that directly lead to positive or negative economic outcomes. Weber (2005), the great German sociologist and political economist, in his classic book *The Protestant Ethic and the Spirit of Capitalism*, provided a systematic and influential study emphasizing the direct positive effects of specific culturally transmitted traits on economic performance. Weber was in part reacting to the Marxist view, which considered cultural beliefs and values, such as religion, as the by-product of underlying economic factors. Instead, Max Weber argued for direct causal effects of culturally transmitted traits on economic outcomes. Specifically, he proposed that the emergence of a new Protestant ethic, which linked "good works" to predestination and salvation, had a direct effect on the rising of the "spirit of capitalism", a new attitude toward the pursuit of economic prosperity. Among Weber's more recent followers is, for example, the economic historian Landes (1998, 2000), who titled one of his contributions *Culture Makes Almost All the Difference*, and opened it with the line "Max Weber was right." Landes' emphasis was also on the direct economic effects of culture, defined as "the inner values and attitudes that guide a population." According to Landes (p. 12): "This is not to say that Weber's 'ideal type' of capitalist could be found only among Calvinists [...]. People of all faiths and no faith can grow up to be rational, diligent, orderly, productive, clean, and humourless. [...] Weber's argument, as I see it, is that in 16th–18th-century northern Europe, religion encouraged the appearance in numbers of a personality type that had been exceptional and adventitious before and that this type created a new economy (a new mode of production) that we know as (industrial) capitalism."

An extensive empirical literature has attempted to directly test Weber's hypotheses, often concluding with a negative assessment of direct effects of Protestant values on economic outcomes. Recent contributors to this literature were Becker and Ludger (2009), who used county-level data from 19th century Prussia, and attempted to estimate the causal effect of Protestantism on economic performance by exploiting the fact that the Lutheran Reform expanded concentrically from Wittenberg, Martin Luther's city. They concluded that Protestantism fostered economic development, but that the main channel was not the spread of a new work ethic associated with religious values, but the expansion of literacy as a consequence of education in reading the Bible.

The direct effects of religious beliefs on economic outcomes were investigated empirically by Barro and McCleary (2003). Barro and McCleary used instrumental variables, such as the existence of a state religion and of a regulated market structure, to identify the direct effect of religion on growth. They concluded that economic growth is positively associated with the extent of religious beliefs, such as those in hell and heaven, but negatively associated to church attendance. They interpreted their results as consistent with a direct effect of religion—a culturally transmitted set of beliefs—on individual characteristics that foster economic performance. Guiso et al. (2003) also studied the effects of religious beliefs on economic attitudes and outcomes, such as cooperation, legal rules, thriftiness, the market economy, and female labor participation. They found that religious beliefs tend to be associated with attitudes conducive to higher income per capita and higher economic growth, and that the effects differ across religious denominations.

While scholars such as Weber have stressed the positive direct effects of cultural traits, such as the Protestant ethic, other scholars have argued that specific culturally transmitted traits and values can be responsible for economic backwardness and underdevelopment. An influential and widely debated example of this view was provided by the political scientist Banfield (1958) in his classic book *The Moral Basis of a Backward Society*, written in collaboration with his wife Laura Fasano, and based on their visit to the southern Italian town of Chiaromonte (called "Montegrano" in the book). Banfield argued that the economic backwardness of that society could be partly explained by the direct effects of inherited values summarized by the term "amoral familism", and consisting in a lack of mutual trust and cooperation, and a disregard for the interests of fellow citizens who were not part of one's immediate family. A theory of intergenerational transmission directly inspired by Banfield's analysis has been provided recently by Tabellini (2008), who also built "on analytical work" on Bisin and Verdier's (2000, 2001) seminal work on the economics of cultural transmission. In Tabellini's model, parents choose which values to transmit to their children, depending on the patterns of external enforcement and expected future transactions. In particular, Tabellini shows that path dependence is possible: adverse initial conditions can lead to a unique equilibrium where legal enforcement is weak and inherited cultural values discourage cooperation.

A recent example of an empirical study of the direct effects of inherited traits on economic growth is Algan and Cahuc (2010). Algan and Cahuc document how the level of inherited trust of descendants of immigrants in the United States is significantly influenced by the country of origin and the timing of arrival of their ancestors. They then use the inherited trust of descendants of immigrants in the US as a time-varying measure of inherited trust in their country of origin, in order to identify the impact of inherited trust on growth, controlling for country fixed effects. Algan and Cahuc find that changes in inherited trust during the 20th century have a large impact on economic development in a panel of 24 countries.

The above-mentioned contributions are examples of a much larger literature on the direct effects of cultural traits on economic outcomes. There is also a smaller but important literature that has extended the analysis to traits that are transmitted biologically, or stem from the interaction of genes and culture (dual inheritance). An example is the contribution by Galor and Moav (2002), who modeled an intergenerationally transmitted trait affecting humans' fertility strategies. They posited that some individuals inherited traits that induced them to follow a quantity-biased strategy, consisting in the generation of a higher number of children, while other individuals followed a quality-biased strategy, consisting in the investment of more resources in a smaller number of offspring. Galor and Moav argued that the evolutionary dynamics of these traits had direct implications for the onset of the Industrial Revolution and the following demographic transition. In the pre-industrial world, caught in a Malthusian trap, selective pressures favored parental investment, which led to higher productivity. In their model, the spread of this inherited predilection for a smaller number of children led endogenously to the transition out of the Malthusian regime. Galor and Moav in their contribution stressed biological transmission. However, their analysis can also be interpreted as a model of cultural transmission of traits influencing fertility strategies, or as the outcome of the interaction of biological and cultural traits.

A more recent contribution that stresses the direct effects of different distributions of inter-generationally transmitted traits on economic development is Ashraf and Galor (2013a). In that study, Ashraf and Galor focus on genetic diversity. While genetic distance refers to genetic differences between populations, genetic diversity is about heterogeneity within populations. In their study, Ashraf and Galor (2013a) document a non-monotonic relationship between genetic diversity and development, and argue that such relation is causal, stemming from a trade-off between the beneficial and the detrimental effects of diversity of traits on productivity. Again, while the focus of Ashraf and

Galor's empirical analysis is on genetic variables, the modes of transmission from intergenerational traits to economic outcomes can operate both through biological and cultural channels, and their interactions. A further discussion of the relation between genetic diversity and ethnic and cultural fragmentation is provided by Ashraf and Galor (2013b).

The interaction of culture and genes is explicitly at the center of the economic analysis of the effects of lactase persistence provided by Cook (2012). Cook argues that country-level variation in the frequency of lactase persistence is positively and significantly related to economic development in pre-modern times—which he measures by using population density in 1500 CE, as we did earlier in this chapter. Specifically, he finds that an increase in one standard deviation in the frequency of lactase persistent individuals (roughly 24% points) is associated with a 40% increase in pre-modern population density. Cook uses instrumental variables (solar radiation) to assess causality, and interprets his results as reflecting the direct effects of inherited cultural and biological traits associated with the introduction of dairying.

Barrier effects. As we already mentioned, most of the contributions on the relation between ancestry and economic performance, including the examples mentioned above, tend to focus on the direct effects of intergenerationally transmitted traits on economic outcomes. However, as we emphasized in the theoretical and empirical analysis presented in the first sections of this chapter, differences in inherited traits can also affect comparative development by acting as barriers to the diffusion of goods, services, ideas, and innovations. A focus on barriers can explain why differences in inherited traits may matter, even though many new ideas and innovations are learned "horizontally," from individuals and populations that are not directly related, rather than "vertically," from one's close relatives and ancestors. The fact is, that when barrier effects do exist, vertically transmitted traits also affect horizontal learning and diffusion. People are more likely to learn new ideas and adopt new technologies from other people who, while not directly related to them, share more recent common ancestors and, consequently, also share, on average, a larger set of inherited traits and characteristics.

The literature on the barrier effects of vertically transmitted traits is not as large as the one on direct effects. In addition to our own contributions, already discussed, a recent example is Guiso et al. (2009), who studied the barrier effects of cultural traits by using data on bilateral trust between European countries. They found that bilateral trust is affected by cultural aspects of the match between trusting country and trusted country, such as their history of conflicts and their religious, genetic, and somatic similarities. Lower bilateral trust then acts as a cultural barrier: it is associated with less bilateral trade, less portfolio investment, and less direct investment between the two countries, even after controlling for other characteristics of the two countries. These findings suggest that culturally transmitted traits can have a significant barrier effect on economic interactions between different societies.

Another study that documents the effects of cultural barriers on trade is provided by Felbermayr and Toubal (2010). Felbermayr and Toubal measure cultural proximity or distance between countries using bilateral score data from the Eurovision Song Contest, a popular European television show. For instance, viewers in Cyprus award Greek singers more points on average than the Greeks receive from viewers in other countries, and vice versa. In contrast, Cypriot and Turkish viewers give each other below-average scores. Felbermayr and Toubal exploit the variation of these scores within-pair and across time to estimate the effects of cultural proximity on bilateral trade, finding significant effects.

An open question concerns the relationship between direct and barrier effects. Of course, in principle, both modes of operation can be at work simultaneously, and some specific traits can play a role along both channels. For example, populations that inherit values and beliefs that make them more open to risk and innovation could benefit directly from such traits, but may also face lower barriers to interactions with other groups. In general, the study of barrier effects stemming from historical and cultural divergence is a promising area of research, still in its infancy, both from a theoretical and empirical perspective. The taxonomy and discussion presented in this chapter are only a first step toward a more complete understanding of this important topic.

3.6. CONCLUSION

In this chapter we provided a theoretical framework and empirical evidence to shed light on a fundamental question: What barriers prevent the diffusion of the most productive technologies from the technological frontier to less developed economies?

In the first part of this chapter, we presented a simple analytical framework to illustrate two basic ideas. The first idea was that genetic distance between populations, which measures their degree of genealogical relatedness, can be interpreted as a summary metric for average differences in traits that are transmitted with variation from one generation to the next. We modeled the transmission of such "vertical" traits—that is, the transmission of characteristics which are passed on vertically across generations within a population over the very long run—and derived the relation between divergence in vertical traits and genetic distance. The second idea was that differences in vertically transmitted traits act as obstacles to horizontal learning and imitation across different populations. We argued that populations that share a more recent common history and are therefore closer in terms of vertical traits tend to face lower costs and barriers to adopting each other's technological innovations.

In the second part of this chapter we brought these ideas to the data. We introduced measures of genetic distance between populations, and used them to test our barrier model of diffusion. We found that, as the model predicts, genetic distance measured relative to the world's technological frontier trumps absolute genetic distance as an explanation for bilateral income differences and for the different usage of specific technological innovations. This was the case both historically, when we measured technological usage on the extensive margin, and for more recent technological developments, when we measured technological usage along the intensive margin. We also documented that, as implied by our model, the effect of genetic distance was more pronounced after a major innovation, such as the onset of the Industrial Revolution, and declined as more populations adopted the frontier's innovation. Overall, we found considerable evidence that barriers introduced by historical separation between populations have played a key role in the diffusion of technological innovations and economic growth.

In the third and final part of this chapter, we discussed our hypotheses and results within the broader context of the growing literature on the deep historical roots of economic development. To organize our discussion we presented a taxonomy based on Spolaore and Wacziarg (2013). The taxonomy provided a conceptual basis for discussing how intergenerationally transmitted traits could conceivably affect economic outcomes. Our taxonomy distinguished possible economic effects of vertical traits along two dimensions. The first dimension referred to the mode of transmission of vertical traits, which could be biological (genetic or epigenetic), cultural (behavioral or symbolic), or resulting from the interaction of genes and culture (dual inheritance). The second dimension defined the mode of operation of these traits, depending on whether they have direct effects on economic outcomes, or operate as barriers to economic interactions between populations. We briefly reviewed examples of economic contributions that focused on different effects—direct effects or barrier effects—of traits transmitted biologically, culturally, or through dual transmission. We argued that most of the literature so far has mainly focused on direct effects, while much less attention has been given to the study of barriers to development stemming from long-term cultural and historical divergence.

The topic of human barriers introduced by historical divergence and their effects on social, political, and economic outcomes is an exciting emerging field of study. Our own work continues to explore the effects of variation in human relatedness on a variety of political economy outcomes. For instance, Spolaore and Wacziarg (2012b) examine the effects of genealogical relatedness on the propensity for interstate militarized conflict, finding that a smaller genetic distance is associated with a significantly higher probability of a bilateral conflict between two countries. This effect, again, is interpreted as evidence of a barrier between societies characterized by distinct norms, values, preferences, and cultures. This time, however, the barrier impedes a costly rather than a beneficial interaction. In ongoing work, we explore the effects of relatedness on trade and financial flows across countries. Finally, we have recently begun an effort to better characterize what genetic relatedness captures, by investigating the relationship between various measures of cultural differences and genetic distance—the goal being to more clearly identify the source of the barriers introduced by a lack of genealogical relatedness. For instance, the barriers could take the form of a lack of trust, differences in preferences or norms, or transactions costs linked to an inability to communicate and coordinate. This chapter provides only an introduction and first step toward a more comprehensive and systematic analysis of such important, unexplored, and promising topics.

APPENDIX 1. TECHNOLOGIES USED IN THE VARIOUS DATASETS

A. 24 Technologies in the CEG 1500 AD Dataset.

1. *Military*: Standing army, cavalry, firearms, muskets, field artillery, warfare capable ships, heavy naval guns, ships (+180 guns).
2. *Agriculture*: Hunting and gathering; pastoralism; hand cultivation; plow cultivation.
3. *Transportation*: Ships capable of crossing the Atlantic Ocean, ships capable of crossing the Pacific Ocean, ships capable of reaching the Indian Ocean, wheel, magnetic compass, horse powered vehicles.
4. *Communications*: Movable block printing; woodblock or block printing; books, paper.
5. *Industry*: Steel, iron.

B. 9 Technologies in the CEG 2000 AD Dataset.

Electricity (in 1990), Internet (in 1996), PCs (in 2002), cell phones (in 2002), telephones (in 1970), cargo and passenger aviation (in 1990), trucks (in 1990), cars (in 1990), tractors (in 1970).

C. 33 Technologies in the CHAT Dataset for 1990–1999.

1. *Agriculture*: Harvest machines, tractors used in agriculture, metric tons of fertilizer consumed, area of irrigated crops, share of cropland area planted with modern varieties (% cropland), metric tons of pesticides.
2. *Transportation*: Civil aviation passenger km, lengths of rail line, tons of freight carried on railways, passenger cars in use and commercial vehicles in use.
3. *Medical*: Hospital beds, DPT immunization before age 1, measles immunization before age 1.
4. *Communications*: Cable TV, cell phones, personal computers, access to the Internet, items mailed/received, newspaper circulation, radios, telegrams sent, mainline telephone lines, television sets in use.
5. *Industry and other*: Output of electricity, Kw Hr, automatic looms, total looms, crude steel production in electric arc furnaces, weight of artificial (cellulosic) fibers used in spindles, weight of synthetic (non cellulosic) fibers used in spindles, weight of all types of fibers used in spindles, visitor beds available in hotels and elsewhere, visitor rooms available in hotels and elsewhere.

REFERENCES

Acemoglu, Daron, Johnson, Simon, Robinson, James A., 2002. Reversal of fortune: geography and institutions in the making of the modern world income distribution. Quarterly Journal of Economics 117 (4), 1231–1294.

Alesina, Alberto, Devleeschauwer, Arnaud, Easterly, William, Kurlat, Sergio, Wacziarg, Romain, 2003. Fractionalization. Journal of Economic Growth 8, 55–194.

Algan, Yann, Cahuc, Pierre, 2010. Inherited trust and growth. American Economic Review 100 (5), 2060–2092.

Ashraf, Quamrul, Galor, Oded, 2011. Dynamics and stagnation in the Malthusian Epoch. American Economic Review 101 (5), 2003–2041.

Ashraf, Quamrul, Galor, Oded, (2013a). The "Out-of-Africa" hypothesis, human genetic diversity, and comparative economic development. American Economic Review 103(1), 1–46.

Ashraf, Quamrul, Galor, Oded, (2013b). Genetic diversity and the origins of cultural fragmentation, American Economic Review 103(3) 528–533.

Banfield, Edward C., 1958. The Moral Basis of a Backward Society. The Free Press, New York, NY.

Barro, Robert J., McCleary, Rachel, 2003. Religion and economic growth. American Sociological Review 68 (5), 760–781.

Barro, Robert J., Sala-i-Martin, Xavier, 1997. Technological diffusion, convergence and growth. Journal of Economic Growth 2 (1), 1–26.

Barro, Robert J., Sala-i-Martin, Xavier, 2003. Economic Growth, second ed. MIT Press, Cambridge, MA.

Becker, Sascha O., Woessmann, Ludger, 2009. Was Weber wrong? A human capital theory of protestant economic history. Quarterly Journal of Economics 124 (2), 531–596.

Bisin, Alberto, Verdier, Thierry, 2000. Beyond the melting pot: cultural transmission, marriage, and the evolution of ethnic and religious traits. Quarterly Journal of Economics 115, 955–988.

Bisin, Alberto, Verdier, Thierry, 2001. The economics of cultural transmission and the dynamics of preferences. Journal of Economic Theory 97, 298–319.

Bloom, David E., Sachs, Jeffrey D., 1998. Geography, demography, and economic growth in Africa. Brookings Papers on Economic Activity 2, 207–273.

Bossert, Walter, D'Ambrosio, Conchita, La Ferrara, Eliana, 2011. A generalized index of fractionalization. Economica 78 (312), 723–750.

Boyd, Robert, Richerson, Peter J., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago.

Cameron, A. Colin, Gelbach, Jonah B., Miller, Douglas L., 2006. Robust Inference with Multi-Way Clustering. NBER Technical Working Paper #T0327.

Caselli, Francesco, 2005. Accounting for cross-country income differences. In: Aghion, Philippe, Durlauf, Steven N. (Eds.), Handbook of Economic Growth, vol. 1A. North-Holland, New York, pp. 679–741.

Cavalli-Sforza, Luigi Luca, Feldman, Marcus W., 1981. Cultural Transmission and Evolution: a Quantitative Approach. Princeton University Press, Princeton.

Cavalli-Sforza, Luigi L., Menozzi, Paolo, Piazza, Alberto, 1994. The History and Geography of Human Genes. Princeton University Press, Princeton.

Chanda, Areendam, Justin Cook, C., Putterman, Louis, 2013. Persistence of Fortune: Accounting for Population Movements, There was No Post-Columbian Reversal, Working Paper. Brown University, February.

Chandler, V., Alleman, M., 2008. Paramutation: epigenetic instructions passed across generations. Genetics 178 (4), 1839–1844.

Comin, Diego, Hobijn, Bart, 2009. The CHAT Dataset. Harvard Business School Working Paper # 10-035.

Comin, Diego, Hobijn, Bart, 2010. An exploration of technology diffusion. American Economic Review 100 (5), 2031–2059 (December).

Comin, Diego, Hobijn, Bart, Rovito, Emilie, 2008. World technology usage lags. Journal of Economic Growth 13 (4).

Comin, Diego, Easterly, William, Gong, Erick, 2010. Was the wealth of nations determined in 1000 B.C.? American Economic Journal: Macroeconomics 2 (3), 65–97.

Cook, C. Justin, 2012. The Role of Lactase Persistence in Precolonial Development. Working Paper, Yale University, August.

Cunha, Flavio, Heckman, James, 2007. The technology of skill formation. American Economic Review 97 (2), 31–47.

Desmet, Klaus, Le Breton, Michel, Ortuño-Ortín, Ignacio, Weber, Shlomo, 2011. The stability and breakup of nations: a quantitative analysis. Journal of Economic Growth 16, 183–213.

Diamond, Jared, 1997. Guns, Germs and Steel: The Fate of Human Societies. Norton & Co., New York.

Fagerberg, Jan, 2004. Innovation: a guide to the literature. In: Fagerberg J., Mowery, D.C., Nelson, R.R. (Eds.), Oxford Handbook of Innovation, Oxford University Press, Oxford (Chapter 1).

Felbermayr, Gabriel J., Toubal, Farid, 2010. Cultural proximity and trade. European Economic Review 54 (2), 279–293.

Galor, Oded, Moav, Omer, 2002. Natural selection and the origin of economic growth. Quarterly Journal of Economics 117 (4), 1133–1191.

Glaeser, Edward L., La Porta, Rafael, Lopez-de-Silanes, Florencio, Shleifer, Andrei, 2004. Do institutions cause growth? Journal of Economic Growth 9 (3), 271–303.

Greenberg, Joseph E., 1956. The measurement of linguistic diversity. Language 32 (1), 109–115.

Guiso, Luigi, Sapienza, Paola, Zingales, Luigi, 2003. People's opium? Religion and economic attitudes. Journal of Monetary Economics 50 (1), 225–282.

Guiso, Luigi, Sapienza, Paola, Zingales, Luigi, 2009. Cultural biases in economic exchange. Quarterly Journal of Economics 124 (3), 1095–1131.

Hall, Robert E., Jones, Charles I., 1999. Why do some countries produce so much more output per worker than others? Quarterly Journal of Economics 114 (11), 83–116.

Henrich, Joseph, McElreath, Richard, 2003. The Evolution of Cultural Evolution. Evolutionary Anthropology 12, 123–135

Hsieh, Chang-Tai, Klenow, Peter J., 2010. Development accounting. American Economic Journal: Macroeconomics 2 (1), 207–223.

Jablonka, Eva, Lamb, Marion J., 2005. Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life. MIT Press, Cambridge, MA.

Kamarck, Andrew M., 1976. The Tropics and Economic Development. Johns Hopkins University Press, Baltimore and London.

Kimura, Motoo, 1968. Evolutionary rate at the molecular level. Nature 217, 624–626.

Landes, David, 1998. The Wealth and Poverty of Nations. Norton, New York.

Landes, David, 2000. Culture makes almost all the difference. In: Harrison, Lawrence E., Huntington, Samuel P. (Eds.), Culture Matters: How Values Shape Human Progress. Basic Books, New York, NY, USA, pp. 2–13.

Lumey, Lambert H., 1992. Decreased birthweights in infants after maternal in utero exposure to the Dutch famine of 1944–1945. Paediatric and Perinatal Epidemiology 6, 240–253.

Maddison, Angus, 2003. The World Economy: Historical Statistics. OECD Development Center, Paris, France.

Masters, William A., McMillan, Margaret S., 2001. Climate and scale in economic growth. Journal of Economic Growth 6(3), 167–186.

McEvedy, Colin, Jones, Richard. 1978. Atlas of World Population History. Penguin Books, Middlesex.

Morgan, Daniel K., Whitelaw, Emma, 2008. The case for transgenerational epigenetic inheritance in humans. Mammalian Genome 19, 394–397.

Olsson, Ola, Hibbs Jr., Douglas A., (2005). Biogeography and long-run economic development, European Economic Review 49(4), 909-938.

Parente, Stephen L., Prescott, Edward C., 1994. Barriers to technology adoption and development. Journal of Political Economy 102 (2), 298–321.

Parente, Stephen L., Prescott, Edward C., 2002. Barriers to Riches. MIT Press, Cambridge.

Putterman, Louis, Weil, David N., 2010. Post-1500 population flows and the long-run determinants of economic growth and inequality. Quarterly Journal of Economics 125 (4), 1627–1682.

Richerson, Peter J., Boyd, Robert, 2005. Not by Genes Alone: How Culture Transformed Human Evolution. University of Chicago Press, Chicago.

Rogers, Everett M., 1962. The diffusion of innovations, first ed. Free Press, New York (fifth edition: 2003).

Simoons, Frederick J., 1969. Primary adult lactose intolerance and the milking habit: a problem in biological and cultural interrelations: I. Review of the medical research. The American Journal of Digestive Diseases 14, 819–836.

Simoons, Frederick J., 1970. Primary adult lactose intolerance and the milking habit: a problem in biological and cultural interrelations: II. A culture historical hypothesis. The American Journal of Digestive Diseases 15, 695–710.

Spolaore, Enrico, 2014. Introduction. In: Spolaore, Enrico (Ed.), Culture and Economic Growth, International Library of Critical Writings in Economics Series. Edward Elgar, Cheltenham.

Spolaore, Enrico, Wacziarg, Romain, 2009. The diffusion of development. Quarterly Journal of Economics 124 (2), 469–529.

Spolaore, Enrico, Wacziarg, Romain, 2012a. Long-term barriers to the international diffusion of innovations. In: Frankel, Jeffrey, Pissarides, Christopher (Eds.), NBER International Seminar on Macroeconomics 2011. University of Chicago Press, Chicago, pp. 11–46 (Chapter 1).

Spolaore, Enrico, Wacziarg, Romain, 2012b. War and Relatedness, Working Paper. Tufts University and UCLA, June 2012.

Spolaore, Enrico, Wacziarg, Romain, 2013. How deep are the roots of economic development? Journal of Economic Literature 51 (2).

Stein, Aryeh D., Lumey, Lambert H., 2002. The relationship between maternal and offspring birth weights after maternal prenatal famine exposure: the Dutch famine birth cohort Study. Human Biology 72, 641–654.

Tabellini, Guido, 2008. The scope of cooperation: values and incentives. Quarterly Journal of Economics 123 (3), 905–950.

Weber, Max, [1905,1930] 2005. The Protestant Ethic and the Spirit of Capitalism (Talcott Parsons, Trans.). Routledge, London and New York (translated from German).

CHAPTER FOUR

Family Ties

Alberto Alesina[*],[†] and **Paola Giuliano**[‡]

[*]Harvard University, USA
[†]IGIER Bocconi, Italy
[‡]UCLA Anderson School of Management, NBER and CEPR, USA

Abstract

We study the role of the most primitive institution in society: the family. Its organization and relationship between generations shape values formation, economic outcomes, and influences national institutions. We use a measure of family ties, constructed from the World Values Survey, to review and extend the literature on the effect of family ties on economic behavior and economic attitudes. We show that strong family ties are negatively correlated with generalized trust; they imply more household production and less participation in the labor market of women, young adult, and elderly. They are correlated with lower interest and participation in political activities and prefer labor market regulation and welfare systems based upon the family rather than the market or the government. Strong family ties may interfere with activities leading to faster growth, but they may provide relief from stress, support to family members, and increased well-being. We argue that the values regarding the strength of family relationships are very persistent over time, more so than institutions like labor market regulation or welfare systems.

Keywords

Family values, Cultural economics, Labor market regulations, Growth, Institutions

JEL Classification Codes

J2, J6, O4, O5, Z1

4.1. INTRODUCTION

Economists, sociologists, and political scientists have long been interested in studying the effect of different family structures on a variety of economic outcomes. There is hardly an aspect of a society's life that is not affected by the family.

The aim of this chapter is to review the role that family ties may play in determining fundamental economic attitudes. The importance of the family as a fundamental organizational structure for human society is of course unquestionable. Historical examples of attempts at eliminating the family as an institution have been a catastrophic failure, think of the cultural revolution in China or Cambodian communism. In this chapter we investigate the effects of different types of family values. In particular, we plan to investigate empirically an idea first developed by political scientists and researchers in the late 1960s

ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53538-2.00004-6

and early 1970s, on the importance of family ties in explaining social capital, political participation, and economic outcomes. The family organization can take different forms, with very tight links between members or a more liberal/individualistic structure even within a well-structured and organized family. The idea that a culture based on too-strong family ties may impede economic development is not new. It goes back at least to Weber (1904), who argues that strong family values do not allow the development of individual forms of entrepreneurship, which are fundamental to the formation of capitalistic societies. Another author who clearly described the relationship between family values and under-development is Banfield (1958). In studying differences between the southern and northern part of Italy, this author suggested that "amoral familism" was at the core of the lower level of development of the south. He depicts "amoral familism" as a particular cultural trait: the "inability of the villagers to act together for their common good, or, indeed, for any end transcending the immediate, material interest of the nuclear family. This inability to concert activity beyond the immediate family arises from an ethos—that of "amoral familism" [. . .] according to which people maximize the material, short-run advantage of the nuclear family; and assume that all others will do likewise." This is of course an extreme, and in a sense degenerate, form of family relationship.

This extreme reliance on the family prevents the development of institutions and public organizations, which, on the contrary, require generalized trust and loyalty to the organization. When people are raised to trust their close family members, they are also taught to distrust people outside the family, which impedes the development of formal institutions.

Strong family ties are not unique to the Italian case, but are also present in many Asian and Latin American countries. Fukuyama (1995) for example argues that "though it may seem a stretch to compare Italy with the Confucian culture of Hong-Kong and Taiwan, the nature of social capital is similar in certain respects. In parts of Italy and in the Chinese cases, family bonds tend to be stronger than other kinds of social bonds not based on kinship, while the strength and number of intermediate associations between state and individual has been relatively low, reflecting a pervasive distrust of people outside the family." In a similar vein, Putnam et al. (1993) refer to many cases in Asia and Latin America where the safety and welfare of the individuals are provided by the family, legal authority is weak and the law resented.

When family ties are so strong, the implications for the economy are pervasive. In this chapter we review the literature on the topic, provide new evidence, and explore macroeconomic implications of the effect of family values. We start with within-country analysis. This will allow us to include country fixed effects to isolate the impact of family values from other confounding effects including national institutions. We analyze the relationship between family values and four different types of societal attitudes that have been shown to be conducive to higher productivity and growth. In particular, we look at political participation and political action; measures of generalized morality; attitudes toward women and society; labor market behavior; and attitudes toward work. We perform our analysis using the combined six waves of the World Value Survey (WVS), a collection of surveys administered to a representative sample of people in more than 80 countries from 1981 to 2010. We find that, on average, familistic values are associated with lower political participation and political action. They are also related to a lower level of trust, more emphasis on job security, less desire for innovation, and more traditional attitudes toward working women. On the positive side, family relationships improve well-being as measured by self-reported indicators of happiness and subjective health.

As a second step, we present cross-country evidence linking stronger family ties to economic and institutional outcomes. One obvious limitation of this evidence is that family values may be an outcome rather than a driver of economic development. While we do not offer any definite answer to the question of causality, we do show that family values are quite stable over time and could be among the drivers of institutional differences and level of development across countries: family values inherited by children of immigrants whose forebears arrived in various European countries before 1940 are related to a lower quality of institutions and lower level of development today. We also show that the relationship between economic and institutional outcomes is fairly robust even after controlling for legal origin, which has been shown to be an important historical determinant of formal institutions across countries.

The chapter is organized as follows. In Section 4.2, we review the literature on family ties. In Section 4.3 we provide a logical framework for the empirical analysis, linking our paper to the theoretical models analyzing the impact of culture on economic outcomes. In Sections 4.4 and 4.5, we describe how family ties and family structures can be measured, and review the deep historical determinants of family ties. Section 4.6 presents results from the within-country analysis. Section 4.7 presents cross-country evidence linking stronger family ties to economic development and institutions and shows the persistence of family values and their effect on institutions and development today. Section 4.8 analyzes the impact of family ties on different measures of well-being and Section 4.9 concludes.

4.2. LITERATURE REVIEW

There is surprisingly little systematic empirical evidence in economics on the role played by different types of family values in determining either economic outcomes or attitudes which, in turn, have an influence over economic development. Most of the research in economics indeed focused its attention on institutions, such as political systems (Acemoglu et al. 2001, 2005), the legal rights of the individual (North, 1990), religion (Guiso et al. 2006), education (Glaeser et al. 2004), social capital (Putnam, 2000; Putnam et al. 1993), ethnic fractionalization (Easterly and Levine (1997), and Alesina and La Ferrara (2005) for a survey) to explain a society's ability to generate innovation, wealth, and growth. Yet, little attention has been devoted to the most primitive societal institution, the family, and how this could be relevant in explaining a variety of socioeconomic outcomes.

The work on the relevance of the family starts with Banfield (1958) and Coleman (1990). Both authors notice that societies based on strong ties among family members, tend to promote codes of good conduct within small circles of related persons (family or kin); in these societies selfish behavior is considered acceptable outside the small network. On the contrary, societies based on weak ties, promote good conduct outside the small family/kin network, giving the possibility to identify oneself with a society of abstract individuals or abstract institutions. This initial intuition has been confirmed recently in an experimental setting by Ermish and Gambetta (2010). The authors used a trust game, played by a representative sample of the British population, and found that people with strong family ties have a lower level of trust in strangers than people with weak family ties.

After the seminal contribution of Banfield (1958) and Coleman (1990), some academics have noted strong patterns of family structures and linked them to significant social and economic outcomes. This includes work by Todd (1985, 1990), Greif (2006b), and Greif and Tabellini (2012). Using data on family structures dating back to the Middle Ages, if not earlier, Todd focuses on the distinction between nuclear and extended family. These two family structures differ in the degree of cooperation between subsequent generations, and in the authority exercised by parents. At one extreme, nuclear families are those in which children are emancipated from their parents and leave the household at the time of marriage or before. At the opposite extreme, the extended family typically consists of three generations living together and mutually cooperating under patriarchal authority.

Todd measures the diffusion of both family types across Western Europe and uses this distinction to explain relative levels of diffusion or resistance to important societal changes such as Protestantism, secularism, or political ideology. His general idea is that the nuclear family's tradition of emancipation increases potential for movement away from the family home which can facilitate the pursuit of independent economic opportunities. Also, the inability to rely on the family for income and housing can generate a more entrepreneurial spirit of self-reliance as well as greater motivation to work. Todd's (1990) definition of family structures has been used more recently (Duranton et al. 2009) to explain contemporary outcomes of European regions. The authors identified important links between family types and regional disparities in household size, educational attainment, social capital, labor force participation, sectoral structure, wealth, and inequality.

Greif (2006a) focuses his attention on the distinction between nuclear families and large kinship groups. Like Todd, he emphasizes the sense of independence typical of nuclear family structures. In particular, he describes how the latter in medieval times facilitated the establishment and growth of corporations: "an individual stands to gain less from belonging to a large kinship group, while the nuclear family structure increases its gains from membership in such a corporation (Greif, 2006a: 1–2)." Greif illustrates a feedback effect where causation works in both directions—on the one hand, nuclear families facilitate the establishment of corporations; on the other, the economic and social transformation related to the development of corporations, encourage the domination of the nuclear family across Europe. Nuclear families encourage both flexibility and independence; corporations substitute for kinship groups and provide a safety net, therefore complementing the nuclear family. Greif and Tabellini (2012) distinguish two different modes of sustaining cooperation in China and Europe. In China, the clan (a common descent group consisting of families tracing their patrilineal descent back to one common ancestor who settled in a given locality) was the fundamental institution, which had prevailed for more than 800 years, beginning with the Song Dynasty. Clan-based organizations provided public goods and social safety nets. In Europe, where the nuclear family was more prevalent, the locus of cooperation became the city, whose members were drawn from many kinship groups. The authors show that in a clan, moral obligations are stronger but are limited in scope, as they apply only toward the kin. In a city, moral obligations are generalized toward all citizens irrespective of lineage, but they are weaker.[1] They refer to this distinction as limited versus generalized morality, which is strongly correlated in our paper to the strength of family ties today. The authors show that the prevalence of one or the other organizational form depends on the distribution of values in society. Like Greif (2006a), they recognize the existence of a feedback effect: subsequent social, legal, and institutional developments evolved in different directions in these two parts of the world, strengthening the clan in China and leading to the emergence of strong and self-governed cities in Europe. The authors interestingly exploit differences in the early family structures across different parts of Europe, taking family structures as indicators of the scope and strength of kin-based relations. As expected, historical patterns of urbanization within Europe reflect these different family traditions, with early urbanization being much more diffused in the European regions, where families with weaker ties were more prevalent.

Alesina and Giuliano (2010) analyze systematically the role of the family as primal institution in a society, showing that the strength of family ties represents a fundamental trait shaping economic behavior and attitudes. The authors do not distinguish between nuclear and extended families, like Greif (2006a) and Todd (1985, 1990), but construct a subjective variable on the strength of family ties using three different questions from the World Value Survey. These questions are meant to measure the importance of the family, the love and respect that children are expected to have for their parents, and the parental duties toward their children.[2] Alesina and Giuliano (2010) show that strong family ties are positively

[1] See Tabellini (2008) for a model of limited versus generalized morality which sustains different types of cooperation.

[2] In Section 4.4, we show that there is indeed a strong correlation across countries between nuclear and extended family and family ties as measured by subjective measures taken from the World Values Survey. Alesina et al. (2013) also show that subjective measures of strong family ties are correlated with Todd's definition of extended families at the regional level, at least in the case of Europe.

correlated with home production (a result consistent with the in-depth case study of Italy by Alesina and Ichino (2009)), lower labor force participation of women and young adults, and negatively with geographical mobility. In a companion paper (Alesina and Giuliano, 2011), the authors also establish an inverse relationship between family ties, generalized trust, and political participation. Strength and weakness of family ties, defined as "cultural patterns of family loyalties, allegiances and authorities," also help explaining living arrangements and geographical mobility of young generations (Reher, 1998; Giuliano, 2007), larger fractions of family firms across countries (Bertrand and Schoar, 2006), and cross-country heterogeneity in employment rates (Algan and Cahuc, 2007).

While all the above-mentioned papers take the strength of family values as given and persistent, Alesina et al. (2013) go one step further and explore the presence of a feedback effect between family ties and labor market institutions. The main idea is that in cultures with strong family ties, individuals are less mobile and prefer more regulated labor markets, while weak family ties are associated with more flexible ones, which then require higher geographic mobility of workers to be efficient. In this setup, individuals inherit strong or weak family ties with a certain probability. Strong family ties provide a certain utility to each individual, which is larger, the larger is the share of individuals with strong family ties in a society. Given their utility function, individuals vote with majority rule on labor market regulation. There are two types of labor market policies: labor market flexibility (i.e. laissez-faire) or regulation of wages and employment. Individuals with weak family ties have a higher utility under flexibility, so this regime is voted for if the society starts from a situation in which the majority of the population has weak family ties. On the other hand, the utility of individuals with strong family ties is always higher under regulation. Finally, firms offer labor contracts. A worker with weak family ties always finds a job where he/she is paid for his/her productivity since he/she has no mobility costs. A worker with strong family ties has a moving cost related to the disutility to live far away from his/her family. Labor market regulations are precisely put in place to protect those workers from the monopsony power of firms. The model generates two stable Nash equilibria. One, where everybody chooses weak family ties and then votes for labor market flexibility. In this case, labor market is competitive, everyone is paid his/her marginal productivity and labor mobility is high. The other, where everyone chooses strong family ties and then votes for stringent labor market regulations (firms have a monopsonistic power because workers have a cost of moving away from their original family). If the majority of the population has strong family ties, it is rational to prefer regulated labor markets. This result explains why these types of regulation are hard to change even though prima facie they appear as suboptimal since they generate lower equilibrium employment and wages.

Although the theoretical model points to the possibility of a feedback effect between labor markets regulation and family ties, the empirical part of the paper presents suggestive evidence that the correlation is more likely to run from cultural values to institutions.

The authors present two sets of evidence to make this point. First, they show a strong correlation between family structures today and family structures in the Middle Ages. As a second step, the authors show that family values inherited by immigrants arrived to the US prior to 1940 are correlated to labor market institutions created after WWII.

Family relationships explain the preferences for other aspects of welfare systems. Focusing on Europe, Esping-Andersen (1999) argues that citizens obtain welfare from three basic sources: markets, family, and government. Where family ties are stronger, social risks are more internalized in the family by pooling resources across generations. His idea is that differences in family relations were at the core of the different evolutions of welfare systems, observed after WWII. In particular, he distinguishes three different types of welfare states: the liberal welfare state (typical of countries like the US), this is a regime that favors small public intervention under the assumption that the majority of citizens can obtain adequate welfare from the market. The second example is the social-democratic regime, characterized by its emphasis on universal inclusion and its comprehensive definition of social entitlements. This model, typical of the Nordic European countries, is internationally unique in its emphasis on de-familizing welfare responsibilities, especially with regard to care for children and the elderly. The third, and somewhat more heterogeneous, regime embraces a large part of Continental European countries: Austria, Belgium, France, Germany, Italy, and Spain. This regime is strongly familistic, assuming that primary welfare responsibilities lie with family members.

Coleman (1988, 1990), also stresses the mutual insurance mechanisms provided by old and young generations in familistic societies. He argues that family ties can strengthen the support received by young generations from the old, while at the same time representing an obstacle for innovation and new ideas. Finally, Galasso and Profeta (2012) show that the strength of family ties is related to the type of pension system chosen by a country. Societies dominated by absolute nuclear families (or weak family ties, such as for example the Anglo-Saxon countries) facilitate the emergence of a pension system which acts as a flat safety net entailing the largest within-cohort redistribution than societies dominated by any other type of family.

4.3. CONCEPTUAL FRAMEWORK

Many authors have stressed the relevance of the historical origins of (under) development (North, 1981; Acemoglu et al. 2001) but a still unanswered question is how differences in historical experiences are perpetuated until today. A recent strand of literature focuses on the importance of individual values to explain this persistence. One reason for why individual values can be relevant is the observation that very often, inside the same country, similar institutions work in a very different way. Putnam et al. (1993) used the example of Italy. They pointed out that for distant historical reasons, local governments, courts, schools, and even the private sector are much less efficient in the south than in the north of Italy despite the presence of national institutions. Guiso et al. (2008) recently pushed forward Putnam et al.'s analysis confirming his basic intuition. The authors show that inhabitants of Italian cities that had the status of free city-states at the beginning of the first millennium, where citizens were deeply involved in political life, today also have a higher level of social and civic capital, as measured by participation in elections and a variety of associations, and a higher level of blood donation.

There are different values that can be relevant to explain the sources of underdevelopment in a country. In this chapter we explore the idea that trust restricted only to family members prevents the formation of generalized trust, which is at the core of many collective good outcomes, from political participation to the formation of institutions to economic outcomes (Banfield 1958; Gambetta, 1988; Putnam, et al. 1993; Fukuyama, 1995; Coleman, 1988, 1990). Also the organization of the family as a strong "production unit" implies certain views about living arrangements and the role of women in market activities versus home production (Alesina and Ichino, 2009).

This chapter is part of a rapidly growing literature which emphasizes the relevance of specific cultural traits for economic and political outcomes. Akerlof and Kranton (2011), Alesina et al. (2013b), Guiso et al. (2006), Fernandez and Fogli (2009), Gorodnichenko and Roland (2013), Spolaore and Wacziarg (2009), and Tabellini (2008, 2010) all provide extensive references and illustrate different applications of this new line of research.

The basic idea underlying the empirical analysis of this chapter is that these normative values evolve slowly over time, as they are largely shaped by values and beliefs inherited from previous generations. In particular, a culture of familism, defined as individual values that stress the link between parents and children and loyalty to the family, is an important channel through which distant history can explain the functioning of current institutions and economic development. We explore this idea in two steps, we first use within-country analysis to study the effect of family values on other types of economic attitudes, which are relevant for growth. Although the issue of reverse causality is an important one, we use established evidence that family values today are related to ancient family structures (see Alesina et al. 2013; Duranton et al. 2009; Galasso and Profeta, 2012; Todd, 1990). As a second step, we discuss aggregate evidence looking at differences in institutions and economic outcomes between weak and strong family-ties societies. The correlations shown are strong and consistent with the microeconomic data. Altogether they suggest that well-functioning institutions and development are often observed in countries or regions where individuals have weak family ties.

Before looking at the empirical evidence, we review a logical framework according to which cultural traits in general and family values in particular are relevant. The economics literature has used the word "culture" with different meanings. According to one definition culture refers to the social conventions and individual beliefs that sustain Nash equilibria as focal points in repeated social interactions (Greif, 1994). In more recent contributions, individuals' beliefs are initially acquired through cultural transmission and then slowly updated through experience from one generation to the next. This line of argument has been pursued by Guiso et al. (2010) who build an overlapping generation model in which children absorb their trust priors from their parents and then, after experiencing the real world, transmit their (updated) beliefs to their own children. An alternative interpretation is that culture refers to more primitive objects, such as individual values and preferences (Akerlof and Kranton, 2000). This latter interpretation is consistent with an emerging literature in psychology, sociology, and evolutionary biology that emphasizes the role of moral emotions in motivating human behavior and regulating social interactions.

Following broadly this last approach, we view cultural beliefs as decision-making heuristics or "rules-of-thumb" that are employed in uncertain or complex environments. Boyd and Richerson (1985) show that if information acquisition is either costly or imperfect, it can be optimal for individuals to develop heuristics or rules-of-thumb in decision-making. By relying on general beliefs about the right thing to do in different situations, individuals may not behave in a manner that is precisely optimal in every instance, but they save on the costs of obtaining the information necessary to always behave optimally. In practice, these heuristics often take the form of deeply held traditional values or religious beliefs (Gigerenzer, 2007; Kanhneman, 2011).

The concept of culture as moral principles, rules-of-thumb or normative values that motivate individuals is particularly appealing. Whereas social conventions sometimes change suddenly because of strategic complementarities, and beliefs are updated as one learns from experience or from others, individual values or rules of thumbs are likely to be more persistent and to change slowly from one generation to the next. The reason is not only that normative values are acquired early in life and become part of one's personality but also that learning from experience cannot logically be exploited to easily modify them. Thus, values are likely to be transmitted vertically from one generation to the next, to a large degree within the family, rather than horizontally across unrelated individuals, and persist over time.

There are a number of reasons why we may observe persistence. First, the underlying cultural traits may be reinforced by policies, laws, and institutions, which reinforce the beliefs. A society with familistic values may perpetuate these beliefs by institutionalizing different forms of welfare state, different maternal leave policies, different pension systems. Another source of persistence can arise from a complementarity between cultural beliefs and industrial structure. Beliefs regarding the importance of the family may cause a society to specialize in family-based industries, which reinforce the attachment to the family, therefore perpetuating this trait. A third explanation that does not rely on these forms of complementarity is that cultural beliefs, by definition, are inherently sticky. The benefit of decision-making rules-of-thumb is that they can be applied widely in a number of environments, saving on the need to acquire and process information with each decision.

Empirically, several studies have investigated the persistence of cultural traits by looking at subnational analysis, therefore holding constant industrial structure, domestic policies, and institutions. More directly, looking at children of immigrants, the literature has held constant the external environment. We follow this tradition. In particular, we use within-country analysis to hold constant the presence of institutions and policies. The concern of reverse causality is limited by the fact that several papers have shown that values toward the family today are related to historical family structures (see Alesina et al. 2013; Galasso and Profeta, 2012). Another part of the literature has also shown that many of the outcomes reviewed in this chapter tend to persist among second generation immigrants in the US and other countries as a result of different values regarding the strength of family ties (Alesina and Giuliano, 2010; Alesina et al. 2013).

4.4. HOW TO MEASURE FAMILY TIES

In this section, we describe different ways of measuring family ties using existing datasets. One uses individual responses from the World Value Survey (WVS) (the measure used for the empirical analysis of this chapter); the other is based upon the classification by Todd (1983, 1990).

4.4.1 Measuring Family Ties Using the World Values Survey

The WVS is a cross-country project carried out for more than 20 years. Each wave has representative national surveys of the basic values and beliefs of individuals in a large cross-section of countries. The questionnaires contain information about demographics (sex, age, education), self-reported economic characteristics (income, social class), and answers to specific questions about religion, political preferences, and attitudes. Bertrand and Schoar (2006), Alesina and Giuliano (2010) and several others since, measure the strength of family ties by looking at three WVS variables capturing beliefs on the importance of the family in an individual's life; the duties and responsibilities of parents and children; and the love and respect for one's own parents. The first question assesses how important the family is in one person's life and can take values from 1 to 4 (with four being very important and 1 not important at all). The second question asks whether the respondent agrees with one of two statements (taking the values of 1 and 2, respectively): (1) one does not have the duty to respect and love parents who have not earned it; (2) regardless of what the qualities and faults of one's parents are, one must always love and respect them. The third question prompts respondents to agree with one of the following statements (again taking the values of 1 or 2, respectively): (1) Parents have a life of their own and should not be asked to sacrifice their own well-being for the sake of their children; (2) it is the parents' duty to do their best for their children even at the expense of their own well-being. The questions can be combined by extracting the first principal component from the whole dataset with all individual responses for the original variables.

Table 4.1 displays the correlation at the country level between the three original measures and the first principal component. All the variables are highly and positively

Table 4.1 Correlation among family values

	Family importance	Respect and love parents	Parental duties	Family ties (princ. comp.)
Family importance	1.0000			
Respect and love parents	0.3446**	1.0000		
Parental duties	0.5518***	0.3495**	1.0000	
Family ties (princ. comp.)	0.7217**	0.7944***	0.7928***	1.0000

**significant at 5%.
***significant at 1%.

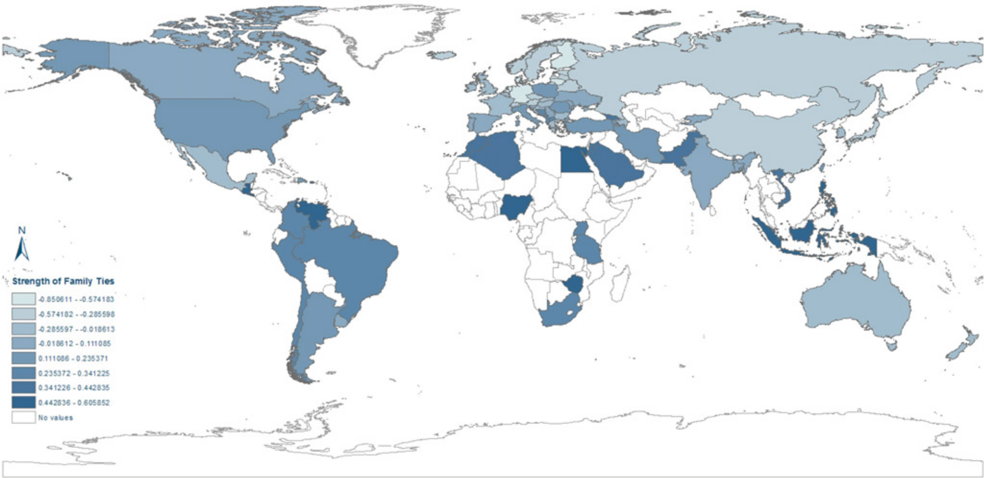


Figure 4.1 Strength of family ties, principal component. *Source: Authors' calculation from the World Value Survey.*

correlated among each other. Figures 4.1–4.4 show maps of each single question and the first principal component, and Figure 4.5 displays the values of the measure of the strength of family ties (expressed using the first principal component) at the country level.[3] The ranking generally reflects priors of the sociological literature. Scandinavian countries and many Eastern European countries tend to have the weakest levels of family ties. In a middle range are France, Canada, the United States, and the United Kingdom. More familistic societies are Italy and many Latin American countries including Colombia, Peru, and Brazil. In the extreme part of the distributions are some Latin American countries like Guatemala and Venezuela; African countries like Egypt and Zimbabwe; and Asian countries like Indonesia, Vietnam, and the Philippines.

[3] The measure is calculated using the six waves from the WVS.

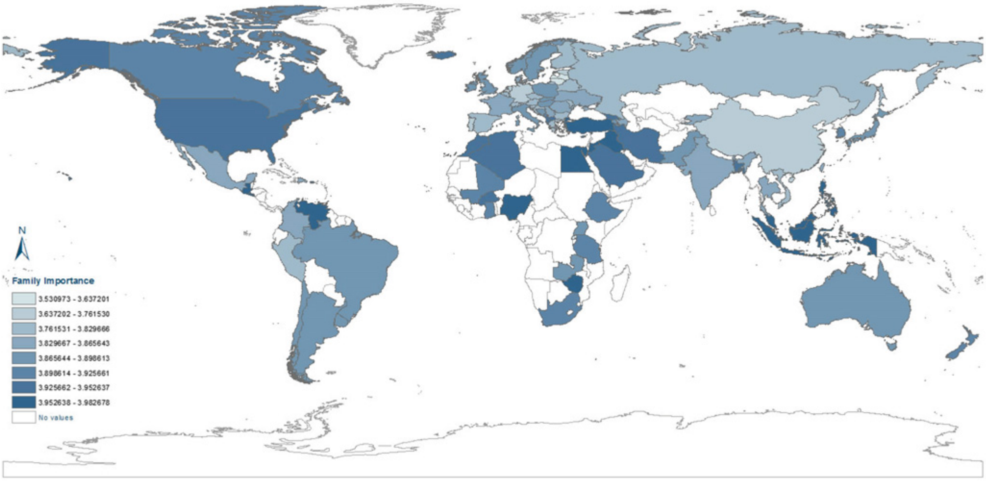


Figure 4.2 Family importance. *Source: Authors' calculation from the World Value Survey.*

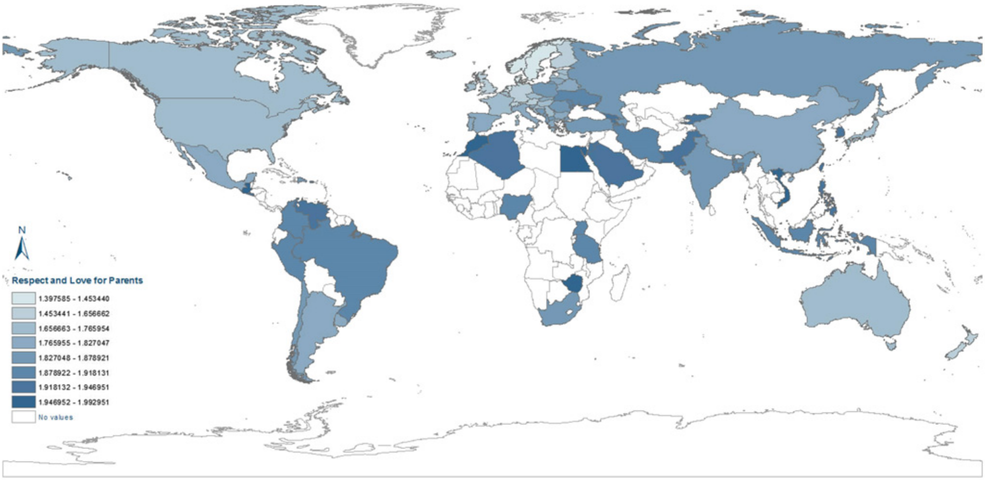


Figure 4.3 Respect and love for parents. *Source: Authors' calculation from the World Value Survey.*

The strength of family ties varies not only across countries, but also across regions of the same country. Figure 4.6 represents the partial correlation of the relationship between generalized trust and the strength of family ties for the case of Europe, after controlling for country fixed effects. As is apparent from the figure, even after controlling for country characteristics, the variation in family ties across Europe is sufficient to explain differences in social capital inside Europe. The difference in the strength of family ties inside the same country can be very pronounced. In Italy, the lowest level of family ties are in the

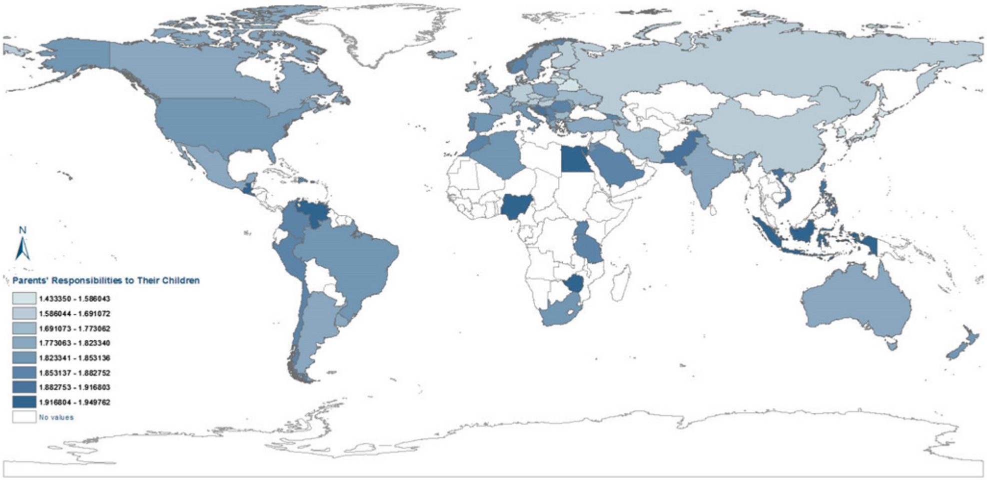


Figure 4.4 Parents' responsibilities to their children. *Source: Authors' calculation from the World Value Survey.*

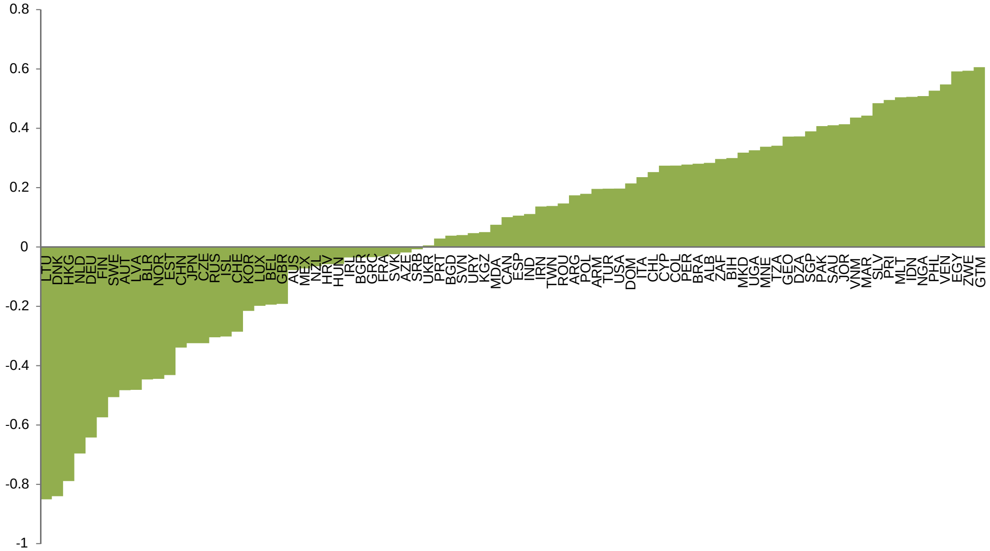


Figure 4.5 Strength of family ties. *Source: Authors' calculation from the World Value Survey.*

northern region of Valle D'Aosta (where it is equal to −0.22, a level similar to some of the Swedish regions), the highest in the southern region of Calabria (where it reaches the high value of 0.44).

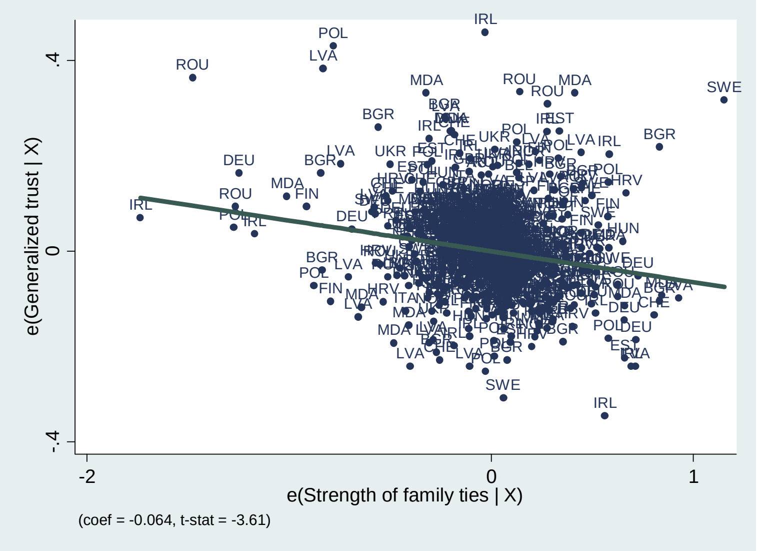


Figure 4.6 Generalized trust and the strength of family ties, regional variation inside Europe.

4.4.2 Todd's Classification of Family Structures

In his books, *The Invention of Europe* (1990) and *The Explanation of Ideology: Family Structures and Social Systems* (1983), Emmanuel Todd classifies family structures according to two main organizing principles. The first principle concerns the vertical relationship between parents and children, the second, the relationship between siblings.

With respect to the vertical relationship between parents and children, the family is defined as "authoritarian" if children are subject to the parental authority even after marriage. The family is defined as "liberal" if children become independent from the parental authority by leaving the parental nest in early adulthood. To measure authoritarian versus liberal families, Todd looked at data on cohabitation between generations within families, in particular between parents and their married children. The family is authoritarian if the eldest son stays in the family when he marries and remains under the authority of the father. Unmarried daughters remain in the family home under the authority of the father or their brothers, when the father dies. In the "liberal" case, children leave the parental home when they reach adulthood or after marriage.

When one looks at the relationship between siblings, the family is defined as "equal" if all siblings are treated equally; it is defined as "unequal" if one particular child (most often the eldest) has a privileged treatment. To measure equality, Todd uses data on inheritance laws and practices. A family is equal when family property is divided evenly between siblings and unequal if primogeniture (or in some cases ultimogeniture) exists. The information on the type of families for both the vertical and the horizontal dimension is obtained by censuses and historical monographs that go back more than 500 years.

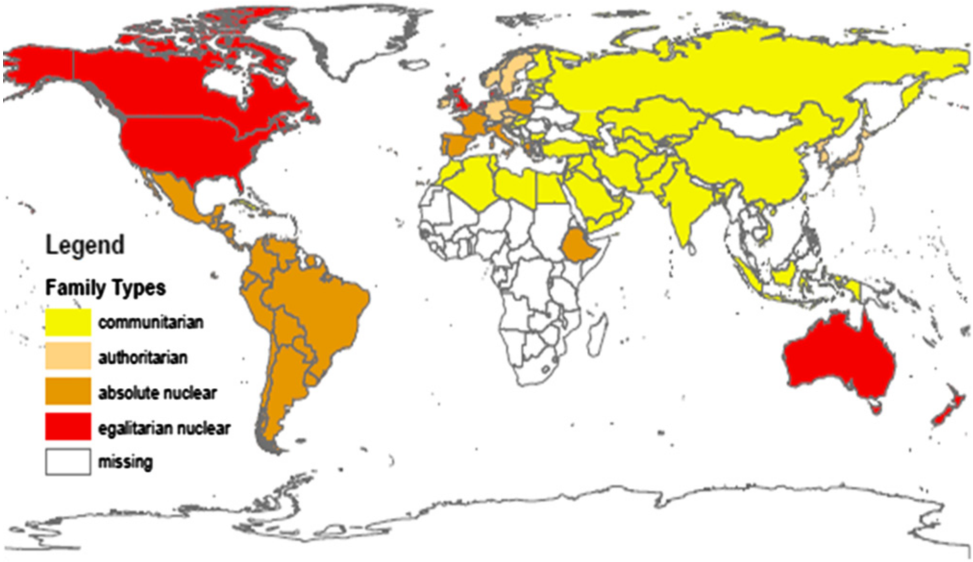


Figure 4.7 Family structures, Todd's classification. *Source:* *Profeta and Galasso (2012)*.

The combination of the authoritarian/liberal vertical relationship with the equal/unequal horizontal relationship gives rise to four types of family structures (depicted in Figure 4.7):

1. *Absolute nuclear family:* this family type is characterized by independent living arrangements (children leave their family in early adulthood either before marriage or to form their own family), and lack of stringent inheritance rules. In this type of family, parents have no obligation to support their adult children; every person is independent and has to rely on his/her individual effort. The United States, the UK, Australia, New Zealand, the Netherlands, and Denmark belong to this group. Interestingly, Laslett (1983) has shown that this family characteristic makes young adults free to take residence where job opportunities are best and thus has favored industrial development.
2. *Egalitarian nuclear family:* this family type is characterized by independent living arrangements, like in the absolute nuclear family. The presence of egalitarian inheritance rules, however, encourages the persistence of a strong relationship between parents and children, who are inclined to stay with their parents longer. To this group belong the southern European countries (Italy, Spain, Greece, and Portugal); Romania, Poland, Latin America, and Ethiopia.
3. *Stem or authoritarian family:* this family type is characterized by the cohabitation of parents and children. Inheritance rules are also not egalitarian. Countries belonging to this group are Austria, Germany, Sweden, Norway, Czech Republic, Belgium, Luxembourg, Ireland, Japan, Korea, and Israel.

Table 4.2 Relationship between the strength of family ties (WVS) and Todd's family structure

	(1) Family important	(2) Respect and love parents	(3) Parental duties
Communitarian family	0.039	−0.135**	0.086***
	(0.040)	(0.065)	(0.031)
Authoritarian family	0.019	0.012	0.163***
	(0.033)	(0.088)	(0.049)
Nuclear egalitarian family	0.018	−0.142**	0.014
	(0.035)	(0.065)	(0.025)
Observations	101,169	94,631	89,011
R-squared	0.007	0.037	0.028

Source: Galasso and Profeta (2012). A higher number in their specification indicates weaker family ties. Data are taken from the WVS. Each specification controls for a quadratic in age, education, income, and political orientation.
*Indicates significance at the 10% level.
**Indicates significance at the 5% level.
***Indicates significance at the 1% level.

4. *Communitarian family:* this type of family is characterized by cohabitation of parents and children and equal inheritance rules. This system characterizes countries like Russia, Bulgaria, Finland, Hungary, Albania, China, Vietnam, Cuba, Indonesia, and India.[4]

Galasso and Profeta (2012) compare Todd's classification of family structures with the one used in this chapter and in Alesina and Giuliano (2010). In particular, they use the three above-described measures of family values taken from WVS and compare them with Todd's classification of family structures. They run a model of the following type:

$$y_i = \alpha + \beta_1 X_i + \beta_2 Communitarian_i + \beta_3 Authoritarian_i + \beta_4 nuclear\_egalitarian_i + \varepsilon_i,$$

where y_i is the answer from the WVS to each of the three family measures, X_i is a set of individual controls (a quadratic in age, income, education, political views). They include dummies for the prevalent type of family in a country, where the absolute nuclear family is the excluded category. Table 4.2 reports the results of their specification. Todd's classification plays no role in explaining the answer to the most general question on the importance of the family (column 1). However, strong children-to-parents links are associated with communitarian and egalitarian nuclear families (column 2). Finally, authoritarian and communitarian families are associated with a prominent role of parents in today's societies. The authors conclude that current survey data broadly confirms the historical types present in Todd's analysis.

[4] Note that Todd (1990) provides regional variations for most European countries, for example the communitarian family was present in the center of Italy. Here we just report the data at the country level. The family type at the country level is based on the type of family present in the majority of the population. For more details on the regional variations of family ties see Duranton et al. (2009) and Todd (1990).

4.5. WHERE FAMILY TIES COME FROM

A large literature in anthropology has documented that the type of family is related to ecological features and means of subsistence in ancient times (Murdock, 1949). Typically, agricultural societies are characterized by large extended families; whereas the small nuclear family is more prevalent among small hunting and gathering societies. The reason for that is that farming requires the help of many people, usually children and kin, who cooperate to cultivate crops. Studies have found that children in agricultural and pastoral societies are taught to be responsible, compliant, obedient, and to respect the elderly and the hierarchy. Hunting or gathering as a means of subsistence, on the other hand, requires moving from area to area. Many hunting and gathering societies do not have a permanent home, but temporary huts or shelters. Mobility means that the small nuclear family is more adaptable for survival under these ecological restraints. Children in hunting and gathering societies tend to be self-reliant, independent, and achievement oriented; and the family is less stratified.

We are not aware of formal tests of whether these ecological features from the distant past tend to persist to the modern times, after industrialization has taken place in many societies. The only work which has studied the correlation between current measures of family ties and long-term historical characteristics is Durante (2010). He proposes a simple explanation of the emergence of trust and different forms of family structures based on the need for subsistence farmers to cope with weather fluctuations. The main idea is that a more variable environment should increase an individual's propensity to interact with non-family members and reduce their dependence on the family for insurance purposes. Durante (2010) tests his prediction in the context of Europe, combining high-resolution climate data for the period 1500–2000 with contemporary survey data on family ties as measured in Alesina and Giuliano (2010), and generalized trust, using the negative expected relationship between these two variables. He finds that regions with greater interannual fluctuations in temperature and precipitation have higher levels of interpersonal trust and weaker family ties. This result is primarily driven by weather variability in the growing-season months, consistent with the effect of climatic risk operating primarily through agriculture. He then replicates the analysis using climate data for the period 1500–1750. The relationship between historical climatic variability and trust and weak family ties is positive and significant, even after controlling for climate variability between 1900 and 2000, which does not appear to have an independent effect on trust or family ties. These findings support an explanation based on the historical formation and long-term persistence of trust and family attitudes.

The results of Durante's specifications for various regions of Europe are reported in Table 4.3. In particular, in panel A we report Durante's results for the period 1900–2000. The left-hand side variable is the principal component of the measures of family

Table 4.3 Family ties and climate variability

	Family ties (principal component from WVS)			
	Precipitation		Temperature	
Climate data 1900–2000	**(1)**	**(2)**	**(3)**	**(4)**
Panel A: Climate data: 1900–2000				
Variability	−0.072**		−0.392*	
(12 months)	(0.033)		(0.214)	
Variability		−0.081***		−0.692***
(growing-season months)		(0.029)		(0.219)
Variability		−0.004		0.063
(non-growing-season months)		(0.024)		(0.130)
Observations	220	220	220	220
Number of clusters	24	24	24	24
R-squared	0.826	0.828	0.826	0.832
Panel B: Climate data: 1500–1750 and 1900–2000				
Variability (growing-season months)	−0.205**	−0.300**	−0.205**	−0.306***
(1500–1750)	(0.085)	(0.112)	(0.081)	(0.100)
Variability (growing-season months)		0.129*		0.138
(1900–2000)		(0.074)		(0.081)
Observations	218	218	218	218
Number of clusters	24	24	24	24
R-squared	0.830	0.833	0.785	0.789

Source: Durante (2010). The regressions control for country fixed effects and for the following regional controls: mean temperature, mean precipitation, average ruggedness index, soil suitability (average and standard deviation), area, dummy for landlocked, distance of the region's centroid from the coast, number of major rivers passing through the region, latitude of the region's centroid. Robust standard errors clustered at the country level in parenthesis.
*Indicates significance at the 10% level.
**Indicates significance at the 5% level.
***Indicates significance at the 1% level.

ties, whereas the dependent variable is the annualized variability calculated using both precipitation (columns 1 and 2) and temperature (columns 3 and 4). The coefficient on precipitation variability is positive and statistically significant at the 5% level (column 1): in regions characterized by a more variable climate, family ties are weaker. The results are primarily driven by variability in precipitation during the growing-season months, whereas variability during the other months displays no significant effect (column 2). The results obtained using temperature are analogous: higher interannual variability, particularly during the growing season, corresponds to weaker family ties (columns 3 and 4).

Panel B reports the test Durante performed to show that differences in the strength of family ties are related to historical rather than contemporary variability. Historical variability in the growing season's precipitation and temperature appear to have a negative, large, and significant effect on the strength of family ties (column 1). This effect remains and becomes even larger when controlling for climate variability over the last century, which appears to have no significant (or even positive effect, for the case of precipitation) effect. The magnitude of the coefficients on historical variability is large: a one standard deviation in growing season variability corresponds to a 0.40 standard deviation decrease in the strength of family ties, for precipitation, and a 0.38 standard deviation decrease for temperature.

4.6. EMPIRICAL ANALYSIS

In this section we examine the relationship between family values and economic attitudes, using within-country analysis drawn from the WVS. Our measure of family ties is defined as the principal component of three subjective measures regarding the role of the family, and the link between parents and children, as described in Section 4.1. We use all available six waves, therefore providing the most comprehensive analysis of the impact of family values on a variety of attitudes.[5] The coverage of countries varies across surveys. The 1981–1984 wave covers 24 countries; the 1989–1993 wave covers 43 countries; 1994–1999, 1999–2004, 2005–2007, and 2008–2010 waves cover, respectively, 54, 70, 57, and 47 countries.

The use of within-country analysis allows us to control for country fixed effects, eliminating the impact of other institutional variables. This approach underestimates the effect of family ties, to the extent that in the distant past they had an impact on current institutions. Nevertheless, the effect can be attributed more credibly to this cultural trait. Omitted variables and reverse causality can still be a problem for this type of regression, for this reason, we prefer to interpret our results as more precisely estimated partial correlations. We divide our dependent variables into four groups.

4.6.1 Measures of Interest in Politics and Political Action

We begin with measures of people's interest in politics and political action. The first variable, which we label interest in politics, is based on the following question: "How interested would you say you are in politics?", the response varies from 1 (not at all interested) to 4 (very interested). Variable 2, which we label discuss politics, asks the respondent "How often do you discuss political matters with friends?" with the response varying from never (1), occasionally (2), to frequently (3). Variables 3 and 4 measure if the respondent belongs to political parties (the first question measures it with a dummy if the person belongs to a political party and zero otherwise; the second question can take values from 0 to 2, with 0 (not a member), 1 (inactive member), and 2 (active member).

[5] Alesina and Giuliano (2010) only used four waves, having a substantially smaller sample size.

The last five questions measure different forms of political action, asking the respondent whether he/she has actually done any of these things (taking the value of 3), whether he/she might do it (2), or whether he/she would never do it (1): signing a petition, joining in boycotts, attending lawful/peaceful demonstrations, joining unofficial strikes, occupying buildings or factories.

Understanding the origin of civic culture and of a well-educated population is an important prerequisite to a well-functioning and stable democracy (Lipset 1959; Almond and Verba, 1963; Glaeser et al. 2004, 2007; Persson and Tabellini, 2009).

4.6.2 Measures of Generalized Morality and Attitudes Toward Society

The second group of questions contains two measures of generalized morality (related to a definition by Tabellini (2008), explained below), one question about trust in the family and three questions about attitudes toward society. Variable one, *trust*, is based on the following question: "Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?", the variable is equal to 1 if participants report that most people can be trusted and 0 otherwise. Variable 2 asks whether obedience is a quality that children can be encouraged to learn at home, taking the value of 1 if the quality is mentioned and 0 otherwise. Variable 3 asks how much the respondent trusts the family from "do not trust the family at all" (1), "do not trust the family very much" (2), "neither trust nor distrust the family" (3), "trust the family a little" (4), "trust the family completely" (5). The last three questions refer to attitudes about the possibility of changing society. The first question asks on a scale from 1 to 10 whether "Ideas that stood the test of time are generally best" (taking the value of 1), or whether "New ideas are generally better than old ones". The second question asks if "One should be cautious about making major changes in life" (taking the value of 1) versus "You will never achieve much unless you act boldly" (taking the value of 10). The third question asks the respondent to choose between three basic kinds of attitudes concerning society: "society must be valiantly defended" (taking the value of 1), "society must be gradually improved by reforms" (taking the value of 2), and "society must be radically changed" (taking the value of 3).

Among all the above variables "trust" measures a fundamental trait in a society. More than 35 years ago, Arrow (1972), recognizing the pervasiveness of mutual trust in commercial and non-commercial transactions, went so far as to state that "it can be plausibly argued that much of the economic backwardness in the world can be explained by the lack of mutual confidence (p. 357)." Since then, Arrow's conjecture has received considerable empirical support. A vast literature investigates the link between aggregate trust and aggregate economic performance, trust also encourage welfare enhancing social interactions, such as anonymous exchange of participation in the provision of public goods, and they are likely to improve the functioning of government institutions. Starting with Banfield, it has also been postulated a negative correlation between trust in a small related circle (like the family) and generalized trust. Platteau (2000) links lack of generalized trust to the distinction between "generalized" versus "limited" morality. In hierarchical societies, trust and honest behavior are often confined to small circles of related people (like members of the family). Outside of this small network, opportunistic and highly selfish behavior is regarded as natural and morally acceptable. These two measures have been defined to distinguish between values consistent with "generalized" versus "limited" morality. Tabellini (2008) has shown that generalized morality is fundamental to understand the origin of economic development across countries and among regions of Europe. We therefore look at the relevance of family ties in the formation of generalized trust and trust toward the family (expecting a negative impact of family ties on generalized trust and a positive impact on trust in the family). In strong family ties societies, individualism is also mistrusted. In familistic societies, the role of parents is to foster obedience. Banfield emphasized the relevance of obedience to claim that such coercive cultural environment reduces individual initiative and cooperation within a group, and can hurt growth and development.

The last three questions are related to the idea put forward by Coleman (1988) that family ties can represent an obstacle for innovation and new ideas.

4.6.3 Labor Market and Attitudes Toward Work

The third group of questions looks at the relationship between family values and the labor market. We explore the correlation between female, youth, and elderly labor force participation and family ties. We also look at questions regarding the relationship between job security and family ties. One question asks the respondent how important is job security in a job. In another, the respondent has to choose the most important thing in looking for a job, where a safe job with no risk is one of five choices (the other four being: a good income, working with people one likes, doing an important job, doing something for the community).

Employment rates vary dramatically across countries, but the bulk of the variation relies on specific demographic groups: women, younger, and older individuals. Looking at micro and macro data for OECD countries, Algan and Cahuc (2007) show that differences in family culture can explain lower female employment and lower level of employment of young and older people in Europe.[6] In the same fashion, Giavazzi et al. (forthcoming) find that culture matters for women's employment rates and for hours worked. In a recent

[6] Although the authors attribute the differences in employment rates to the presence of the nuclear versus the extended family in different OECD countries, the effect on employment is not studied using different family structures but considering some subjective measures. In particular, they look at three questions: whether the respondent agrees with the statement "When jobs are scarce, older people should be forced to retire from work early"; the second asking the respondent whether they agree with the statement "Adult children have a duty to look after their elderly parents"; and finally, "Independence is a quality that children should be encouraged to learn at home."

paper, Alesina et al. (2013) looked at the relationship between family ties and the labor market. The main idea is that in cultures with strong family ties, moving away from home is costly. Thus individuals with strong family ties choose regulated labor markets to avoid moving and limiting the monopsony power of firms, even though regulation generates lower employment and income. We look at within-country analysis on preferences for job security that further limit the possibility that the results are driven by other country characteristics.

4.6.4 Measures of Attitudes Toward Women

The fourth group of variables contains measure of people's attitudes toward women. The first question asks the respondent whether he/she agrees with the statement "When jobs are scarce, men should have more right to a job than women." The other six variables come from the answer to the question "For each of the following statements I read out, can you tell me how much you agree with each? Do you agree strongly, agree, disagree, or disagree strongly?" The statements are: "A working mother can establish just as warm and secure a relationship with her children as a mother who does not work"; "Being a housewife is just as fulfilling as working for pay"; "On the whole, men make better political leaders than women do"; "A university education is more important for a boy than for a girl"; "A pre-school child is likely to suffer if his or her mother works"; "A job is alright but what most women really want is a home and children." We recode the questions so that a higher number means a more traditional perception of the role of women in society.

Gender role attitudes are relevant in explaining differences in female labor force participation across countries (see Fortin, 2005; Fernandez and Fogli, 2009). In strong family ties societies (Esping-Andersen, 1999; Ferrera, 1996; Castles, 1995; Korpi, 2000), family solidarity is based on an unequal division of family work between men and women (what has been called the "male-breadwinner hypothesis"): weak family ties will foster an egalitarian gender role in which men and women participate equally in employment and housework, whereas strong family ties are based on the "male-breadwinner hypothesis" in which men work full-time and women dedicate themselves to housework. In the more traditional, strong family ties societies, is the woman who is supposed to fulfill the family obligations and as such, participate less in the market. According to Esping-Andersen (1999), this gender relationship has been helped by a welfare state model that has historically delegated family care services for children and the elderly to the family sphere and has protected the male-breadwinner figure. Alesina and Ichino (2009) provide an in-depth analysis of this type of family organization with respect to Italy.

4.6.5 The Impact of Family Ties

In Tables 4.4–4.7, we present our results on the overall effects of family ties. Each attitude is regressed on our measure of family ties, some control variables (age, education, marital

Table 4.4 Family ties and political participation

Variables	(1) Interest in politics	(2) Discuss politics	(3) Belong to political parties	(4) Membership political party	(5) Sign petition	(6) Join in boycotts	(7) Attend demonstrations	(8) Join unofficial strikes	(9) Occupy buildings
Family ties	−0.010***	−0.006***	−0.002***	−0.004**	−0.029***	−0.046***	−0.036***	−0.041***	−0.026***
	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
Age	0.016***	0.018***	0.002***	0.006***	0.012***	0.008***	0.009***	0.004***	−0.000
	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Age squared	−0.000***	−0.000***	−0.000***	−0.000***	−0.000***	−0.000***	−0.000***	−0.000***	−0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	−0.277***	−0.189***	−0.034***	−0.083***	−0.088***	−0.123***	−0.155***	−0.099***	−0.058***
	(0.004)	(0.003)	(0.001)	(0.004)	(0.004)	(0.003)	(0.004)	(0.003)	(0.002)
Married	−0.008	−0.016**	0.010**	0.009	0.002	0.023***	0.018**	0.044***	0.029***
	(0.011)	(0.007)	(0.004)	(0.008)	(0.009)	(0.009)	(0.009)	(0.008)	(0.006)
Education dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes
Country dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes
Wave dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	212,931	220,148	133,684	66,407	131,066	127,491	131,408	126,513	125,180
R-squared	0.136	0.115	0.060	0.181	0.278	0.182	0.143	0.096	0.096

Coefficients are reported with robust standard errors in brackets.

*Indicates significance at the 10% level.

**Indicates significance at the 5% level.

***Indicates significance at the 1% level.

Table 4.5 Family ties, generalized morality, and attitudes toward society

	(1)	(2)	(3)	(4)	(5)	(6)
			Children		Major	Society
		Trust the	qualities:	New and	change	changed/society
	Trust	family	obedience	old idea	in life	defended
Family ties	−0.006***	0.069***	0.024***	0.050***	0.112***	0.017***
	(0.001)	(0.008)	(0.001)	(0.009)	(0.010)	(0.002)
Age	0.002***	0.002	−0.004***	0.027***	0.015***	−0.002***
	(0.000)	(0.003)	(0.000)	(0.003)	(0.004)	(0.001)
Age squared	−0.000***	−0.000	0.000***	−0.000	0.000	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	−0.006***	−0.002	−0.003	0.090***	0.086***	0.032***
	(0.002)	(0.014)	(0.002)	(0.019)	(0.022)	(0.003)
Married	−0.013***	−0.083*	0.002	−0.002	−0.182***	−0.003
	(0.005)	(0.043)	(0.005)	(0.042)	(0.052)	(0.008)
Education dummies	yes	yes	yes	yes	yes	yes
Country dummies	yes	yes	yes	yes	yes	yes
Wave dummies	yes	yes	yes	yes	yes	yes
Observations	217,647	9,802	220,639	81,640	69,736	110,077
R-squared	0.104	0.057	0.111	0.131	0.083	0.050

Coefficients are reported with robust standard errors in brackets.
*Indicates significance at the 10% level.
**Indicates significance at the 5% level.
***Indicates significance at the 1% level.

Table 4.6 Family ties, labor market, and attitudes toward work

	(1) Female LFP	(2) Youth LFP	(3) Elderly LFP	(4) Job security	(5) Job security in looking for job
Family ties	−0.013***	−0.012**	−0.006**	0.017***	0.022***
	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)
Age	0.063***	−0.043***	−0.050	0.003***	0.004***
	(0.001)	(0.007)	(0.043)	(0.000)	(0.000)
Age squared	−0.001***	0.001***	−0.000	−0.000***	−0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Married	−0.006	−0.060***	0.028	−0.000	−0.001
	(0.007)	(0.007)	(0.020)	(0.005)	(0.007)
Female		−0.268***	−0.264***	−0.004**	−0.003
		(0.003)	(0.005)	(0.002)	(0.003)
Education dummies	yes	yes	yes	yes	yes
Country dummies	yes	yes	yes	yes	yes
Wave dummies	yes	yes	yes	yes	yes
Observations	98,218	44,336	26,974	213,576	99,749
R-squared	0.224	0.269	0.251	0.106	0.049

Coefficients are reported with robust standard errors in brackets.
*Indicates significance at the 10% level.
**Indicates significance at the 5% level.
***Indicates significance at the 1% level.

status, and a gender dummy[7]), country-specific effects, and wave dummies. The sample size differs across regressions and ranges from a minimum of 26,974 to a maximum of 212,931[8]; therefore always providing substantial variation in time period and number of countries.

Before we comment on the results of the impact of family ties, it is useful to discuss the effect of our control variables. The results, which are of independent interest, are very reasonable and provide credibility to the measure of family ties we are going to use. There is a hump-shaped relationship in age between interest in politics, political participation, and political action, and between age and job security. There is also a hump-shaped relationship between age and trust, whereas the level of trust in the family does not change with age. Emphasizing obedience is less important among young people and it has a U-shaped relationship with age. The same U-shaped relationship also exists for

[7] We do not include income in our regressions since in the next section we do find that family ties could explain part of the differences in GDP per capita across countries. Our results are, however, robust to its inclusion.

[8] The smallest sample is for labor force participation of the elderly (26,974), therefore the smaller sample size depends on the fact that the regressions are not run on the whole population. The variable trust in the family is the one with substantially lower sample size, of around 10,000 observations.

Table 4.7 Family ties and attitudes toward women

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				Men	University	Child	Women
		Working	Housewife	political	important	working	home
	Job scarce	mother	fulfilling	leaders	for girls	mother	children
Family ties	0.015***	−0.002	0.044***	0.023***	0.008***	0.043***	0.075***
	(0.001)	(0.002)	(0.002)	(0.003)	(0.003)	(0.004)	(0.004)
Age	0.001***	−0.001	0.001*	0.002	0.001	0.006***	0.003**
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Age-squared	0.000**	0.000***	0.000**	0.000	0.000*	−0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	−0.117***	−0.154***	−0.068***	−0.279***	−0.221***	−0.109***	−0.068***
	(0.003)	(0.005)	(0.005)	(0.006)	(0.005)	(0.009)	(0.009)
Married	0.019***	0.009	−0.014	0.007	0.032***	−0.005	0.016
	(0.006)	(0.012)	(0.012)	(0.013)	(0.012)	(0.043)	(0.049)
Education dummies	yes	yes	yes	yes	yes	yes	yes
Country dummies	yes	yes	yes	yes	yes	yes	yes
Wave dummies	yes	yes	yes	yes	yes	yes	yes
Observations	118,200	133,811	130,836	100,679	103,027	29,929	29,153
R-squared	0.234	0.086	0.092	0.203	0.123	0.169	0.190

Coefficients are reported with robust standard errors in brackets.
*Indicates significance at the 10% level.
**Indicates significance at the 5% level.
***Indicates significance at the 1% level.

the attitudes looking at whether society should be defended versus whether it should be dramatically changed. Not surprising, young people believe that new ideas are better than old ones and are more open to major changes in life. Attitudes toward women are not systematically related to age. Gender and education also have the expected effects. Women are generally less interested and involved in politics than men. They also trust less (gender, like age, however, is not systematically related to the level of trust in the family, a more universal value that does not change with specific demographics). Not surprisingly, women have less traditional beliefs about the role of women in society compared to their male counterparts (an indication that they most likely suffer from the presence of traditional gender role attitudes). Education is positively related to political interest and political action, a result supporting the model by Glaeser et al. (2007). More educated people have a higher level of trust, less traditional attitudes about the role of women in society; they also believe obedience is not an important quality to teach children. Finally, they are in support of new ideas but more conservative with respect to major changes in life and in society.[9]

Let's now consider the effect of family ties. Table 4.4, which relates to political participation and political action, shows that family ties have a negative and highly statistically significant coefficient. Regarding the magnitude of the effect, the beta coefficient of family ties on political participation (the first four columns of Table 4.4) is equal for the four different measures to 0.01 (roughly to 1/5 of the magnitude of the beta coefficient of the middle level of education, which ranges between 0.04 and 0.05).[10] The magnitude of the beta coefficient for family ties is larger for the measures of political actions. In this case the coefficient goes from 0.04 to 0.08 and it is between 1/3 or even the same effect of the middle level of education.

Table 4.5, which includes the same controls of Table 4.4, refers to those variables of "generalized morality" (as in Tabellini, 2008) and openness to new ideas. The results are as expected. Particularly important is the result of column 1 which shows a negative effect of family ties on generalized trust, but positive on trusting family members (column 2). Strong family ties imply teaching more obedience to children (column 3) and being relatively conservative in terms of personal and social change (columns 4, 5, and 6). As for the magnitude of the effects: the beta coefficients of family ties on trust is equal to −0.016 (half the coefficient of middle level of education, which has a positive effect compared to the lower level of education). The impact of family ties on trusting the family is three

[9] When we control for income as one of our robustness checks, we do find that income is positively correlated with trust and trust in the family, like education. Similarly, income is inversely correlated with the importance of obedience. Income is however inversely correlated with the importance of new ideas and major changes in life, but positively correlated with the belief that society should be changed.

[10] We include two dummies for education: one for middle and one for upper level (the excluded group is lower level of education). The sign of the middle and upper level of education coefficient is positive, as the excluded group is lower level of education.

times the effect of middle level of education; the magnitudes of middle level of education and family ties are equivalent (but of opposite sign) for obedience and the three attitudes on personal and social change (columns 4–6).

Table 4.6 looks at the labor market of women, young adults, and the elderly. Individuals coming from strong family ties have a lower level of labor force participation for women, young adults, and older people. This is consistent with the male-breadwinner hypothesis according to which, women are supposed to stay at home and take care of the family, together with older and younger people. Consistent with the relationship postulated by Alesina et al. (2013), individuals with familistic values consider job security the most important characteristic in a job. The impact of family ties on the labor force participation of the three groups is small compared to the impact of education (the beta coefficient is 1/10 when compared to the one on middle level of education). This is not surprising: family ties are very relevant in the determination of labor market institutions (see Alesina et al. 2013) and the country fixed effects are most likely capturing part of that channel. The impact of family ties on job security (columns 4 and 5) on the other hand is six times larger than the effect of middle level of education.

Table 4.7 refers to the attitudes towards women. With the exception of column (2), in all other columns the variable family ties has the expected sign and it implies a more traditional role of women in the family. Indeed, this makes sense: with close family ties, the family needs someone who organizes it, and keeps it together, typically the wife and mother. In this sense, the family becomes a formidable producer of goods and services which are not counted in standard measure of GDP, like childcare, care of the elderly, and various other forms of home production.[11] As for the magnitude of the effects, it goes from roughly $\frac{1}{4}$ of the effect of middle level of education (for the first four columns) to being more or less of equivalent magnitude (for the last three columns).

Overall, we find that different beliefs about the importance of the family in one person's life and the relationship among generations are relevant for the determination of values, which have been proven to promote employment, innovation, and growth. If values about the family are transmitted from generation to generation and they move slowly over time, they could provide an explanation on how the distant past can affect the current functioning of institutions. Indeed, several papers have provided evidence that attitudes toward the family and different forms of family structures are transmitted from generation to generation and affect the behavior of second generation immigrants, who still maintain the values and behavior of their parents despite living in an institutional environment which is very different than their ancestors' country of origin.[12] It is also

[11] See Alesina and Ichino (2009) for an empirical estimate of the size of home production in a few countries with weak or strong family ties.

[12] See Alesina and Giuliano (2010, 2011) and Alesina et al. (2013). All these papers show that family ties have an effect on the behavior of second generation immigrants in the US and a large set of European countries. This evidence hints at the possibility that the partial correlations established in Section 4.4 can have causal nature.

worth noticing that all the results presented in this section are most likely a lower bound of the effect of family ties. If family values become part of the national culture, this is captured by the country fixed effects together with the impact of institutions and all other time invariant characteristics.

4.7. FAMILY TIES, DEVELOPMENT, AND INSTITUTIONS

In this section, we provide some suggestive evidence in support of the idea that family ties are correlated with fundamental determinants of economic outcomes at the aggregate level. We document a strong correlation between the strength of family ties, economic development, and quality of institutions. Countries with strong family ties have lower levels of per capita GDP and lower quality of institutions.

We do our analysis in two steps. As a first step, we establish a basic correlation between the strength of family ties, economic development, and the quality of institutions. As a second step, a small one toward establishing causation, we show that family values brought by immigrants who arrived in several destination countries before 1940 are correlated with the level of development and the quality of institutions today.

We measure economic development with real GDP per capita. As a measure of institutional quality we use the Worldwide Governance Indicators (WGI) of the World Bank. The WGI reports on six broad dimensions of governance for over 200 countries for the period 1996–2011. These dimensions are: voice and accountability (the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media); political stability and absence of violence (measuring perceptions of the likelihood that the government will be destabilized or overthrown by unconstitutional or violent means, including politically motivated violence and terrorism); government effectiveness (the quality of public services; the quality of the civil service and the degree of its independence from political pressures; the quality of policy formulation and implementation; and the credibility of the government's commitment to such policies); regulatory quality (the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development); rule of law (capturing perceptions of the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, property rights, the police, and the courts, as well as the likelihood of crime and violence); and control of corruption (the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as "capture" of the state by elites and private interests).

4.7.1 The Correlation Between Family Ties, Economic Development, and Institutional Quality

We first establish that countries with stronger family ties have lower economic development on average, measured by GDP per capita (Table 4.8). We run cross-country

Table 4.8 Family ties and per capita GDP

Variables	(1) Log GDP	(2) Log GDP	(3) Log GDP	(4) Log GDP
Family ties	−1.984***	−0.969**		
	(0.383)	(0.441)		
Inherited family values			−0.860**	−0.786***
			(0.428)	(0.285)
Log (years of schooling)		2.414***		2.350***
		(0.498)		(0.307)
Observations	80	73	122	100
R-squared	0.221	0.409	0.064	0.522

Coefficients are reported with robust standard errors in brackets.
*Indicates significance at the 10% level.
**Indicates significance at the 5% level.
***Indicates significance at the 1% level.

regressions of GDP per capita on our measures of family values.[13] We show that the coefficient from a regression of logarithm of GDP per capita on the strength of family ties is highly negative and significant. A one standard deviation increase in the strength of family ties (0.36) is associated with a reduction of the log of GDP per capita of 0.71 (roughly equal to 44% of its standard deviation). The second column controls for human capital, measured by the logarithm of the average schooling years in the total population over age 15. By adding this variable, we might be overcontrolling since educational choices might themselves be an outcome of family values. The strength of family ties is still very strong although is magnitude is, not surprisingly, reduced.

The cross-sectional correlations leave open the possibility that other omitted variables can explain both the strength of family ties and differences in economic development across countries. Using the combined waves of the WVS we can limit this possibility by looking at the correlation between regional income and regional family ties, after controlling for country fixed effects. The results are reported in Table 4.9. In order to maintain a very large sample (more than 1000 regions) we constructed the income measure by collapsing the income variable from the WVS, instead of using estimates of regional GDP which are available only for a limited European sample.[14] In column 1, we report the correlation between regional income and the strength of family ties. Similar to the cross-country regressions, the correlation is negative and significant at the 1% level. This correlation also exists once we control for country fixed effects with a smaller but

[13] The measure of GDP is averaged between 1980 and 2010, the years in which the World Value Survey was taken. In particular, before taking the average, we match each country with the GDP corresponding to the year in which the survey was taken.

[14] The income variable in the dataset indicates income scales and is coded as a variable going from one to eleven, where one indicates the lower step in the scale of incomes and 11 the highest.

Table 4.9 Family ties and regional income

	(1) Whole sample	(2) Whole sample	(3) Europe	(4) Africa	(5) Asia	(6) North America and Oceania	(7) South America
Family ties	−0.540***	−0.349***	−0.287**	−1.383***	−0.498**	−0.327	0.133
	(0.078)	(0.111)	(0.127)	(0.398)	(0.201)	(0.408)	(0.444)
Country fixed effect	no	yes	yes	yes	yes	yes	yes
Observations	1,197	1,197	661	103	255	83	86
R-squared	0.047	0.526	0.466	0.691	0.482	0.731	0.354

Unit of analysis is a region in the World Value Survey. Coefficients are reported with robust standard errors in brackets.
*Indicates significance at the 10% level.
**Indicates significance at the 5% level.
***Indicates significance at the 1% level.

still relevant magnitude: a one standard deviation increase in the strength of family ties (0.44) is associated with a reduction in income of 0.152 (roughly equal to 14% of its standard deviation). It is also interesting to note that the correlation exists in all different continents. Columns 3–7 indeed show that the correlation is quite strong not only inside Europe but also inside Africa and Asia.[15]

The next question is whether the negative relationship between GDP and family values is also reflected in a negative relationship between family values and institutions. We explore this question in Table 4.10. We find that the strength of family ties is associated with lower quality of institutions. The effect is always negative and significant for all different types of institutions. The effect is also sizeable: a one standard deviation increase in the strength of family ties (0.35) is for example associated with a reduction of the control of corruption measure of 0.61 (roughly equivalent to 54% of its standard deviation).

A recent literature has suggested that one important driver of many formal institutions is legal origin. For example, English (common) law countries have been shown to have higher levels of investor protection, superior protection of property rights, and a more efficient judicial system. When we control for legal origin (Table 4.11), the negative association between family ties and the quality of institutions stay virtually the same.

4.7.2 Inherited Family Values and Current Institutions and Development

Our implicit assumption in all the empirical analysis is that family values change slowly. They are transmitted from generation to generation and they have persisted through history to the present day. This form of persistence seems intuitively likely given the probability that children are brought up to consider the attachment to the family, the

[15] The results on North America and Oceania are not significant, most likely due to the small sample size. Similarly for South America.

Table 4.10 Family ties and institutions

	(1) Control of corruption	(2) Government effectiveness	(3) Political stability	(4) Rule of law	(5) Regulatory quality	(6) Voice and accountability
Family ties	−1.729***	−1.575***	−1.576***	−1.595***	−1.199***	−1.428***
	(0.308)	(0.266)	(0.212)	(0.281)	(0.239)	(0.239)
Observations	80	80	80	80	80	80
R-squared	0.288	0.292	0.374	0.291	0.230	0.288

Coefficients are reported with robust standard errors in brackets.
*Indicates significance at the 10% level.
**Indicates significance at the 5% level.
***Indicates significance at the 1% level.

Table 4.11 Family ties and institutions, controlling for legal origin

	(1) Control of corruption	(2) Government effectiveness	(3) Political stability	(4) Rule of law	(5) Regulatory quality	(6) Voice and accountability
Family ties	−1.572***	−1.504***	−1.368***	−1.490***	−1.205***	−1.334***
	(0.395)	(0.357)	(0.278)	(0.370)	(0.309)	(0.286)
Legal origin dummies	yes	yes	yes	yes	yes	yes
Observations	80	80	08	80	80	80
R-squared	0.401	0.375	0.394	0.379	0.265	0.308

Coefficients are reported with robust standard errors in brackets.
*Indicates significance at the 10% level.
**Indicates significance at the 5% level.
***Indicates significance at the 1% level.

respect for parents, and the belief that they will do everything for their children as the natural state of the world. As a result, children will most likely reproduce the same values and beliefs with their own children. The persistence may develop and it can be facilitated through intermediate factors, such as the nature of political or economic institutions, shaped first by family structures which, in turn, have continued to influence our society today in a path-dependent manner.

In this section, we isolate the impact of cultural values on today's institutions. Ideally we would like to have measures of family values observed much before the measure of current institutions. Family values going so far back in time cannot be observed directly, since there is no survey available for that period of time. However, following Algan and Cahuc (2010) we can detect family ties by looking at family values inherited by children of immigrants in several European countries whose forbears arrived in Europe before 1940.

The idea behind this exercise is that parental values are a good predictor of the values of children. For this reason, we can use the family values that European descendants have inherited from their forebears who migrated to Europe from different countries before 1940 to know the values for the period preceding the quality of institutions today. This method allows us to cope with the lack of information on historical family values, by using the values that the descendants of various immigrants groups have inherited from their ancestors' countries of origin. This strategy is very useful because by using the values that European immigrants have inherited from the home country instead of the average values of the residents today, we can exclude reverse causality.

To perform our exercise we use data from the European Social Survey (ESS). The ESS is a biennial cross-sectional survey administered in a large sample of mostly European nations. The survey was conducted in 2002/2003, 2004/2005, 2006/2007, 2008/2009, and 2010/2011. The number of countries surveyed varies by wave. There are 22 countries included in the first round, 26 in the second, 25 in the third, 29 in the fourth, and 20 in the fifth. The sample size for a survey differs by country depending on its size. They range from 579 for Iceland to 2,870 for Germany.

Our primary sample consists of children of immigrants. We define children of immigrants as individuals born in a certain country but whose fathers were born abroad.[16] In order to get enough observations, we use information on second generation immigrants born before 1940. In the presence of cultural transmission children of immigrants should have inherited attitudes toward the families from their parents (who should have arrived in the destination countries not later than 1940 but possibly much earlier), who came to the destination countries with cultural attitudes from their countries of origin. Let's consider for example the case of France. To calculate the historical family values, we do consider children of French immigrants in a certain destination country. We do restrict the sample to children of immigrants born before 1940 and calculate their family values. These values are a reflection of their parental values who arrived from France before 1940, therefore the values of children of immigrants are a reflection of French family values before 1940.

The European Social Survey does not contain the same variables on family ties as those of the World Values Survey. To measure the strength of family ties we use a question asking the respondent his/her level of agreement with the following statement: "A person's family should be the main priority in life" the answer can go from "disagree strongly" to "agree strongly" on a scale from one to five.

There is a strong correlation between the inherited family ties of the children of immigrants born before 1940 (as measured by the ESS question) and current family ties in the countries of origin of their parents (as proxied by our measure of family ties calculated from the WVS). The correlation is about 0.35, showing that there is a strong inertia in family values across countries.

[16] When this information is not available we use the country of origin of the mother, if she is an immigrant. Natives are excluded from the analysis.

Table 4.12 Inherited family values and institutions

	(1) Control of corruption	(2) Government effectiveness	(3) Political stability	(4) Rule of law	(5) Regulatory quality	(6) Voice and accountability
Inherited family values	−0.664***	−0.622***	−0.558***	−0.630***	−0.477**	−0.613***
	(0.197)	(0.221)	(0.184)	(0.213)	(0.201)	(0.201)
Observations	128	129	129	129	128	129
R-squared	0.090	0.081	0.068	0.083	0.053	0.082

Coefficients are reported with robust standard errors in brackets.
*Indicates significance at the 10% level.
**Indicates significance at the 5% level.
***Indicates significance at the 1% level.

Table 4.13 Inherited family values and institutions, controlling for legal origin

	(1) Control of corruption	(2) Government effectiveness	(3) Political stability	(4) Rule of law	(5) Regulatory quality	(6) Voice and accountability
Inherited family values	−0.529***	−0.509***	−0.529***	−0.525***	−0.382**	−0.499***
	(0.148)	(0.174)	(0.157)	(0.163)	(0.160)	(0.164)
Legal origin	yes	yes	yes	yes	yes	yes
Observations	122	122	122	122	122	122
R-squared	0.340	0.309	0.260	0.320	0.235	0.263

Coefficients are reported with robust standard errors in brackets.
*Indicates significance at the 10% level.
**Indicates significance at the 5% level.
***Indicates significance at the 1% level.

We next discuss the correlation between the inherited family ties dating back to at least 60 years ago and current regulations in the home countries. Tables 4.12 and 4.13 show the OLS estimations, with and without the inclusion of legal origin dummies. We do find a robust and significant negative relationship between inherited family values and current institutions. The relationship holds even after controlling for legal origin. We do the same exercise with the level of development finding again a stable negative relationship between current development and inherited family values (columns 3 and 4 of Table 4.8). Overall, we do find that there is a long-lasting effect of family ties on the quality of current institutions.

4.8. FAMILY TIES AND WELL-BEING

Strong family ties countries are characterized by less favorable economic outcomes and attitudes. Unemployment rate, labor force participation, and income per capita are worse in strong family ties countries. Such unfavorable outcomes however do not seem to lead to dramatic situations of economic need in the population or to social unrest. This observation seems to suggest that in some sense those negative economic outcomes are less painful in strong family ties societies. In this part, we review existing evidence on the positive effects of familistic societies and provide some additional evidence on the conjecture that family ties could indeed improve well-being.

Bentolila and Ichino (2008) study the relationship between unemployment and consumption in four different countries: Spain, Italy, Great Britain, and the US. Their empirical results indicate that an increase in the duration of unemployment spells of male household heads is associated with smaller consumption losses in Spanish and Italian households. They conclude that extended family networks constitute the social institution which plays the crucial role of reducing the cost of unemployment near the Mediterranean. In Spain and Italy, the family appears to supplement for the lack of generosity of the welfare system and for the imperfection of capital markets. In this sense, the Mediterranean family-based solution seems to produce a desirable outcome from a welfare point of view since it allows for more consumption smoothing.

Along similar lines, Alesina and Giuliano (2010) look at the amount of home production in strong family ties societies. Societies with strong family ties are associated with more time spent at home by wives/mothers, and young adults living at home longer. This implies more home production (in the form of childcare, home cooking, caring for the elderly, house cleaning, etc.). In addition, according to a more traditional role attributed to women in societies in strong family ties societies, these activities should be mostly performed by wives and daughters. The authors indeed find that the strength of family ties is relevant for the determination of home production of women, but not of men, as expected.

Alesina and Ichino (2009) present some detailed calculations of the value of home production in four different countries: Spain and Italy with relatively strong family ties, the United States with an intermediate level, and Norway with a low level of family ties. They use two procedures in order to estimate the value of home production: the opportunity cost and the market value. They first calculate how much market income is lost by various individuals by working a certain number of hours at home rather than in the market, based upon characteristics such as their age, level of education, and wage value in the market. The second method is based upon how much it would cost to hire from the market individuals to perform household duties like cooking, cleaning, etc.[17] The authors find that home production is very large: it increases measured market income by a minimum of 53% to a maximum of 121% depending on the country and method of calculations. But more interestingly for our purposes the authors uncover a very large difference between strong and weak family ties countries. For instance, using the opportunity cost method, Italian families exactly double their market income by working

[17] The authors discuss in the detail the properties and the pros and cons of the two methods.

Table 4.14 Family ties and happiness

	(1) Happiness	(2) Satisfaction with life	(3) State of health
Family ties	0.057***	0.143***	0.025***
	(0.001)	(0.005)	(0.002)
Age	−0.006***	−0.027***	−0.011***
	(0.000)	(0.002)	(0.001)
Age-squared	0.000***	0.000***	−0.000***
	(0.000)	(0.000)	(0.000)
Female	0.014***	0.033***	−0.114***
	(0.003)	(0.010)	(0.004)
Married	−0.013	−0.128***	−0.036***
	(0.008)	(0.026)	(0.010)
Education dummies	yes	yes	yes
Country dummies	yes	yes	yes
Wave dummies	yes	yes	yes
Observations	222,197	221,458	187,053
R-squared	0.141	0.179	0.221

Coefficients are reported with robust standard errors in brackets.
*Indicates significance at the 10% level.
**Indicates significance at the 5% level.
***Indicates significance at the 1% level.

at home contrary to an increase of about 74% in the US. Using the market cost of services, Italians more than double their market income (+121%) while Norwegians increase it by 80%. These results suggest that a market income measure tends to underestimate the well-being of strong family ties countries, given that home production is not included in this measure.

These considerations open the question of the relationship between the strength of family ties and alternative measures of well-being to which we now turn. Table 4.14 illustrates this relationship using measures of subjective happiness and self-reported health. The first question asks the respondent, on a scale from 1 to 4, whether "Taking all things together, would you say you are:" very happy (taking the value of 4), quite happy, not very happy, not at all happy (taking the value of 1)? The second question asks "All things considered, how satisfied are you with your life as a whole these days?" The answer goes from dissatisfied (taking the value of 1) to satisfied (taking the value of 10). The last question asks the respondent, "All in all, how would you describe your state of health these days?" Would you say it is very good (taking the value of 5), good, fair, poor, or very poor (taking the value of 1). The results in Table 4.14 clearly show that, all in all, although strong family ties can harm societies in a variety of ways, they can also have positive effects in an individual's life, as measured by happiness and self-reported measures of health. The magnitude of the effect is also sizeable: the beta coefficients of family ties on the three measures of well-being are equal to 0.08, 0.06, and 0.03 respectively (for a comparison, the impact of the highest level of education is equal to 0.09, 0.04, and 0.08).

How can one interpret these results? One interpretation could be that well-being depends essentially on the quality of social relationships and not only on individual income. From this perspective, if social relationships are particularly good among family members, we should expect a strong correlation between family ties and well-being. Second, these results on well-being may capture the effect of stress and harder work (reflected in higher per capita income) in environments with weaker family ties. Alesina and Ichino (2009) make this argument with explicit reference to Italy. In a sense, the strong family ties of this county may explain both its relative decline in a globalized world but also the relatively high life satisfaction (at least for now) of Italians.

4.9. CONCLUSION

We show that differences in family values have an impact on attitudes and outcomes that are relevant to explain differences in growth across countries and the quality of institutions. We study attitudes toward working women, the society, generalized morality, and civic engagement. Our findings confirm an idea first developed by political scientists and sociologists: trust in the family prevents the formation of generalized trust, which is at the core of many collective good outcomes, from political participation to the formation of institutions to economic development. This should not be taken of course as a "criticism" of the family as a fundamental institution of society but as an analysis of the effect of different family arrangements. Our analysis indeed shows that family ties are related to different measures of happiness, life satisfaction, and self-reported health.

ACKNOWLEDGMENT

We thank Andrea Passalacqua for excellent research assistantship.

REFERENCES

Acemoglu, D., Johnson, S., Robinson, J., 2001. The colonial origins of comparative development: an empirical investigation. American Economic Review 91, 1369–1401.

Acemoglu, D., Johnson, S., Robinson, J.A., 2005. Institutions as the fundamental cause of long-run economic growth. In: Aghion, Philippe, Durlauf, Stephen (Eds.), Handbook of Economic Growth. Elsevier BV, Amsterdam, pp. 385–472.

Akerlof, G., Kranton R., 2011. Identity Economics: How Our Identities Affect Our Work, Wages and Well-Being. Princeton University Press, Princeton, New Jersey.

Akerlof, G., Kranton, R., 2000. Economics and identity. Quarterly Journal of Economics 115, 715–753.

Alesina, A., La Ferrara, E., 2005. Ethnic diversity and economic performance. Journal of Economic Literature 43 (3), 762–800.

Alesina, A., Giuliano, P., 2010. The power of the family. Journal of Economic Growth 15, 93–125.

Alesina, A., Giuliano, P., 2011. Family ties and political participation. Journal of the European Economic Association 9 (5), 817–839.

Alesina, A., Ichino, A., 2009. L'Italia fatta in casa. Mondadori, Milano.

Alesina, A., Algan, Y., Cahuc, P., Giuliano, P., 2013. Family Values and the Regulation of Labor. Mimeo.

Alesina, A., Giuliano, P., Nunn, N., 2013b. On the origin of gender roles: Women and the plough, Quarterly Journal of Economics 128 (2), 469–530.

Algan, Y., Cahuc, P., 2007. The Roots of Low European Employment: Family Culture? NBER Macroeconomic Annual. MIT Press, Cambridge, MA.

Algan, Y., Cahuc, P., 2010. Inherited trust and growth. American Economic Review 100, 2060–2092.

Almond, G., Verba, S., 1963. The Civic Culture: Political Attitudes and Democracy in Five Nations. Princeton University Press.

Arrow, K., 1972. Gifts and exchanges. Philosophy and Public Affairs 1 (4), 343–362.

Banfield, E., 1958. The Moral Basis of a Backward Society. Free Press, New York.

Bentolila, S., Ichino, A., 2008. Unemployment and consumption near and far away from the Mediterranean. Journal of Population Economics 21, 255–280.

Bertrand, M., Schoar, A., 2006. The role of family in family firms. Journal of Economic Perspectives 20 (2), 73–96.

Boyd, R., Richerson, P.J., 1985. Culture and the Evolutionary Process. University of Chicago Press, London.

Castles, F., 1995. Welfare state development in Southern Europe. Western European Politics 18, 201–213.

Coleman, J.S., 1988. Social capital in the creation of human capital. American Journal of Sociology XCIV, S95–S120.

Coleman, J.S., 1990. Foundations of Social Theory. Harvard University Press, Cambridge, MA.

Durante, R., 2010. Risk, Cooperation and the Economic Origins of Social Trust: an Empirical Investigation, Science Po. Mimeo.

Duranton, G., Rodriguez-Pose, A., Sandall, R., 2009. Family types and the persistence of regional disparities in Europe. Economic Geography 85, 23–47.

Easterly, W., Levine, R., 1997. Africa's growth tragedy: Policies and ethnic divisions. Quarterly Journal of Economics 112 (4), 1203–1250.

Ermish, J., Gambetta, D., 2010. Do strong family ties inhibit trust? Journal of Economic Behavior and Organization 75, 365–376.

Esping-Andersen, G., 1999. Social Foundation of Post-Industrial Economies. Oxford University Press, Oxford.

Fernandez, R., Fogli, A., 2009. Culture: an empirical tnvestigation of beliefs, work and fertility. American Economic Journal: Macroeconomics 1 (1), 146–177.

Ferrera, M., 1996. The Southern model of welfare in social Europe. Journal of the European Social Policy 1, 17–37.

Fortin, N., 2005. Gender role attitudes and the labour market outcomes of women across OECD countries. Oxford Review of Economic Policy 21, 416–438.

Fukuyama, F., 1995. Trust: The Social Virtues and the Creation of Prosperity. Free Press, New York

Galasso, V., Profeta, P., 2012. When the State Mirrors the Family: the Design of Pension Systems. Bocconi University, Mimeo.

Gambetta, D. (Ed.), 1988. Trust: Making and Breaking Cooperative Relations. Blackwell.

Giavazzi, F., Schiantarelli, F., Serafinelli, M., forthcoming. Attitudes, policies and work. Journal of the European Economic Association.

Gigerenzer, G., 2007. Gut Feelings: the Intelligence of the Unconscious. Penguin Group, New York.

Giuliano, P., 2007. Living arrangements in Western Europe: does cultural origin matter? Journal of the European Economic Association 5 (5), 927–952.

Glaeser, E., La Porta, R., Lopez de Silanes, F., Shleifer, A., 2004. Do institutions cause growth. Journal of Economic Growth 9, 271–304.

Glaeser, E., Ponzetto, G., Shleifer, A., 2007. Why democracy need education? Journal of Economic Growth 12 (2), 77–99.

Gorodnichenko, Y., Roland, G., 2013. Culture, Institutions and the Wealth of Nations. UC Berkeley, Mimeo.

Greif, A., 1994. Cultural beliefs and the organization of society: a historical and theoretical reflection on collectivist and individualist societies. Journal of Political Economy 5 (102).

Greif, A., 2006a. Family structure, institutions, and growth: The origins and implications of western corporations. American Economic Review, 96 (2), 308–312.

Greif, A., 2006b. Institutions and the Path to the Modern Economy: Lessons from Medieval Trade. Cambridge University Press, Cambridge.

Greif, A., Tabellini, G., 2012. The Clan and the City: Sustaining Cooperation in China and Europe. Mimeo, Stanford.

Guiso, L., Sapienza, P., Zingales, L., 2006. Does culture affect economic outcomes? Journal of Economic Perspectives (Spring).

Guiso, L., Sapienza, P., Zingales, L., 2008. Long Term Persistence, NBER WP 14278.

Guiso, L., Sapienza, P., Zingales, L., 2010. Social capital as good culture. Journal of the European Economic Association 6 (2–3), 295–320.

Kanhneman, D., 2011. Thinking, Fast and Slow. Farrar, Straus and Giroux, New York.

Korpi, W., 2000. Faces of Inequality: Gender, Class and Patterns of Inequalities in Different Types of Welfare States. Social Politics 7, 127–191.

Laslett, P., 1983. Family and household as work group and kin group: Areas of traditional Europe compared. In: Wall R., Robin J. (Eds.), Family Forms in Historic Europe. Cambridge University Press, Cambridge, pp. 513–563.

Lipset, S.M., 1959. Some social requisites of democracy: economic development and political legitimacy. American Political Science Review 53, 69–105.

Murdock, P.M., 1949. Social Structure. Free Press, New York.

North, D., 1981. Structure and Change in Economic History. Norton, New York.

North, D.C., 1990. Institutions, Institutional Change and Economic Performance. Cambridge University Press, New York.

Persson, T., Tabellini, G., 2009. Democratic capital: the nexus of political and economic change. American Economic Journal: Macroeconomics 1 (2), 88–126.

Platteau, J.P., 2000. Institutions, Social Norms, and Economic Development. Academic Publishers and Routledge.

Putnam, R., Leonardi, R., Nanetti, R.Y., 1993. Making Democracy Work: Civic Traditions in Modern Italy. Princeton University Press.

Putnam, R., 2000. Bowling Alone: the Collapse and Revival of American Community. Simon and Schuster, New York, NY.

Reher D.S., 1998. Family ties in Western Europe: Persistent contrasts. Population and Development Review 24, 203–234.

Spolaore, E., Wacziarg, R., 2009. The diffusion of development. Quarterly Journal of Economics 124 (2), 469–529.

Tabellini, G., 2008. The scope of cooperation: values and incentives. Quarterly Journal of Economics 123 (3), 905–950.

Tabellini, G., 2010. Culture and institutions: economic development in the regions of Europe. Journal of the European Economic Association 8 (4), 677–716.

Todd, E., 1983. The Explanation of Ideology: Family Structures and Social Systems. Basic Blackwell, New York.

Todd, E., 1990. L'invention de l'Europe. Seuil, Paris.

Weber, Max, 1904. The Protestant Ethic and the Spirit of Capitalism. Scribner's Press, New York.

CHAPTER FIVE

The Industrial Revolution

Gregory Clark
University of California, Davis, CA 95616, USA

Abstract

The Industrial Revolution decisively changed economywide productivity growth rates. For successful economies, measured efficiency growth rates increased from close to zero to close to 1% per year in the blink of an eye, in terms of the long history of humanity, seemingly within 50 years of 1800 in England. Yet the Industrial Revolution has defied simple economic explanations or modeling. This paper seeks to set out the empirical parameters of the Industrial Revolution that any economic theory must encompass, and illustrate why this makes explaining the Industrial Revolution so difficult within the context of standard economic models and narratives.

Keywords

Industrial revolution, Economic growth, Growth theory

JEL Classification Codes

N13, O33, O43, O47

5.1. INTRODUCTION

The economic history of the world is surprisingly simple. It can be presented in one diagram, as in Figure 5.1 below. Before 1800, income per capita for all the societies we observe fluctuated. There were good and bad periods. But there was no upward trend. The great span of human history—from the arrival of anatomically modern man to Confucius, Plato, Aristotle, Michelangelo, Shakespeare, Beethoven, and all the way to Jane Austen—was lived in societies caught in the Malthusian trap. Jane Austen may write about refined conversation over tea served in China cups, but for the mass of people, as late as 1813, material conditions were no better than their ancestors of the African savannah. The Darcys were few, the poor plentiful.[1]

Around 1780 came the Industrial Revolution in England. Incomes per capita began a sustained growth in a favored group of countries around 1820. In the last 200 years, in the most fortunate countries, real incomes per capita rose 10–15-fold. The modern world was born. The Industrial Revolution thus represents the single great event of world economic history, the change between two fundamentally different economic systems.

[1] Clark (2007) extensively reviews the evidence for this assertion.

Handbook of Economic Growth, Volume 2A
ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53538-2.00005-8

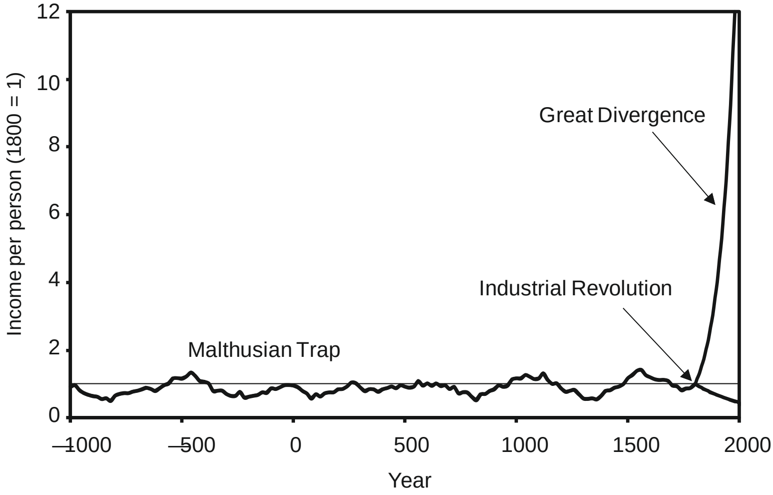


Figure 5.1 A schematic history of world economic growth. *Source: Clark (2007), Figure 1.1, p. 2.*

The puzzle is why it occurred only around 1780, and why it occurred in a modest island nation on the northwest shores of the European continent.

At one level the transformation the Industrial Revolution represents is very simple. Beginning with the Industrial Revolution, successful modern economies experience steady rates of efficiency advance. Every year more output is produced per unit of input. At a proximate level, the growth of income per work-hour in modern societies can be represented as:

$$g_y = ag_k + g_A, \tag{5.1}$$

where g_k is the rate of growth of capital per worker hour, a is the share of capital payments in national income, and g_A is the growth rate of efficiency. Since the Industrial Revolution, the capital stock has grown about as rapidly as output. Also, the share of capital in all earnings is about a quarter. Thus, only about a quarter of all modern growth in income per person comes directly from physical capital. The rest is an unattributed rise in the measured efficiency of the economy, year by year.

But while Equation (5.1) suggests that efficiency growth and physical capital accumulation are independent sources of growth, in practice, in market economies there has been a strong correlation between the two sources of growth. Economies with significant efficiency growth are also those with substantial growth rates of physical capital. Something links these two sources of growth.

Some economists, most notably Paul Romer, have theorized that this correlation stems from external benefits associated with physical capital accumulation (Romer, 1986, 1987, 1990). For this explanation to work, there would have to be $3 of external benefit accruing to physical capital investments for every $1 of private benefit. Most of the modern physical capital stock, however, is still mundane things such as houses, buildings, roads, water and sewer systems, and bridges. These types of investment do not seem to be associated with substantial external benefits. So if productivity advance is systematically associated also with the growth of the stock of such physical capital there must be another mechanism.

The most plausible one is that the association of physical capital accumulation with efficiency advance stems just from the effects of efficiency advance on increasing the marginal product of capital. In a world with relatively constant real interest rates since the Industrial Revolution, such a rising marginal product will induce more investment. And indeed if the economy is roughly Cobb-Douglas in its production structure, efficiency advances will induce a growth of the physical capital stock per person at a rate equal to the growth of output per person, so that the capital-output ratio is constant. This is roughly what we observe.

Thus, at a deeper level all modern growth seemingly stems from this unexplained rise in economic efficiency, as a product of a rise in knowledge about production processes. Somehow after 1780 investment in such knowledge increased, or enquiry became much more effective in creating innovation.

Before the Industrial Revolution we find no sign of any equivalent efficiency advances. This is true globally all the way from 10,000 BC to 1800, where we can measure the implied rate of productivity advance just from the rate of growth of population. In this long interval, average estimated rates of efficiency advance are 0.01% per year or less. We know this because we can assume before the Industrial Revolution, because of the Malthusian Trap, that output per person and capital per person was, in the long run, constant. In that case, any gains in efficiency will be absorbed by population growth according to the formula:

$$g_A = cg_N. \tag{5.2}$$[2]

We can thus approximate efficiency growth rates from population growth rates if we look at sufficiently long intervals. Table 5.1 shows these calculations at a world level. Implied rates of technological advance are always extremely slow.

But it is also true that implied rates of technological advance are also slow for those economies where we can measure actual efficiency levels before 1800 through measurements of the real payments to factors. Figure 5.2 shows the implied efficiency in England 1250–2000. As can be seen, there is, surprisingly, in England no sign of any significant improvement in the efficiency of the economy all the way from 1250 to 1800. Only around 1800 does the modern age of steady efficiency advance appear. Before that, the measured efficiency of the economy fluctuated, peaking around 1450, but with almost no upward trend.

[2] For a more detailed explanation see Clark (2007, 379–82).

Table 5.1 Population and technological advance at the world level, 130,000 BC to 1800

Year	Population (millions)	Population growth rate (%)	Technology growth rate (%)
130,000 BC	0.1	–	–
10,000 BC	7	0.004	0.001
1 AD	300	0.038	0.009
1000 AD	310	0.003	0.001
1250 AD	400	0.102	0.025
1500 AD	490	0.081	0.020
1750 AD	770	0.181	0.045

Source: Clark (2007), Table 7.1.

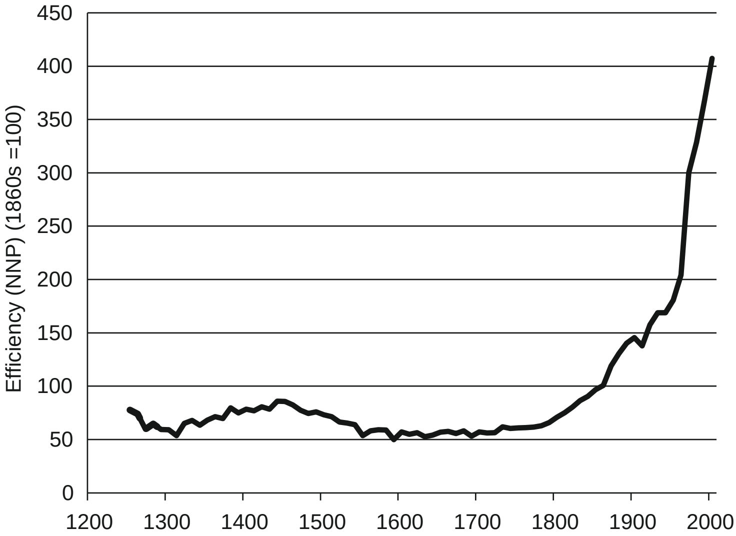


Figure 5.2 Estimated efficiency of the English economy, 1250–2000. *Source:* *Clark (2010)*.

The Industrial Revolution thus seems to represent a singularity. A unique break in world history. But also an event where we know clearly what we have to explain. Why did the rate of expansion of knowledge about production efficiency increase so dramatically in England around 1800? Figure 5.3 shows that the upturn in productivity growth rates can be located to the 1780s/1790s. That upturn is preceded by seven decades in which the average annual productivity growth rate was a mere 0.14% per year. Fast by the standards of the pre-industrial world, but glacially slow in modern terms. Overall productivity growth rates 1780–1789 to 1860–1869, averaged 0.58% per year, about half way to fully modern levels.

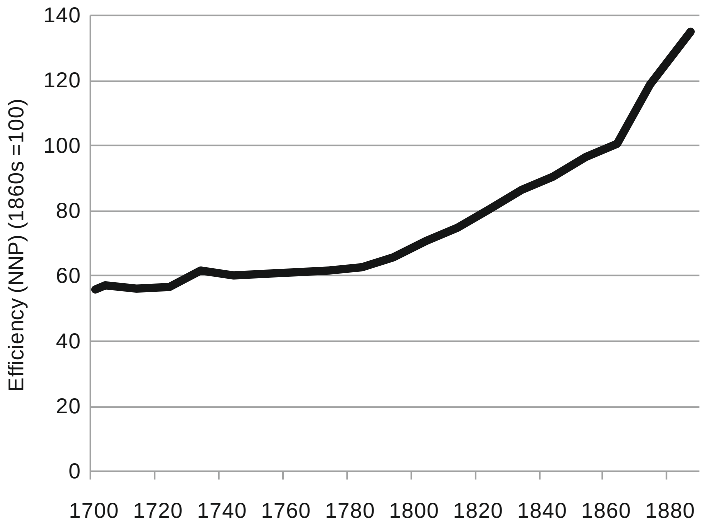


Figure 5.3 Efficiency levels, England, 1700–1880. *Source:* *Clark (2010)*.

We also know what sectors contributed most of the productivity advance 1780–1789 to 1860–1869. National productivity growth will be related to productivity advance in individual sectors through the equation:

$$g_A = \sum \theta_j g_{Aj}, \tag{5.3}$$

where g_{Aj} is the growth rate of productivity by sector, and θ_j is the share of j in total value added in the economy. These results are shown in Table 5.2.

Textiles contributed nearly half, 43%, of all measured productivity advance. Improvements in transport, mainly the introduction of the railway, were the next biggest source of advance, contributing 20%. Agriculture, ironically, contributed almost 20% also. Coal and

Table 5.2 Sources of industrial revolution efficiency advance, 1780s–1860s

Sector	Efficiency growth rate (%)	Share of value added	Contribution to national efficiency growth rate (% per year)
All textiles	2.3	0.11	0.25
Iron and steel	1.8	0.01	0.02
Coal mining	0.2	0.02	0.00
Transport	1.5	0.08	0.12
Agriculture	0.4	0.30	0.11
Identified advance	–	0.51	0.49
Whole economy	–	1.00	0.58

Source: Clark (2007), Table 12.1.

iron and steel were in themselves minor contributions despite the fame of these sectors and their innovations in this period. Productivity growth in the half of the economy not covered in Table 5.2 was modest, less than 0.20% per year.

The decomposition in Table 5.2 established some things already. The Industrial Revolution has been thought of by some as essentially consisting of the arrival of the first of what have been called general purpose technologies, the steam engine. General purpose technologies, a rather nebulous concept, have been variously defined. They can be loosely thought of as innovations that have pervasive application throughout the economy, that go through a prolonged period of improvement, and that spawn further innovation in the sectors they are employed in.[3] Various GPTs have been identified, such as the introduction of steam power during the Industrial Revolution, and the introduction of electricity, and the recent IT revolution.

Steam power in England certainly touched a number of areas in the Industrial Revolution. It was important in coal mining, on the railroads, and in powering the new textile factories. The steam engine itself underwent a long process of improvement in thermal efficiency, and in the ratio of power to weight, from its first introduction by Thomas Newcomen in 1707–1712, to the 1880s. The earliest engines had a thermal efficiency as low as 0.5%, while those of the 1880s could achieve thermal efficiencies of 25%. The steam engine was associated also with the widespread use of fossil energy in the economy to replace wind, water, and animal power sources in transport, home heating, and manufacturing.

Table 5.2 suggests, however, that whatever role steam power played in economy-wide productivity advance after the 1860s, its role up to then in the new productivity advance of the Industrial Revolution was minor. Coal mining and iron and steel production contributed very little to Industrial Revolution productivity advance, and most of their productivity advance did not stem from the introduction of steam power.[4] Even in transport, a substantial part of the productivity advance is attributable to the improvement of the traditional road transport system, the introduction of canals, and improvements in sailing ships. The textile factories of the Industrial Revolution could, if necessary, have still been powered by water wheels even as late as the 1860s. Advances in textiles and agriculture explain the majority of the Industrial Revolution.

The diverse nature of productivity advance in this era makes the Industrial Revolution all the more puzzling. The revolution in textiles came through mechanical innovations that can be traced to a number of heroic individual innovators: John Kay, Richard Arkwright, James Hargreaves, Edmund Cartwright. But the improvements in agriculture stem from the advances of thousands of anonymous farmers in improving yields, mainly involving non-mechanical changes.

[3] Bresnahan and Trajtenberg (1996).

[4] Clark and Jacks (2007).

Another important element in the Industrial Revolution era is the unimportance of traditional investments in physical capital in explaining the growth of output per worker. Capital per worker rose no faster than output per worker, so that right from the onset of modern growth efficiency growth dominated.

Thus, any satisfying account of the Industrial Revolution has to do the following things. First explain why NO society before 1800—not ancient Babylon, Pharaonic Egypt, China through countless centuries, Classical Greece, Imperial Rome, Renaissance Tuscany, medieval Flanders, the Aztecs, Mogul India, the Dutch Republic—expanded the stock of knowledge by more than 10% a century. Then explain why, within 50 years of 1800, the rate of growth of knowledge rose to modern rates in one small country on the margins of Europe, Britain. Then we will understand the history of man.

5.2. THEORIES OF THE INDUSTRIAL REVOLUTION

The drama and the centrality of the Industrial Revolution has ensured that there is a steady supply of new or recycled theories of this great transition. These theories mostly fall into a number of discrete categories.

Bad equilibrium theories seek to explain the Malthusian stagnation as a product of a self-reinforcing system of poor economic incentives. The desires and rationalities of people in all human societies are essentially the same. The medieval peasant in Europe, the Indian coolly, the bushman of the veld, share a common set of aspirations, and a common ability to act to achieve those aspirations. What differs across societies, however, are the institutions that govern economic life. Thus

> *In fact, most societies throughout history got "stuck" in an institutional matrix that did not evolve into the impersonal exchange essential to capturing the productivity gains that came from the specialization and division of labor that have produced the Wealth of Nations* (North, 1994, 364).

Thus, there is a caricature of the pre-industrial world that many economists intuitively hold, which is composed of a mixture of all the bad movies ever made about early societies. Vikings pour out of long ships to loot and pillage defenseless peasants and burn the libraries of monasteries. Mongol hordes flow out of the steppe on horseback to sack Chinese cities. Clerical fanatics burn at the stake those who dare to question arcane religious doctrines. Peasants groan under the heel of rapacious lords whose only activity is feasting and fighting. Aztec priests cut out the hearts from screaming, writhing victims with obsidian knives. In this brutal and chaotic world, who has the time, the energy, or the incentive, to develop new technology?

The advantage of a theory which relies on some exogenous shock to the economic system is that it can hopefully account for the seemingly sudden change in the growth rate of measured efficiency, around 1800. Institutions can change suddenly and dramatically—witness the French Revolution, the Russian Revolution, and the Iranian Revolution that overthrew the Shah.

These theories of an institutional shift in appropriability face two major difficulties, however, one conceptual, one empirical. The conceptual difficulty is that if modern economic growth can be produced by a simple institutional change, then why in all the varied and various societies that the world has seen since 10,000 BC and before was there none which stumbled upon the right set of institutions that made knowledge property? Societies varied markedly in what could be property and how property was transferred between owners. For example, in civil cases over possession of land in the legal system established by the Normans in medieval England after 1066, the party whose right to land was contested could elect to prove his or her title through armed combat with his opponent! This may seem a crazy way of settling property disputes to us, but the point is that societies have made all kinds of different choices about institutional forms. Why did some not stumble upon the right set of institutions? It seems that we cannot rely on chance here in institutional choice. There must be something that is keeping the institutions of the pre-industrial world in the "bad" state.

Thus, a slightly more sophisticated version of the "bad institutions" theory are those which seek to explain through the political economy of institutions why systematically early societies had institutions that discouraged economic growth (see, for example, Greif, 2006; North and Thomas, 1973; North and Weingast, 1989; North, 1994; Jones, 2002; Acemoglu et al. 2001, 2002; and Acemoglu and Robinson, 2012).

The common feature that Douglass North and other such institutionalists point to in early societies is that political power was not achieved by popular elections. In pre-industrial societies, as a generalization, the rulers ultimately rested their political position on the threat of violence. Indeed there is a close empirical association between democracy and economic growth. By the time England achieved its Industrial Revolution it was a constitutional democracy where the king was merely a figurehead. The USA, the leading nation in the world in economic terms since the 1870s, has always been a democracy also.[5]

For economic efficiency in any society property rules have to be chosen to create the maximum value of economic output. In such a case a disjuncture can arise between the property rules in the society that will maximize the total value of output, and the property rules that will maximize the output going to the ruling elite. Indeed, North and others have to argue that such a disjuncture systematically arises in all societies before 1800. This idea has been restated recently as the idea that economic growth is the replacement of extractive economic institutions, designed just to secure income for a ruling clique, with inclusive economic institutions, designed to maximize the output of societies as a whole (Acemoglu and Robinson, 2012).

One subset of such theories that has shown amazing persistence, despite its inability to account for the most basic facts of the Industrial Revolution, is that which links the Industrial Revolution to the earlier Glorious Revolution of 1688–1689. Thus the recent

[5] The recent rise of China is, however, an exception to the general association of growth and democracy.

widely read book by Acemoglu and Robinson, *Why Nations Fail*, has a chapter titled *How a political revolution in 1688 changed institutions in England and led to the Industrial Revolution*.

The Glorious Revolution established the modern political system of the UK, a system that has been continuously modified, but not fundamentally changed since then. The new political system created Parliament, the representative of the propertied classes in England in 1689, as the effective source of power in what is nominally a monarchy.

A basic problem with placing political developments at the heart of the Industrial Revolution is that changes in political regime before 1800 have no discernible impact on the efficiency level of the economy, even 80 years later. The Glorious Revolution had no discernible impact on economic efficiency before 1770, two or three generations after the institutional change, as Figure 5.4 shows. It is also clear in the figure that even the earlier political and military disruptions of the Civil War of 1642–1649, and the Interregnum of 1649–1660, were not associated with any decline in the efficiency of operation of the economy.

Further, there is no sign that private investors in England perceived a greater security of property even as a result of the Glorious Revolution. The return to private capital in the economy did not deviate from trend after 1689. Private investors seem to have looked at the political changes with indifference (Clark, 1996). The return to government debt did eventually decline significantly after 1689, and had fallen to modern levels by the 1750s. This decline was no doubt driven in part by the enhanced taxing power of the government after 1689. But almost all of the money raised from those taxes went to finance the British Navy in the long struggle with France that ended only with the

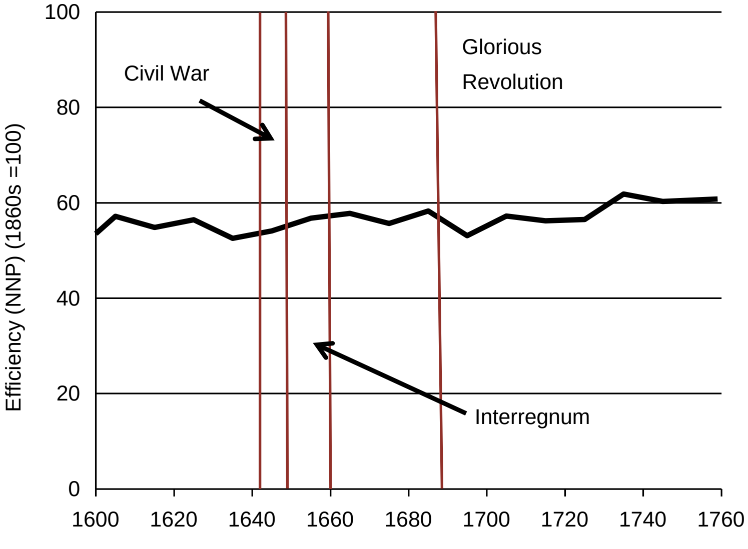


Figure 5.4 Economic efficiency and political changes, England, 1600–1770. *Source:* *Clark (2010)*.

defeat of Napoleon in 1815. Almost none of it went into the subsidization of innovation or education.

And we do see long before the Glorious Revolution or the Industrial Revolution societies that had stable representative political systems, the inclusive institutions of Acemoglu and Robinson, but little or no productivity advance. The Dutch Republic of 1588–1795 was one such regime.[6] Under the political arrangements of the Republic, the Netherlands experienced its Golden Age. Despite its modest size and lack of substantial domestic natural resources, it conquered a substantial colonial empire in the East, possessing for a while the premier navy in the world, dominating world trade in the 17th century. It developed sophisticated systems of banking and public finance, allowing substantial borrowing to develop a modernized transportation system internally, and support the most urbanized society in Europe. But because productivity advance stagnated in the Netherlands 1650–1795, these political and institutional achievements led to no sustained growth, and no break from the pre-industrial world.

From 1223 to 1797,Venice operated as a republic, with the government under control of a balance of popular and patrician representatives. Policy was geared toward the needs of a trading and commercial empire.Venice again developed an important trading empire in the eastern Mediterranean, with colonies and dependencies such as Dalmatia, Crete, and Cyprus. It also developed important manufacturing activities such as its glass industry. But again, none of this was reflected in the kind of sustained productivity advance seen in the Industrial Revolution.

Similarly, the free cities of the Hanseatic League were from the Middle Ages dominated by a politics that emphasized the needs of trade and commerce. Lübeck, for example, became a free city in 1226, and retained city state status until 1937. After gaining its freedom, Lübeck developed a system of rule and government called Lübeck Law that spread to many other Baltic cities of the Hanseatic League in the Middle Ages such as Hamburg, Kiel, Danzig, Rostock, and Memel. Under Lübeck Law, the city was governed by a council of 20 that appointed its own members from the merchant guilds and other town notables. It was thus government by the leaders of the commercial interests of the cities (Lindberg, 2009). Though not democracy, this was government by interests that should have fostered commerce and manufacturing. Under such rule the Hansa cities became rich and powerful, engaging in substantial manufacturing enterprises, such as shipbuilding and cloth production, as well as trade. But again, this was not associated with sustained technological advance.

It is true that the early societies that we know of in detail seem to have lacked the legal notion that you could own property in ideas or innovations. Thus, in both the Roman and Greek worlds when an author published a book there was no legal or practical way

[6] The Dutch Act of Abjuration of 1581 has been argued by some to be the precursor to the Declaration of Independence of the USA of 1776.

to stop the pirating of the text. Copies could be freely made by anyone who acquired a version of the manuscript (on papyrus rolls), and the copier could amend and alter the text at will. It was not uncommon for a text to be reissued under the name of a new "author".[7] It was common to condemn such pirating of works or ideas as immoral. But writings and inventions were just not viewed as *commodities* with a market value.[8]

While the ancients may have lacked them, there were systems of intellectual property rights in place, however, long before the Industrial Revolution. The earliest established foundations of a modern patent system were found in the 13th century in Venice. By the 15th century in Venice, true patents in the modern sense were regularly being awarded. Thus, in 1416, the Council of Venice gave a 50-year patent to a foreigner, Franciscus Petri from Rhodes, for a new type of fulling mill. By 1474, the Venetian patent law had been codified. There is also evidence of the awarding of patents in Florence in the 15th century. The Venetian innovation of granting property rights in knowledge, which was very important to the famous Venetian glass industry, spread to Belgium, the Netherlands, England, Germany, France, and Austria in the 16th century as a consequence of the movement of Italian glassworkers to these other countries. Therefore, by the 16th century all the major European countries, at least on an ad hoc basis, granted property rights in knowledge to innovators. They did this in order to attract skilled craftsmen with superior techniques to their lands. The spread of formal patent systems thus predates the Industrial Revolution by at least 350 years.

The claims of North and his associates for the superiority of the property rights protections afforded by the patent system in 18th century England thus stem from the way in which the system operated after the Glorious Revolution of 1688–1659 established the supremacy of Parliament over the King. Under the patent system introduced in the reign of Elizabeth I (1568–1603) the system was supervised by government ministers. Political interference led to the creation of spurious monopolies for techniques already developed, or the denial of legitimate claims. After the Glorious Revolution, Parliament sought to avoid this by devolving the supervision of patents to the courts. Generally the courts would allow any patent to be registered as long as no other party objected. No other major European country had a formal patent system as existed in England before 1791. But, as Figure 5.5 shows, while the Glorious Revolution produced a brief increase in patent rates, there was no sustained increase in patenting rates until the 1760s, 75 years after the Glorious Revolution.

There also existed other institutions in, for example, medieval European society, which we would think would promote innovation better than the modern patent system. Producers in many towns were organized into guilds which represented the interests of the

[7] This problem continued into at least the 17th century in England, where publishers quite freely pirated the works of authors.

[8] See Long (1991, pp. 853–7).

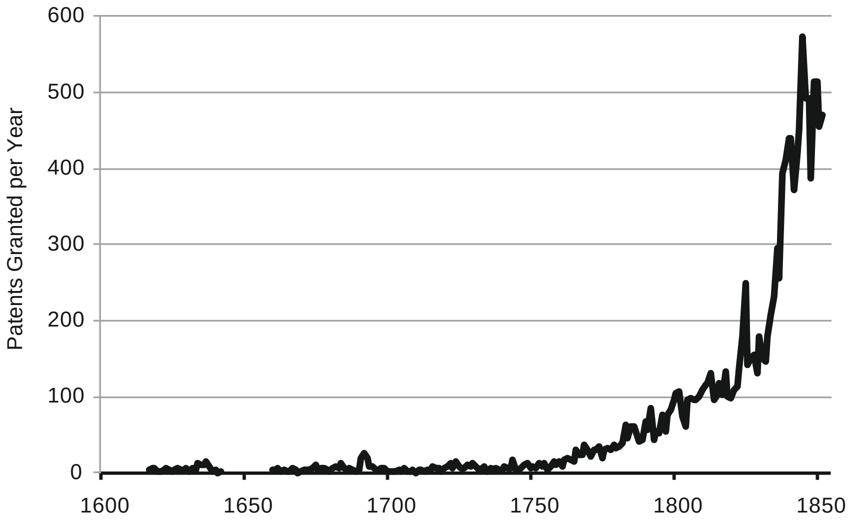


Figure 5.5 Patents per year, England, 1660–1851. *Source:* *Mitchell (1988), p. 438.*

trade. These guilds were in a position to tax members to facilitate lump-sum payments to innovators to reveal productive new techniques to the members.

The empirical difficulty with the appropriability argument is the appallingly weak evidence that there was any great gain in the returns to innovators in England in the 1760s and later. The textile industry for example was at the forefront of technological change during the Industrial Revolution. Figure 5.6 shows TFP in the production of cotton cloth, taking cotton as a basic input. From 1770 to 1869, TFP rose about 22-fold.

Yet the gains of the textile innovators were modest in the extreme. The value of the cotton textile innovations alone by the 1860s, for example, was about £115 million in extra output per year. But a trivially small share of this value of extra output ever flowed to the innovators. Table 5.3, for example, shows the major innovators in cotton textiles and the gains accruing to the innovators through the patent system or other means. Patents mostly provided poor protection, the major gains to innovators coming through appeals post hoc to public beneficence through Parliament. Also, the patent system shows none of the alleged separation from political interference. The reason for this is that Parliament could, on grounds of the public good, extend patents beyond the statutory 17 years to adequately reward those who made significant innovations. James Watt was the beneficiary of such a grant. But such grants depended on social and political protection just as much as in the old days.

The profit rates of major firms in the industry also provide good evidence that most of the innovation in the textile industry was quickly leaking from the innovators to other producers with no rewards to the innovators. Knick Harley has reconstructed the profit rates being made by some of the more successful cotton spinning and weaving firms in

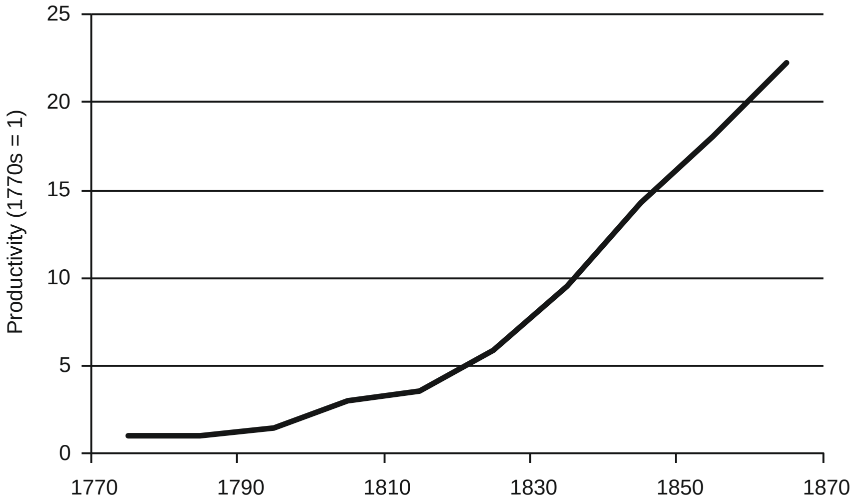


Figure 5.6 Cotton spinning and weaving productivity, 1770–1869. *Note*: The years 1862–1865 were omitted because of the disruption of the cotton famine. *Sources: Cotton cloth prices, Harley (1998). Labor costs, return on capital, Clark (2010).*

the early Industrial Revolution period (Harley, 2010). The cotton spinners Samuel Greg and Partners earned an average profit from 1796 to 1815 of 11.4% per year, just the normal commercial return for a risky venture such as manufacturing. Given the rapid improvements in cotton spinning productivity going on in the industry in these years it suggests that whatever innovations were being introduced were spreading from one firm to another very quickly. Otherwise leading firms such as Samuel Greg would have made large profits compared to their competitors. Similarly, the firm of William Grey and Partners made less than 2% per year from 1801 to 1810, a negative economic profit rate. The innovations in the cotton spinning industry seem to have mainly caused prices to fall, leaving little excess profits for the firms that were innovating. Thus, a third firm, Richard Hornby and Partners, in the years 1777 to 1809 was in a sector of the industry, hand loom weaving, which had not yet been transformed by any technological advance. Yet, its average profit rate was 11.4%, as high as Samuel Greg in the innovating part of the industry. The conclusion is that the host of innovations in cotton textiles does not seem to have particularly rewarded the innovators. Only a few such as Arkwright and the Peels became noticeably wealthy. Of the 379 people probated in 1860–1869 in Britain who left estates of £0.5 million or more, only 17 (or 4%) were in the textile industry, even though, as noted from 1760–1769 to 1860–1869, this one sector generated nearly half the productivity growth in the economy (Rubinstein, 1981, 60–7). The Industrial Revolution economy was spectacularly bad at rewarding innovation. Its innovators captured little of the rewards. The Industrial Revolution did not make paupers into princes. This is why

Table 5.3 The gains from innovation in textiles in the Industrial Revolution

Innovator	Device	Result
John Kay	Flying Shuttle, 1733	Impoverished by litigation to enforce patent. House destroyed by machine breakers in 1753. Died in poverty in France.
James Hargreaves	Spinning Jenny, 1769	Patent denied. Forced to flee by machine breakers in 1768. Died in workhouse in 1777.
Richard Arkwright	Water Frame, 1769	Worth £0.5 million at death in 1792. By 1781, other manufacturers refused to honor patents. Made most of his money after 1781.
Samuel Crompton	Mule, 1779	No attempt to patent. Grant of £500 from manufacturers in the 1790s. Granted £5000 by Parliament in 1811.
Reverend Edmund Cartwright	Power Loom, 1785	Patent worthless. Factory destroyed by machine breakers. Granted £10,000 by Parliament in 1809.
Eli Whitney (USA)	Cotton Gin, 1793	Patent worthless. Later made money as a government arms contractor.
Richard Roberts	Self-Acting Mule, 1830	Patent revenues barely covered development costs. Died in poverty in 1864.

Source: Clark (2007), Table 12.2

Britain has few foundations to rival the great private philanthropies and universities of the USA.

A similar tale can be told for the other great nexus of innovation in Industrial-Revolution England: coal mining; iron and steel; and railroads. Coal output, for example, exploded in England in the Industrial Revolution era. This coal heated homes, made ore into iron, and powered railway locomotives. Yet there were no equivalents of the great fortunes made in oil, railways, and steel, in America's late 19th century industrialization. The coal fields in the northeast yielded modest mineral rents to the owners throughout the years 1700 to 1870. Coal rents were typically 10% or less of the value of coal at the pithead, and 5% or less of the retail price of coal to the consumer in places like London throughout these years (Clark and Jacks, 2007, 48). The operators of the pits again seem to have generated modest results on their investments in shafts, underground roads, and winding gear. The technological gains which made an enormous expansion of coal output possible, such as the steam engine, seem to have been relatively modest. These new techniques, which allowed access to ever deeper coal seems, were available to all coal mines in areas like the northeast coalfield, without any return to the pioneers.

The new industrial priesthood, the engineers who developed the English coalfields, railways, and canals, made prosperous but typically moderate livings. Though their names survive—Richard Trevithick; George and Robert Stevenson; Humphrey Davy—they again captured very little of the social rewards for their enterprise. Richard Trevithick, the pioneer of locomotives, died a pauper in 1833. George Stevenson, whose famous locomotive, the *Rocket*, in a trial in 1829 ran loaded at 15 miles an hour, an unheard of speed for land travel in this era, did much better; but his country house in Chesterfield was, however, a pittance compared to his substantial contributions to railway engineering. But other locomotives competed in the famous trial, and soon a swarm of locomotive builders were supplying the railway network.

Innovation during the Industrial Revolution era typically benefited consumers in the form of lower prices. As coal output exploded, real prices to consumers steadily declined: the real price in the 1700s was 60% greater than in the 1860s. Coal, iron and steel; and rail carriage all remained highly competitive in England in the Industrial Revolution era. The patent system offered little protection to most of the innovations in these sectors, and innovations quickly leaked from one producer to another.

The rise in innovation rates in Industrial Revolution England was not induced by unusual rewards for innovation, but by a greater supply of innovation at still modest rates of reward. The institutionalist perspective is that the rewards offered by the market shifted upwards compared to all previous pre-industrial economies. There is no evidence of any such change. The last significant reform of the patent system was in 1689, more than 100 years before efficiency gains became common. And the patent system itself played little role for most innovation in Industrial Revolution England.

Instead the upsurge in innovation in the Industrial Revolution period reflected a surge in supply. With the benefits to innovation no greater than in earlier economies, the supply still rose substantially. Facing the same challenges and incentives as in other economies British producers were more likely to attempt novel methods of production.

Productivity growth in cotton textiles in England from 1770 to 1870, for example, far exceeded that in any other industry. But the competitive nature of the industry, and the inability of the patent system to protect most technological advances, kept profits low. Cotton goods were homogenous. Yarn and cloth sold in wholesale markets where quality differences were readily perceptible to buyers. The efficient scale of cotton spinning and weaving mills was always small relative to the market. New entrants abounded. By 1900, Britain had about 2000 firms in the industry. Firms learned improved technique from innovating firms through hiring away their skilled workers. The machine designers learned improved techniques from the operating firms. The entire industry—the capital goods makers and the product producers—over time, clustered more and more tightly in the Manchester area. By 1900, 40% of the entire world output of cotton goods was produced within 30 miles of Manchester. The main beneficiaries of this technological advance were the consumers of textiles all across the world, and the owners of land in the cluster of textile towns that went from being largely worthless agricultural land to valuable building sites.

The greatest of the Industrial Revolution cotton magnates, Richard Arkwright, is estimated to have left £0.5 million when he died in 1792.[9] His son, also Richard Arkwright, inherited his father's spinning mills. But though his son had managed his own mills and had much experience in the industry which was still showing rapid productivity growth, he soon sold most of his father's mills, preferring to invest in land and government bonds. By 1814, he owned £0.5 million in government bonds alone. He prospered mainly on government bonds and real estate, leaving £3.25 million when he died in 1843 despite sinking much money into a palatial country house for his family.[10] But Arkwright Senior accumulated less wealth than Josiah Wedgwood, who left £0.6 million in 1795, even though Wedgwood operated in the pottery sector, which enjoyed far less technological progress (pottery was still handmade, by and large, even in the late 19th century).

Though the first great innovations of the Industrial Revolution era did not offer much in the way of supernormal profits because of the competitive nature of the industry, the second, railroads, seemed to offer more possibilities. Railways are a technology with inherent economies of scale. At the very minimum, one line has to be built between two cities, and once it is built a competitor has to enter with the minimum of a complete other line. Since most city pairs could not profitably support multiple links, exclusion, and hence profits, thus seemed possible.

The success of the Liverpool-Manchester line in 1830 (by the 1840s equity shares on this line were selling for twice their par value) inspired a long period of investment in railways. Figure 5.7 shows the rapid growth of the railway network in England from 1825 to 1869, by which time more than 12,000 miles of track had been laid across the tiny area of England. This investment and construction was so frenetic that so-called railway manias struck in 1839 and 1846.

But again the rush to enter quickly drove down profit rates to very modest levels, as Table 5.4 shows. Real returns, the return on the capital actually invested, by the 1860s were no greater than for very safe investments in government bonds or agricultural land. While railway lines had local monopolies, they ended up in constant competition with each other through roundabout routes.

Therefore, while, for example, the Great Western may have controlled the direct line from London to Manchester, freight and passengers could cross over through other companies to link up with the East Coast route to London. Again, profits inspired imitation which could not be excluded and the profit was squeezed out of the system. Consumers were again the main beneficiaries.

[9] Fitton (1989, p. 219).

[10] Ibid (p. 296).

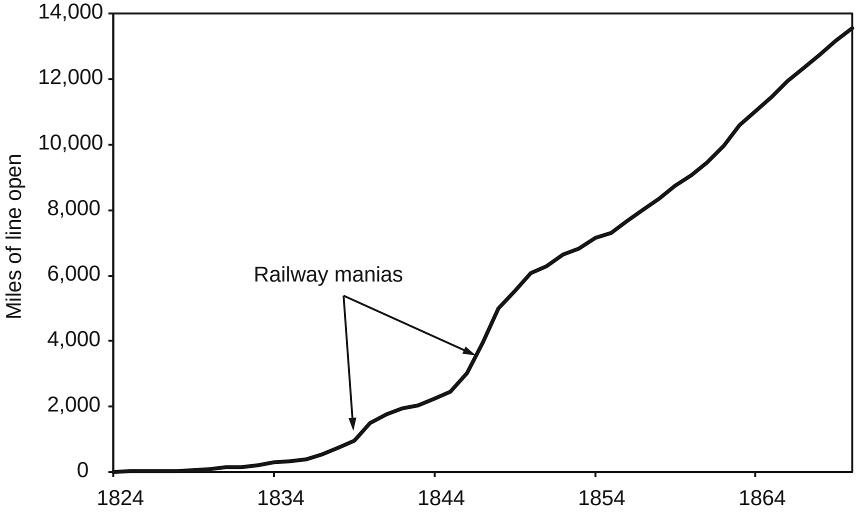


Figure 5.7 English railroad construction, 1825–1869. *Source:* *Mitchell and Deane (1971, p. 225)*.

Table 5.4 Profit rates on the capital invested in British-owned railways, 1860–1912

Period	Rate of return, UK (%)	Rate of return, British Empire (%)	Rate of return, foreign lines (%)
1860–9	3.8	–	4.7
1870–9	3.2	–	8.0
1880–9	3.3	1.4	7.7
1890–9	3.0	2.5	4.9
1900–9	2.6	1.6	4.4
1910–13	2.6	3.1	6.6

Source: Clark (2007), Table 14.7.

It is for this reason that in Britain, unlike in the USA, there are very few universities and major charities funded by private donors.[11] The Industrial Revolution did not result in great individual or family fortunes in England. By the 1860s, the rich were still by and large the descendants of the landed aristocracy. Of 379 men dying between 1860 and 1879 in Britain, who left at least £0.5 million, 256 (68%) owed their wealth to inherited land. As noted above, only 17 (4%) were textile magnates, despite textiles being the driving industry in Industrial Revolution productivity advance.[12]

[11] The industrialization of the United States created much greater private and family fortunes.
[12] Rubinstein (1981, pp. 60–7).

The unsatisfactoriness of conventional institutional accounts—which emphasize returns to innovation and to investment in general—has led to exploration of other avenues by which institutions may matter. Avner Greif, Murat Iyigun, and Diego Sasson wrote a recent paper which argues that the Industrial Revolution was underpinned by English welfare institutions, dating to the early 16th century, which insured against failure (Greif et al. 2012). It was not the size of the rewards on the upside that distinguished England from other societies such as China, but the cushion against failure for those who tried and did not succeed. James Hargreaves, inventor of the spinning jenny, may have died in the workhouse in 1777, but at least he did not die in the street. However, this is a bit like saying that New York has developed a high risk, high rewards financial sector because it allows for financial support for adults without minor dependents in a way not found, for example, in Texas. Presumably, the Harvard graduates in the financial sector have backup plans, other than general relief, if their hedge fund fails.

One thing that is striking about institutionalist explanations in general is the absence of any agreed metric for institutional quality. There is a belief in the physical sciences that a basic element in any scientific analysis of any phenomenon is to have a defined objective, and shared system of measurement. Institutionalists on this standard are still in the pre-science world of phlogiston and other early theories.

5.3. CHANGES IN PEOPLE

The modest signs of any increase in returns to innovation at the time of the Industrial Revolution suggest as an alternative that the transition was instead driven by changes in the aspirations and capabilities of economic agents. And this has been the theme for another set of explanations of the Industrial Revolution. In this extensive set of theories, a rise in human capital investment, and consequent improvement in the capabilities of economic actors, is key to the transition between the Malthusian regime and the modern (Becker et al. 1990; Lucas, 2002; Galor and Weil, 2000; Galor and Moav, 2002; Galor, 2011).

We certainly see that the English population on the eve of the Industrial Revolution had characteristics that differed from most pre-industrial societies. In particular, the levels of literacy and numeracy were high by the standards of the pre-industrial world. Even the great civilizations of the past, such as the Roman Empire, or the city states of the Italian Renaissance, had general levels of literacy and numeracy that were surprisingly low by the standards of Industrial Revolution England. And we know as a general feature that modern, high-income, fast-growth economies are distinguished by high levels of human capital. So increases in human capital that created knowledge externalities, at the gross level, would seemingly be a candidate source of the Industrial Revolution.

We find interesting evidence that the average numeracy and literacy of even rich people in most earlier economies was surprisingly poor. A prosperous landowner in Roman Egypt, Isidorus Aurelius, for example, variously declared his age in legal documents in a less than two-year span in 308–309 AD as 37, 40, 45, and 40. Clearly, Isidorus had no clear idea of his age. Other sources show he was illiterate (Duncan-Jones, 1990, p. 80). A lack of knowledge of their true age was widespread among the Roman upper classes as evidenced by age declarations made by their survivors on tombstones. In populations where ages are recorded accurately, 20% of the recorded ages will end in 5 or 10. We can thus construct a score variable Z, which measures the degree of "age heaping" where $Z = \frac{5}{4}(X - 20)$, and X is the percentage of age declarations ending in 5 or 10 to measure the percentage of the population whose real age is unknown. This measure of the percentage of people who did not know their true age correlates moderately well in modern societies also with the degree of literacy.

Among those wealthy enough to be commemorated by an inscribed tombstone in the Roman Empire, typically half had unknown ages. Age awareness did correlate with social class within the Roman Empire. More than 80% of office holder's ages seem to have been known by their relatives. When we compare this with death records for modern Europe we find that by the eve of the Industrial Revolution age awareness in the general European population had increased markedly, as Table 5.5 shows.

We can also look at the development of age awareness by looking at censuses of the living, as in Table 5.6. Some of the earliest of these are for medieval Italy, including the famous Florentine *Catasto* of 1427. Even though Florence was then one of the richest cities of the world, and the center of the Renaissance, 32% of the city population did not know their age. In comparison, a census in 1790 of the small English borough of Corfe Castle in Dorset, with a mere 1239 inhabitants, most of them laborers, shows that all but 8% knew their age. In 1790, again awareness correlates with measures of social class, with universal knowledge among the higher status families, and lower age awareness among

Table 5.5 Age heaping, Rome versus later Europe

	Social group	Sample size	Innumeracy rate
Imperial Rome			
Rome	All	3708	48
Italy outside Rome	All	1395	43
Italy outside Rome	Town Councilors	75	15
Modern Europe, death records			
Geneva, 1560–1600	All	–	54
Geneva, 1601–1700	All	–	44
Geneva, 1701–1800	All	–	23
Liege, 1740	All	–	26
Paris, c. 1750	All	–	15

Source: Duncan-Jones (1990, pp. 84–90).

Table 5.6 Age heaping among living populations (23–62)

Place	Date	Type of community	Sample size	Z
Town of Florence	1427	Urban	–	32
Florentine Territory	1427	Rural	–	53
Pistoia	1427	Urban	–	42
Pozzuoli	1489	Urban	–	72
Sorrento	1561	Urban	–	67
Corfe Castle, England	1790	Urban	352	8
Ardleigh, England	1796	Rural	433	30
Terling, England—Poor relief recipients	1801	Rural	79	19

Notes: The total population of Corfe Castle was 1239, and of Ardleigh 1145.
Source: Duncan-Jones (1990). Terling, Essex Record Office D/P 299/12/3. Ardleigh, Essex Record Office, D/P 263/1/5.

the poor. But the poor of Corfe Castle or Terling in Essex had as much age awareness as office holders in the Roman Empire.

Another feature of the Roman tombstone age declarations is that ages seem to be greatly overstated for many adults. Thus, while we know that life expectancy in ancient Rome was probably in the order of 20–25 at birth, tombstones record people as dying at ages as high as 120. For North African tombstones, for example, 3% of the deceased are recorded as dying at age 100 or more.[13] Almost all of these 3% must have been 20–50 years younger than was recorded. Yet their descendants did not detect any implausibility in recording these fabulous ages. In contrast, the Corfe Castle census records a highest age of 90, well within the range of possibilities given life expectancy in rural England in these years.

Therefore, another explanation for the Industrial Revolution is that while the incentives to innovate were not greater, the capabilities and aspirations of economic agents had improved. This raises two important issues. First, why did history move in a general direction toward increasing levels of literacy and numeracy? What internal dynamic drove this move? Second, was England sufficiently distinct from earlier societies in terms of the abilities of its economic agents to account for the transition to modern growth?

Figure 5.8 shows, for example, literacy rates, measured by a person's ability to sign his or her name, in England 1580–1920. Two things stand out: the first is that literacy rates for men rose substantially long before the Industrial Revolution. If mass literacy was the key to growth then seemingly the Industrial Revolution would have again appeared 100 years before the 1780s. The second is that dramatic increases in literacy rates are a phenomenon only of the late Industrial Revolution period, the years 1850–1900. Literacy in the Industrial Revolution period itself rose by modest amounts.

[13] Hopkins (1966, p. 249).

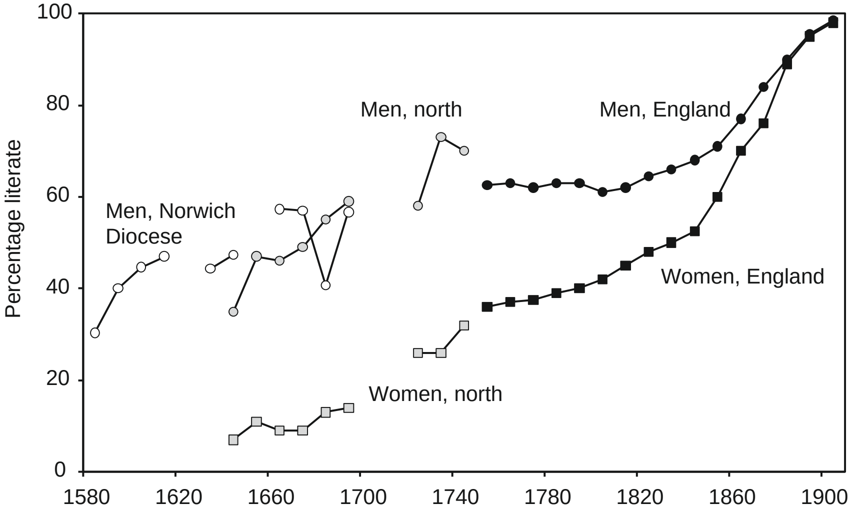


Figure 5.8 Literacy in England, 1580–1920. *Sources: 1750s–1920s, Schofield, 1973, men and women who can sign marriage resisters. The north, 1630s–1740s, Houston (1982), witnesses who can sign court depositions. Norwich Diocese, 1580s–1690s, Cressy (1977), witnesses who can sign ecclesiastical court declarations. Source: Clark (2007), Figure 9.3, p. 179.*

Also, literacy rates in England in 1780 were not high by the standards of many other parts of northwest Europe. Literacy rates then exceeded those of England, in Scotland, the Netherlands, much of Germany, and in Scandinavia. But with those caveats we can ask what might have driven the trend all across northern Europe to greater levels of numeracy and literacy by the eve of the Industrial Revolution.

Another caveat about the role of numeracy and literacy is that given the observed rates of those returning to schooling, the increased investment in countries like England in the Industrial Revolution period can account little for faster productivity growth rates. Thus we can modify Equation (5.1) to allow for investment in human capital to:

$$g_y = a_k g_k + a_h g_h + g_A, \tag{5.4}$$

where a_h is the share of income attributable to human capital investments and g_h is the growth rate of the stock of human capital. But the growth rate of the human capital stock in England 1760–1860 implied by Figure 5.8 is very modest: less than 0.4% per year. And even if we allowed one third of all the 60% share of wage payments in income in Industrial Revolution England to be attributed to human capital, this would entail human capital investments increased income growth rates by a mere 0.08% per year. If human capital lies at the heart of the Industrial Revolution it must be because there are significant external benefits associated with human capital investments, as Lucas (1988), hypothesized.

Why then did education levels rise in the centuries leading up to the Industrial Revolution? A theme of many economic models of the transition from Malthusian stagnation to modern growth listed above is that there was a switch from quantity (or at least desired quantity) to quality, in families as we moved to the modern world. This theme has been driven by the observation in modern cross-sections, looking across countries, that high income, high education societies are those with few children per woman. Also within high income societies there was a period between 1890 and 1980 where again, lower income families were those with more children.

Such theories face a number of challenges in modeling the actual world of Industrial Revolution England. The first challenge is that these theories are expressed always in terms of children surviving to adulthood. In the modern world, in most societies, child survival rates are high, and so in practice, births and surviving children are closely equivalent. But in all known pre-industrial societies, including pre-industrial England, large numbers of children did not survive even to their first year. In these cases, the distinction between births and surviving children becomes important. Measured in terms of births, Malthusian societies witnessed high fertility, with the average woman surviving to age 50 giving birth to five children. But in such societies the average number of children surviving to adulthood was only two.

Further, since children who died in the pre-industrial world tended to do so fairly early, the numbers of children in any household at any time in the pre-industrial world would typically be three or less. For example, of 1000 children born in England in 1700–1724, nearly 200 would be dead within 6 months (Wrigley et al. 1997). Pre-industrial families would look similar to the families of America in the high growth 1950s and 1960s. Pre-industrial families thus faced remarkably similar trade-offs between the number and quality of children as do modern families. In some sense there has been no change in fertility from the pre-industrial to the modern world, measured in net as opposed to gross terms.

The second challenge these theories face is that in England the transition from high births per woman to lower levels of births per woman did not occur at the onset of the Industrial Revolution, but only 100 years later in the 1880s.[14] Fertility in England did not show any decline at the aggregate level prior to 1880. Indeed the opposite occurred, as Figure 5.9 illustrates. Births per woman, and also net fertility, rose precisely in the period of the Industrial Revolution in England.

The third challenge is that in cross-section in pre-industrial England there was a strong positive association between net fertility and the wealth or occupational status of families. Figure 5.10, for example, shows in twenty-year periods, the numbers of children alive at the time wills were made for married men in England marrying 1520–1879, where those leaving wills are divided into wealth terciles defined across the whole sample. The lowest tercile in wealth would still be men of above median wealth at death. Their implied net fertility is similar to that for men as a whole in England, as revealed by Figure 5.9.

[14] France was the only country to experience a decline in fertility starting in the late 18th century, and France of course lagged Britain in terms of the onset of modern growth.

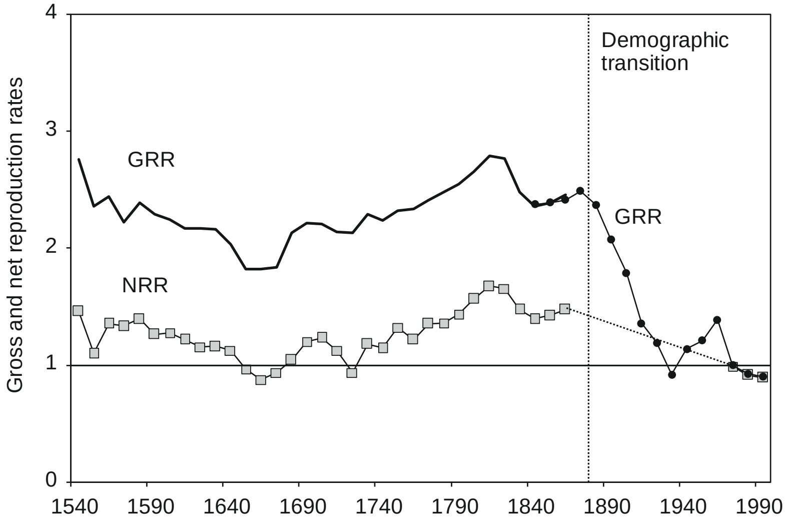


Figure 5.9 The fertility history of England, 1540–2000. *Source:* Clark (2007), *Figure 14.6, p. 290.*

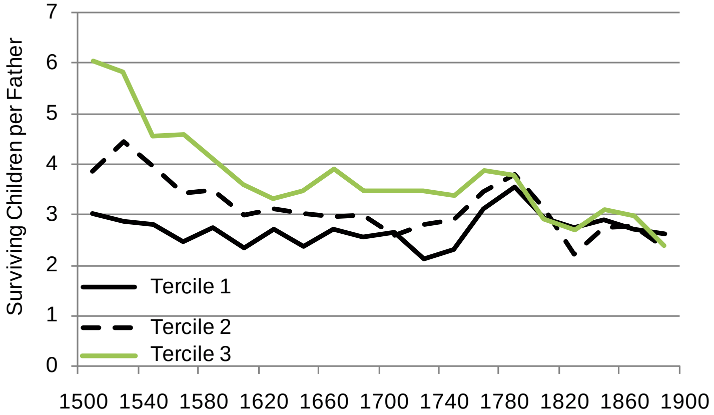


Figure 5.10 Net fertility by wealth terciles, marriage cohorts, 1520–1879. *Source:* Clark and Cummins *(2013a)*.

But the men of the top wealth tercile marrying before 1780 were leaving on average 3.5–4 surviving children. The most educated and economically successful men in pre-industrial England were those with the largest numbers of surviving offspring. Matching these men to parish records of births shows that this advantage in numbers of surviving children stems largely from the greater fertility of the wives of richer men. Their gross fertility was equivalently higher. This positive association of economic status and fertility

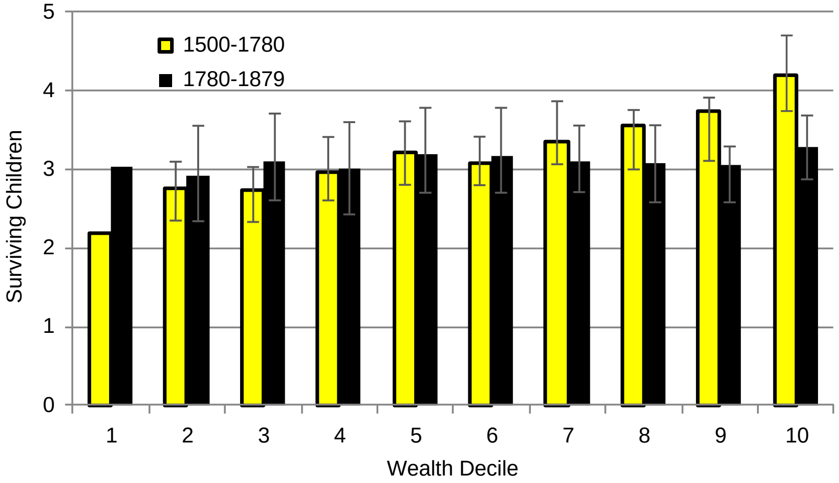


Figure 5.11 Net marital fertility by wealth decile, marriages 1500–1779 and 1780–1879. *Note*: The lines at the top of the columns indicate the 95% confidence interval for the net fertility of these groups relative to the decile of lowest asset income. All assets normalized by the average wage in the year of death from Clark (2010). *Source: Clark and Cummins (2013a)*.

pre-1780 has been confirmed in an independent study of gross fertility in parish records in England 1538–1837 by Boberg-Fazlic et al. (2011).

For marriages between 1780–1879, this pattern of high fertility by the rich and educated disappears. Instead, we have for most of the Industrial Revolution period, an interval where fertility is unlinked to education, status, or wealth. Figure 5.11 shows the dramatic shift in pattern this represents, grouping married men by wealth deciles. Another feature revealed in Figure 5.11 is that the pattern of higher net fertility with wealth before 1780 continues all through the wealth spectrum. There is no wealth level at which we observe any decline in net fertility.

The delay in the decline in aggregate fertility levels in England till after the Industrial Revolution represents a formidable challenge for theories that seek to explain the Industrial Revolution through a quality-quantity trade-off, and rising levels of human capital. However, these recent findings that richer families did indeed reduce their fertility just at the time of the onset of the Industrial Revolution offers some hope for models based on heterogenous agents as opposed to a single representative agent. But if richer families were changing their behaviors in response to economic signals, we would expect to find in this period signs of greater returns to human capital investments. Another problem for quality-quantity models of the Industrial Revolution is that such evidence is lacking. Figure 5.12, for example, shows the earnings of building craftsmen—carpenters, masons, bricklayers, plasterers, painters, plumbers, pavers, tilers, and thatchers—relative to unskilled building laborers and assistants. The skill premium is actually at its highest

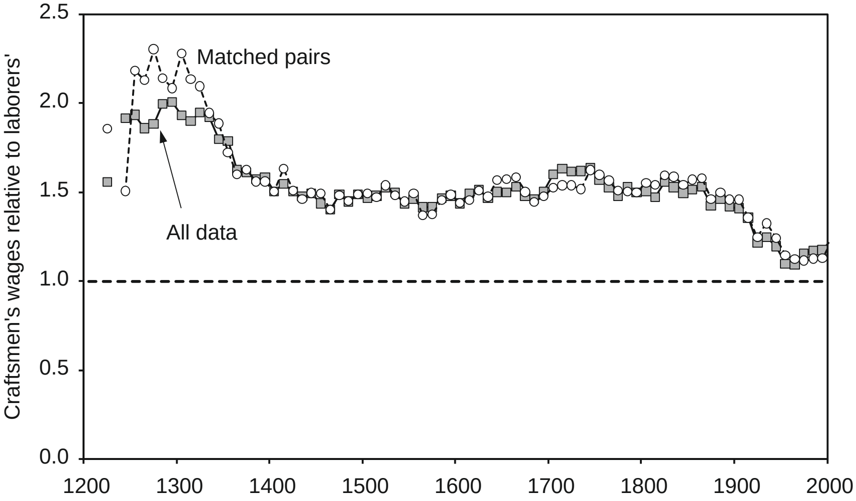


Figure 5.12 The skill premium, building workers, England, 1220–2000. *Source:* Clark (2005).

in the interval 1220–2000 in the earliest years, before the onset of the Black Death in 1348, when a craftsman earned nearly double the wage of a laborer. If there was ever an incentive to accumulate skills it was during the early economy. Thereafter, it declines to a lower but relatively stable level from about 1370 until 1900, a period of over 400 years, before declining further in the 20th century. Thus, the time of the greatest market reward for skills and training was long before the Industrial Revolution. And the period of the demographic transition in England, the switch toward smaller family sizes circa 1880, is not associated with any rise in the skill premium.

The information on the skill premium in building may be criticized as showing only the returns to a very limited form of human capital. What about wider measures of the impact of quantity of children before and after the Industrial Revolution on child outcomes? Do we find that for marriages prior to 1780 there is little or no cost in terms of child outcomes where richer families have more children, but that after 1780 a quality-quantity trade-off becomes evident?

The same source that was used above to measure net fertility as a function of wealth and socio economic status, men's wills, can also be employed to measure the effects of the number of children on the outcomes for children before and during the Industrial Revolution (Clark and Cummins, 2013b).

In measuring the quality-quantity trade-off in the modern world the problem has been that "high quality" families tend to have fewer children. The observed relationship between quality and quantity may thus reveal no underlying causal relationship. In capturing the true quality-quantity trade-off, researchers have had to control for the inherent endogeneity of family size. We can thus portray parent influences on child "quality" as

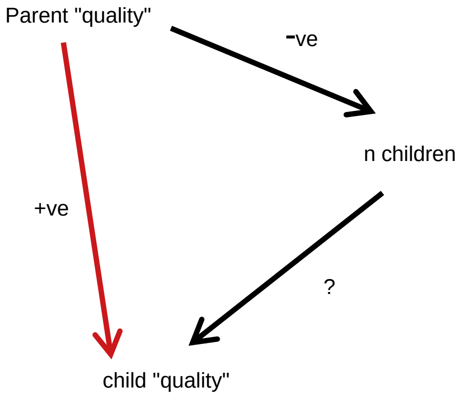


Figure 5.13 Parent influences on child quality.

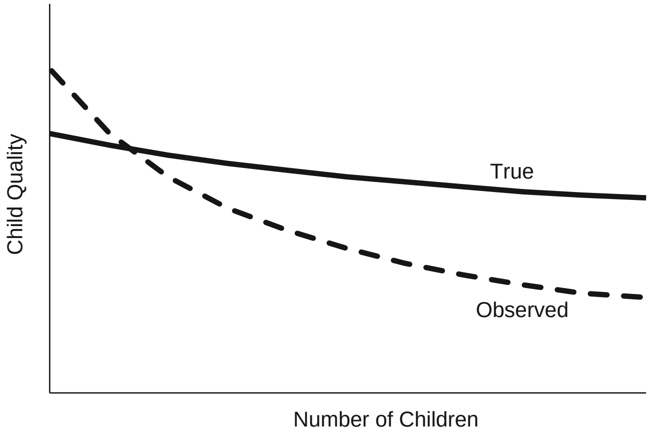


Figure 5.14 The true and observed quality-quantity trade-off.

following two potential routes, as in Figure 5.13. Since, in the modern world, high quality parents also tend to have smaller numbers of children, the observed negative correlation between n and child quality may stem just from the positive correlation of parent and child quality. As Figure 5.14 shows, the estimate of the trade-off between quantity and quality will be too steep using just the observed relationship. Estimates of $\hat{\beta}$ in β the regression

$$q = \beta n + u, \tag{5.5}$$

where q is child quality, n child numbers, and u the error term, are biased toward the negative because of the correlation between n and u.

To uncover the true relationship investigators have followed a number of strategies. The first is to look at exogenous variation in family size caused by the "accident" of twin births (e.g. Rosenzweig and Wolpin, 1980; Angrist et al. 2010; Li et al. 2008). In a world where the modal family size is 2, there are a number of families who accidentally end up with 3 children because their second birth is of twins. What happens to the quality of these children compared to those of two-children families? These studies find the uncontrolled relationship between quantity and quality decreases. Indeed, it is often insignificant and sometimes positive (Schultz, 2007, 20). For instance, Angrist et al. (2010), find "no evidence of a quality-quantity trade-off" for Israel using census data. Li et al. (2008), however, do report the expected relationship instrumenting using twins in China, but only in the Chinese countryside. But in China there are government policies designed to penalize couples who have more than the approved number of children, so we may not be observing anything about the free market quality/quantity trade-off.

In summary, there is a clear raw negative correlation in modern populations between child numbers and various measures of child quality. However, once instruments and other controls to deal with the endogeneity of child quality and quantity are included, the quality-quantity relationship becomes unclear. The quality-quantity trade-off so vital to most theoretical accounts of modern economic growth is, at best, unproven.

However, we see above that in the period 1540–1780 in England the modern negative relationship between child numbers and parent quality is reversed and is instead positive. Thus in this period in estimating β in Equation (5.5) we will find that $\hat{\beta}$ is in this case biased instead toward 0. Figure 5.15 shows this effect. Any negative effects of quantity on quality found will be underestimated, as opposed to the bias in estimating β in modern studies. Then there is the intermediate fertility regime in England, with marriages formed 1780–1880, where parent quality and numbers of children are uncorrelated, so that $\hat{\beta}$ will be unbiased.

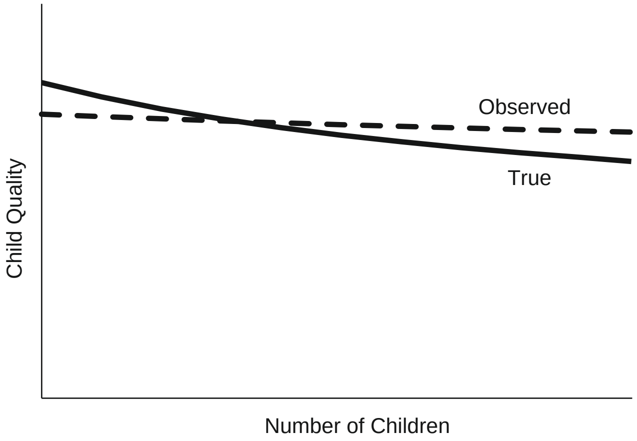


Figure 5.15 The true and observed quality-quantity trade-off, marriages pre-1780.

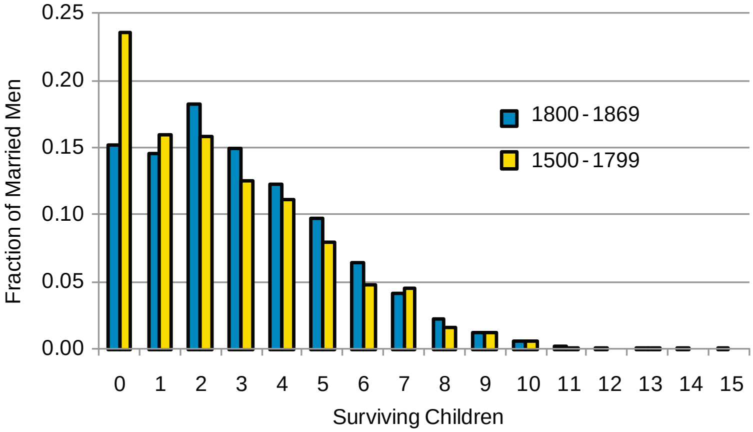


Figure 5.16 The distribution of net family sizes in pre-industrial England. *Note*: Number of observations before 1800, 6940; after 1800, 1418. *Source: Clark and Cummins (2013b)*.

The second advantage of the pre-industrial data from England for observing the quality-quantity trade-off, is the much greater variation in family sizes before 1870 than in the modern world, and the evidence that this variance was largely the product of chance, like modern twin births. Figure 5.16 shows the distribution of the number of surviving children per father, at the time of the father's will, for fathers marrying 1500–1799, and 1800–1869. This number will include children from more than one wife, where a first wife died and the husband remarried.

As noted above, we can measure family size in two ways. A second is the number of births per family, gross fertility. This is shown in Figure 5.17, giving the distribution of births per mother for the wives of men marrying in England 1500–1799, where the husband had only one wife. Thus, despite the average of five births per wife, in 10% of all marriages there was only one child born, in about 20%, only two. The number of baptisms is the overwhelming explanator of the number of surviving children per man. The R^2 of the regression predicting numbers of surviving children from the number baptized is 0.73. On average, 0.62 of each child born would be alive at the time of the will. If we include in this regression indicators for location, social status, wealth, and time period then the R^2 increases only marginally to 0.75. At the individual family-level both gross fertility, births, and net fertility, the number of surviving children, were largely random variables. Only a tiny fraction of the variation in each is explained by correlates such as wealth, occupation, literacy, and location.

When the coefficient β in the equation:

$$q = \beta_n + u$$

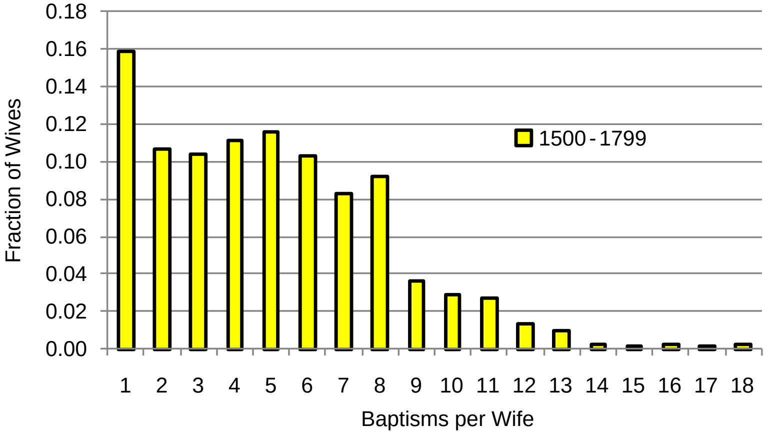


Figure 5.17 The distribution of number of baptisms per wife, 1500–1799. *Note*: Number of observations before 1800, 818. *Source: Clark and Cummins (2013b).*

is estimated by OLS estimate of β will be, in the limit,

$$\hat{\beta} = \beta + \frac{\text{cov}(n, u)}{\text{var}(n)}.$$

But in pre-industrial England, the degree of bias this will impart will be small because n was largely a random variable, so the bias in estimating β will be correspondingly very slight.

Thus, suppose $n = \theta u + e$. Then:

$$\frac{\text{cov}(n, u)}{\text{var}(n)} = \frac{\theta \text{var}(u)}{\theta^2 \text{var}(u) + \text{var}(e)}.$$

The greater is var(e), the random component in n, then the less the bias in the estimate of β. We show below that for marriages formed before 1870, var(e) was enormous relative to θ^2var(u). We can thus use the observed correlation between quality and quantity in this period as a measure of the true underlying causal connection between quantity and quality in the years before and during the Industrial Revolution.

We have three measures of child quality for sons born over the years 1500–1879: the wealth of those probated, the socioeconomic status of those probated, and the probate rates of all sons. The likelihood of a man being probated was strongly linked to their wealth and social status. Probate was only required if the estate of the deceased exceed a certain limit. In 1862, 65% of men of high socioeconomic status (professionals and gentlemen) were probated, compared to 2% of laborers (Clark and Cummins, 2013a).

Table 5.7 Social distribution among will makers, and father-son pairs

Social group	*N* all wills	% all wills	*N* father-son	% father-son
Gentry/Independent	405	7	220	15
Merchants/Professionals	506	9	167	11
Farmers	1906	33	605	41
Traders	883	15	152	10
Craftsmen	1132	19	217	15
Husbandmen	708	12	99	7
Laborers/Servants	268	5	16	1

Source: Clark and Cummins (2013b).

The sample of father-son pairings is very much biased toward the rich. As Table 5.7 shows, will makers in the years 1500–1920 were disproportionately from the upper social groups. In 1862, the bottom two social groups in the table were 40% of men dying, but they represent only 8% of fathers and sons where both were probated (Clark and Cummins, 2013a, Table A.12). In contrast, the top three social groups represented 13% of men dying in 1862, but a full 67% of those where both father and son were probated. Thus, what we are principally looking at here is the effects of family size on the outcomes for children of the upper third of the population in pre-industrial England. But this is the group whose behavior was changing first around 1780, then around 1880, in the two-stage demographic transition observed in Industrial Revolution England.

The effect of family size on wealth is estimated from the size of the coefficient b_2 in the expression:

$$\ln W_s = b_0 + b_1 \ln W_f + b_2 \ln N + b_3 \mathrm{DFALIVE} + e, \tag{5.6}$$

where N is the number of surviving children; $\ln W_s$ the average log wealth of sons of a given father; and DFALIVE the fraction of sons for whom the father was alive at the time of son's probate. DFALIVE is a control for the effects of sons who die before fathers, and are therefore likely to receive smaller transfers of wealth from fathers. Such sons will also tend to be younger. And in this data wealth rises monotonically with age until men are well past 60. Since some fathers had more than one probated son, we averaged wealth across the probated sons and treated each family as the unit of observation.

With this formulation, b_3 is the elasticity of a son's asset income as a function of the number of surviving children the father left. N varies in the subsample of fathers and sons from 1 to 13. The coefficient b_2 shows the direct link between fathers' and sons' wealth, independent of the size of the fathers' family.

Table 5.8 shows the estimated coefficients from Equation (5.6) for fathers dying 1500–1920. The results are reported for the data pooled across all years, and for fathers dying 1500–1819 (who would have sons born up until 1800, typically), those dying 1820–1880,

Table 5.8 Sons' wealth and family size

	All	Pre–1820	1820–1880	1880–1920
LnWf	.502***	.560***	.527***	.457***
	(.030)	(.051)	(.073)	(.046)
LnN	−.311***	−.241***	−.312	−.390**
	(.082)	(.090)	(.227)	(.176)
DFALIVE	−.868***	−.710**	−.611	−.866*
	(.258)	(.314)	(.643)	(.448)
Constant	2.032***	1.929***	2.024***	1.696***
	(.158)	(.210)	(.502)	(.341)
Obs	1,029	610	175	244
R-squared	.292	.306	.281	.302

Source: Clark and Cummins (2013b).
Robust standard errors in parentheses.
*Significantly different from 0 at the 10% level.
**Significantly different from 0 at the 5% level.
***Significantly different from 0 at the 1% level.

and post-1880. The link between fathers' and sons' wealth as revealed by the estimate of b_1 is highly significant and stable across the subperiods.

The estimated coefficient on the log of surviving children is negative in all three periods, as would be implied by a quality-quantity trade-off. So this study is unusual in finding for the early period a quantitatively and statistically significant effect of family size on son outcomes. However, even though it will be potentially biased toward zero for fathers dying before 1820, the value in these earlier years is estimated as being similar to that in 1820–1880.[15] There is no indication in this data of a substantially more adverse quality-quantity trade-off with the arrival of the Industrial Revolution. There is nothing in the estimates of Table 5.8 to suggest that changing family sizes among the wealthy and educated in Industrial Revolution England were driven by a changing quality-quantity trade-off. Again the economic environment seems stable as the dramatic changes of the Industrial Revolution were occurring.

The predicted quantitative effects of sibling size on wealth at death are shown in Figure 5.18, where wealth at a family size of 1 fixed at 1. Pooling all the data, the effects of family size on the outcomes for children measured in terms of wealth are actually reasonably modest. Moving from a family of one child (with our data by definition a boy), to one of 10 children, reduces the average wealth of sons by only 51%. This is demonstrated visually in Figure 5.18.

This is not a very strong effect if the main transmission of wealth was through division of a fixed pie of wealth among children (the red line in Figure 5.18). For in that case

[15] The bias, as argued above, will be small before 1880 because of the randomness of family sizes.

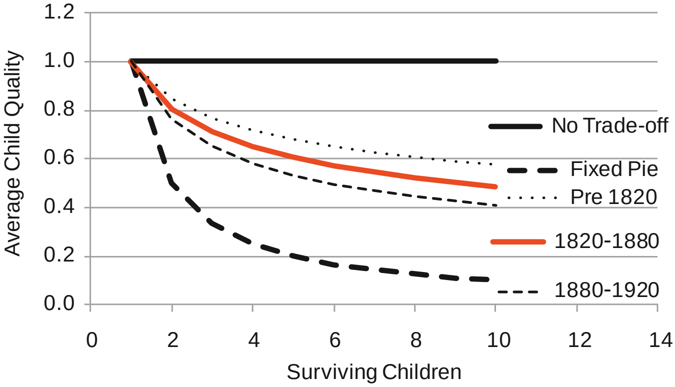


Figure 5.18 The empirical Quality-Quantity effect, 1500–1920. *Source:* *Clark and Cummins (2013b)*.

the expected coefficient on ln$N$ should be -1. The average wealth of the children of a family of 10 would be only 10% of that of a family with only one sibling. We can derive similar estimates of the effect of family size by period on the chances of being probated, and on occupational status. In each case, the effects are in the right direction, but even more modest than for wealth (Clark and Cummins, 2013a,b).

The facts above, regarding the transition from pre-industrial to modern fertility in England in the Industrial Revolution era, represent a formidable challenge to those trying to model the Industrial Revolution in a child quality-quantity framework. Since some of these patterns were discovered only in the last few years, such as the strong positive association of wealth and fertility in pre-industrial England, many of these models fail to capture essential features of the fertility transitions (Clark and Hamilton, 2006; Clark and Cummins, 2013a; Boberg-Fazlic et al. 2011).

Some of the theory papers mentioned above, such as that of Becker et al. (1990), fail at the first challenge. They posit a pre-industrial world that never existed of high net fertility and rapid population growth. And while they model a world with two equilibria—one where parents invest nothing in the human capital of their children, and the other where they invest considerable human capital—the escape from the zero human capital Malthusian trap is exogenous to the model. "Technological and other shocks" (Becker et al. 1990, p. S32) somehow raise the level of human capital far enough above zero to lead to a convergence to the high growth equilibrium. These shocks are conceived to be "improved methods to use coal; better rail and ocean transports; and decreased regulation of prices and foreign trade" (Becker et al. 1990, p. S33). But how such shocks get translated into human capital is never specified. With the arrival of highly paid unskilled work in textile factories during the Industrial Revolution, for example, we would expect, in the Becker, Murphy, and Tamura model, a reduction in educational investment.

Robert Lucas creates a Malthusian trap with many of the same characteristics of Becker, Murphy, and Tamura (Lucas, 2002), but which tries to model better pre-industrial fertility, measured as surviving children, so that it increases in income. In the low-level equilibrium there is again no human capital investment. This arises because Lucas specifies a land-using sector where human capital plays no role, and a "modern" sector where human capital enters with constant returns. Goods production is thus (simplifying slightly):

$$F(x, H, l) = \max_{\theta} \left[x^{\alpha}\theta^{1-\alpha} + BH(l - \theta) \right], \tag{5.4}$$

where x is land per person, H is human capital per person, l is the labor devoted to production, and θ is the labor devoted to the land-using sector. However, the assumption that there is a crucial difference in character between the farm sector and other areas of the economy is unsupportable both for the pre-industrial and for the modern eras. We see above in Table 5.2 that agriculture in England during the Industrial Revolution era experienced unusually fast productivity growth rates also. And agriculture had as much demand for skills and human capital as other sectors of the economy.[16]

In Lucas (2002), parents' utility depends on goods consumption, the number of children, and the utility of the children, but with the slightly different functional form:

$$V_t = c_t^{1-\beta} n_t^{\eta} V_{t+1}^{\beta}. \tag{5.5}$$

Human capital evolves according to:

$$H_{t+1} = H_t \varphi(h_t), \tag{5.6}$$

where h is the labor invested in education. This means that in the Malthusian equilibrium there is no investment in human capital since H starts as 0. Thus, all production is conducted using the land-using technology. Since there is a land constraint, now there will only be a constant output Malthusian equilibrium if $n = 1$, so that the population stabilizes. To ensure this, Lucas assumes that each child requires a fixed investment of *goods*, k. As population increases, so that output per person declines, the relative cost of children thus rises. Eventually, n will be driven to 1.

In the contrasting endogenous growth regime, H is large, so that nearly all output comes from the technology where there are constant returns to H. Consumption and

[16] Hansen and Prescott (2002), is another model which produces an industrial revolution by positing a difference between the farm and non-farm sectors. The inherent rate of productivity growth in the non-farm sector is assumed to be higher. This means that wherever the economy starts there will eventually be an industrial revolution. Why that industrial revolution does not occur in 1800 BC as opposed to 1800 AD is not explained. Also, productivity growth rates in the industrial sector in England in reality increased at the time of the Industrial Revolution. The Industrial Revolution was not the result of composition effects only. And, as noted, productivity growth rates in the farm sector also increased in the Industrial Revolution era, and since then, have been as rapid as those in the rest of the economy.

human capital grow at the same rate, and fertility and educational investment per child is constant. The number of children per parent chosen in this steady-state growth path will depend on the weights in the utility function for children η versus their utilities β, and on the form of $\varphi(h)$.

But like Becker et al. (1990) Lucas gives no mechanism that gets the economy from the Malthusian trap to the sustained growth regime. Instead he has to assume that somehow enough human capital, H, accumulates for non-economic reasons to push the economy far enough from the Malthusian equilibrium for convergence on the modern growth regime to begin. The Industrial Revolution is again the *deus ex machina*.

We therefore see a very poor match between the elements that would seem to go into a human capital story of the Industrial Revolution—the Industrial Revolution itself, the average size of families, and the premium paid in the labor market for skills. If human capital is the key to the Industrial Revolution, the trigger for its expansion in pre-industrial England remains mysterious if we assume a universal set of preferences for all societies.

Endogenous growth theories such as those of Galor and Weil (2000) and Galor and Moav (2002), seek to avoid the need for some exogenous shock to trigger the switch to higher human capital investment and the consequent Industrial Revolution. This requires that some elements of the economy must be evolving endogenously within the pre-industrial era. Since incomes and consumption are predicted to be static within the Malthusian regime, it is not these. Instead, Galor and Weil (2000) rely on the accumulation of population in the pre-industrial era to drive up the rate of innovation and the return to human capital. In this they rely on an interesting paper by Michael Kremer which argues for population size as a driver of rates of productivity advance (Kremer, 1993).

Kremer assumes that the social institutions that provide the incentives to individuals to create knowledge are the same in all societies. Each person has a given probability of producing a new idea. In this case the growth rate of knowledge will be a function of the size of the community. The more people you are in contact with the more you get to benefit from the ideas of others. There was substantial but slow productivity growth in the world economy in the years before 1800, and that all got translated into a huge expansion of the world population, through the effects of Equation (5.2). That larger population produced more ideas and more rapid growth. Sheer scale is what produces modern economic growth.[17]

Kremer supports the argument with two sorts of evidence:

a. The first is population growth rates for the world as a whole in the pre-industrial era. World population growth rates are faster the greater the size of populations. That implies, since population growth rates and the rate of technological advance

[17] Diamond (1997) contains many of the same ideas, merged also with consideration of the role of geography in creating the community that benefits from knowledge expansion.

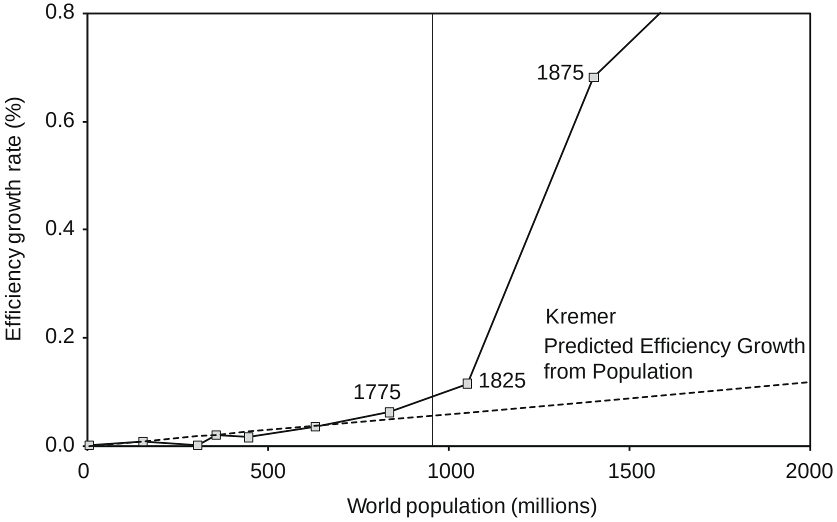


Figure 5.19 Population and the rate of technological advance—actual versus predicted.

are proportionate, that productivity growth rates were speeding up over time as population grew. This is shown in Figure 5.19.

b. The second is population density, as an index of the level of technology in the pre-industrial world, for major isolated geographic areas—Eurasia, the Americas, and Australia—as a function of the land area. The prediction is that the smaller the land area, and hence the potential population, the lower will be the rate of technological advance. In this case, at any given time population density will depend on land area. This is found for the three cases examined.

One immediate implication of the Kremer argument, however, would be that *ceteris paribus*, the Industrial Revolution should have occurred in China. Chinese population in the pre-industrial world was large relative to that of Europe. Even as late at 1800 it has been estimated that China contained 260 million people, while Europe outside Russia had only 130 million, half as many as China. Thus Galor and Weil (2000) have no insight to offer on why the Industrial Revolution was British, as opposed to Chinese. It is a more general theory about the world transition to growth.

Interesting though Kremer's ideas are, no matter how much population is a driver of the rate of technological advance, population alone cannot produce a discontinuity in the rate of technological advance circa 1800, of the magnitude indicated in Figure 5.19. Therefore, a simple specification for the effect of population on changes in productivity would be:

$$\Delta A = \delta N, \tag{5.7}$$

where A is now the stock of knowledge (the number of ideas). If every person has some chance of producing a new idea then the expansion of the idea stock will be at best, proportional to the population size.[18] This implies that the rate of growth of ideas (= productivity) will be:

$$g_A = \frac{\Delta A}{A} = \delta \frac{N}{A}. \tag{5.8}$$

But integrating Equation (5.2) above this is equivalent to the condition:

$$N = \theta A^{1/c}, \tag{5.9}$$

where θ is just a parameter. That is, the population size depends on the existing level of the technology. Substituting from (5.9) for A in (5.8) gives:

$$g_A = cN\left(\frac{\theta}{N}\right)^c = c\theta N^{1-c}. \tag{5.10}$$

This formula implies that the rate of efficiency growth, g_A, rises less than proportionately with population. Yet, what we see in Figure 5.19 is that the rate of technological advance seems to rise faster than population growth. Figure 5.19 also shows the rate of technological advance predicted by this Kremer argument (the lowest curve). The increase of the rate of technological advance as we move to modern population sizes is just not fast enough to explain what we observe.

Technology growth rates would be more responsive to population if instead of (5.8) we posit:

$$\Delta A = \delta NA. \tag{5.11}$$

This says that the stock of ideas grows as a product of the number of people, and the existing stock of ideas (with again no duplication of ideas). This in turn implies that:

$$g_A = \frac{\Delta A}{A} = cN. \tag{5.12}$$

This predicted growth rate of technology as a function of population is also shown in Figure 5.19. Now the fit is closer before 1800, but there is still no close fit with modern productivity growth rates. At best, productivity growth rates would be proportionate to population under the Kremer assumptions.

This feature of the Kremer model, that it is hard to produce with an endogenous growth model, a discontinuity of the magnitude seemingly observed in the Industrial Revolution, is a general problem for all such endogenous growth models. Thus, the Galor and Weil endogenous growth model, which uses the Kremer population size driver as the

[18] Assuming that there is no duplication of ideas with a larger population, where the same thing is discovered by multiple people. In actual fact, we would expect that the gains in idea production would rise less than proportionately with population.

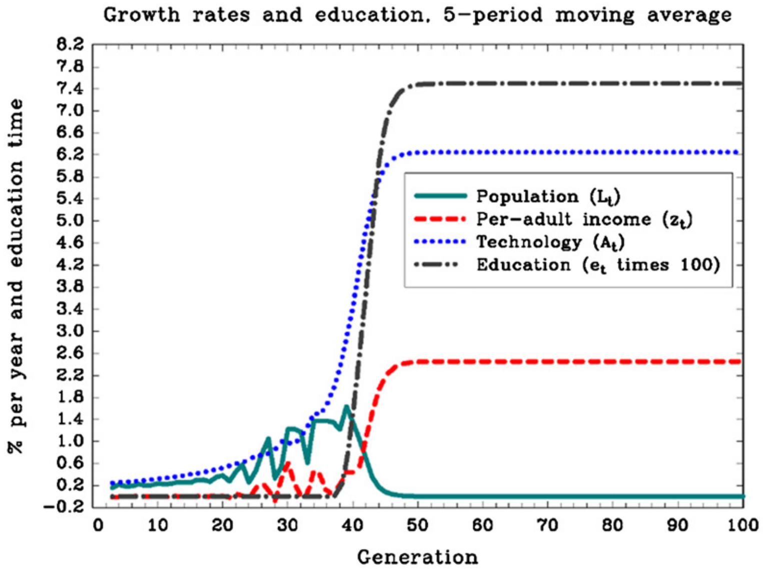


Figure 5.20 Simulation of the Galor and Weil (2000) endogenous growth model. *Source: Lagerlöf (2006), Figure 5, p. 130.*

spark for the Industrial Revolution (and is described further below), has been simulated in Lagerlöf (2006). Figure 5.20 shows the outlines of that simulation, where time is measured on the horizontal axis in terms of generations. In the Galor-Weil model, there is a transition period between the Malthusian regime and modern growth in which technology advances more quickly, incomes rise above subsistence, and population expands. But this transition period here lasts 20 generations, which would be 500–600 years.[19]

But we do not see at the world level in 1200–1800 any signs of the income growth rates, or the population growth rates predicted in this simulation of the Galor-Weil model. Table 5.1, for example, shows that at the world level population growth rates remained in the range of 0.1–0.2% per year, far slower than Figure 5.20 implies. Clark (2007) shows that there is no sign on a world scale that incomes per person had risen above those of the hunter-gatherer era, despite the prediction of Figure 5.20 that by then, incomes per capita in the world would have risen to three times their Malthusian level by 1800. Also, at least in England, we see no sign of the abrupt rise in human capital coincident with declining fertility portrayed in Figure 5.20. Measurements of human capital, as in Figure 5.8, suggest a much more modest transition starting hundreds of years before the Industrial Revolution and continuing through it.

[19] Lagerlöf assumes a generation length of 20 years, but this is too short for any pre-industrial society, where 25–30 years would be more realistic.

Galor and Weil (2000), as noted above, marry the key idea of Kremer that the rate of technological progress depends on population size with the Beckerian human capital approach. They posit a utility function of the form:

$$V_t = c_t^{1-\gamma}(y_{t+1}n_t)^{\gamma}. \tag{5.13}$$

Utility now is a weighted average of the consumption of the parents and the aggregate potential income of their children, y_{t+1}, in the next period. While in the Lucas model children have a fixed cost in goods, in the Galor and Weil model they have a fixed cost only in time. That means that at low incomes, when time is cheap, people would have more children, as in the Becker et al. (1990) model, and we would not get a Malthusian steady state. To get a Malthusian equilibrium where income per capita is stable, the authors make an additional assumption that there is a minimum physical consumption level, $\tilde{c}$. This means that as long as potential income is below some level $\tilde{y}$, increases in income are associated with increases in fertility. As income falls low enough we must reach a state where there is surplus enough beyond $\tilde{c}$ to allow for 1 and only one child per family (treating families as having one parent).[20]

Potential income per worker is of the form:

$$y_t = A_t x_t^{1-\alpha} H_t^{\alpha}, \tag{5.14}$$

where x is land per person, and A is related to the efficiency of goods production. Now human capital is required even in the Malthusian equilibrium. H evolves according to the time invested in educating each child, h, through a function of the form:

$$H_{t+1} = H(h_t, g_{At}), \tag{5.15}$$

where H increases in g_{At}. The TFP variable A evolves according to a function of the form:

$$g_{At} = g(h_t, N_t), \tag{5.16}$$

where N_t is the total population size. Efficiency thus grows more rapidly in large economies with more time resources devoted to each child. And the growth of efficiency increases the human capital per child and the subsequent output per person. Galor and Weil (2000) at least try to preserve some distinction between human capital and the TFP of the economy, but it is not clear whether there is any real substance to the formal mathematical separation. There is no way, observationally, to distinguish

[20] A feature of these theoretical models is that the preferences specified over goods and children in all of them have no function other than allowing the modelers to get the desired outcome in terms of child numbers and human capital in a constrained maximization setting. They do not better explain the world, or offer further insight or predictions about fertility behavior. They are just ways of reproducing, in a desired mathematical format, observed behavior.

economies which have high output because TFP is high, or those that have high output because the human capital stock, as opposed to educational input stock, is large.

The functional form chosen for the utility function is such that the share of time devoted to raising children is always γ once families have achieved the subsistence consumption. Thus there is a built-in trade-off between the quality and quantity of children. Any move to more education must be associated with lower fertility. Therefore, the authors build in an inverse U-shape to fertility as potential incomes rise—with an increase caused by the subsistence constraint on the lower end, and then a decline caused by the rising value of investment in education at higher potential incomes. Again, the utility function here does no real explanatory work. It captures an observed empirical regularity.

The system is constructed so that the amount of time invested in each child increases with the expected rate of technological progress, and the rate of technological progress increases with the time investment per child. At the Malthusian equilibrium the parents spend the minimum possible amount per child, and the only determinant of technological progress is the population size N. By the assumption that g_A is positive, even without any educational investments, population grows in the Malthusian equilibrium, so that the steady-state potential income is maintained by the balance of declining land per person and increasing technological efficiency.

But as population increases, so does the base rate of technological progress, leading parents eventually to invest more than the minimum time in educating their children. At moderate population levels this creates a Malthusian regime with still the minimum consumption per person, but more children each getting some education and a faster rate of technological progress. Eventually population is sufficiently large so that education is productive enough that parents choose fewer high quality children, the population growth rates decline, and potential incomes begin a continuous increase.

Galor and Weil (2000) still face the fundamental problem of the earlier human capital models, however, in that what drives parents to invest more in education in the Industrial Revolution era is a rising perceived return to education. This, as we noted, we do not observe. Nor do we observe any more adverse trade-off between quantity and quality as we move from 1500 to 1920.

Galor and Moav (2002) employ many of the modeling elements of Galor and Weil (2000) except that the Kremer driver for the Industrial Revolution, technological progress being a positive function of population, is replaced. The new driver is a natural selection, either through genes or cultural transmission, of individuals of a certain type in the Malthusian era. Individuals of type i are assumed to choose between consumption, the number of children, and the quality of the children according to a utility function of the form:

$$V_t^i = (c_t^i)^{1-\gamma}(n_t^i(H_{t+1}^i)^{\beta^i})^{\gamma}. \tag{5.16}$$

Now individuals care not about the potential income of their children, but the amount of human capital they possess. The weight individuals give the human capital of their children, indexed by β^i, thus varies with their type. High β families thus produce children with more human capital and more earnings potential. There are assumed, for simplicity, to be just two types of individuals, high β and low β. The potential earnings of each type, y_t^i, are a function of the land-labor ratio, x; the level of technology, A; and human capital, H_t^i, where:

$$y_t^i = A_t x_t^{1-\alpha} (H_t^i)^{\alpha}. \tag{5.17}$$

Now some of the return to education becomes externalized. Low β types gain from the increases in A generated by the investments of the high β types. But the idea is still that once efficiency starts growing more quickly, a given amount of time spent on education produces more human capital. You get more for each year of education. Again this would seem to imply that the wage premium of skilled workers would have to rise in the Industrial Revolution era, which, as noted above, we do not observe.

Again in the Malthusian era, a minimum consumption level, $\tilde{c}$, binds and all gains in potential income go to child rearing. The "high quality" types choose to endow their children with more human capital, however; and this means that they have higher potential incomes in the following period, which results in their descendants having not only higher quality children, but also more children. Thus, the composition of the population changes in the Malthusian period toward individuals with the "high quality" values.[21] This rise in average education inputs increases the private return to education by speeding up the rate of technological advance inducing both high β and low β types to invest in more education and fewer children.

The Galor and Moav (2002) model does have one potentially useful feature, which is that the change in the composition of the population can proceed for generations in a Malthusian state where rates of population growth and levels of income remain low. It would be potentially consistent with the slow rise of education levels in Europe in the 300 years preceding the Industrial Revolution.

The Galor and Moav (2002) model therefore fits the positive association of fertility with wealth and socioeconomic status in pre-industrial England detailed above. However, if we were to elaborate the model to a large number of types we would see that English demography before 1800 is inconsistent with this model. For in Galor and Moav (2002), the positive relation between income and fertility will only be found at lower levels of income close to the consumption constraint, $\tilde{c}$. Once income gets high enough in the pre-industrial period we would see a negative connection between income and fertility, as in the modern era. The highest quality types would die out in the pre-industrial era along with the lowest quality types. Selection in Galor and Moav (2002) is for those whose quality type leads to income just modestly above the subsistence consumption constraint.

[21] Interestingly, the composition of the population in the post-Malthusian period switches back toward the "low quality" types since once potential income for even the low quality types passes a certain boundary they begin to have more children since they invest the same time as the high quality families in child rearing but invest less in each child.

Thus, while the empirical evidence is clear that for at least 500 years before the Industrial Revolution there was differential fertility in England toward those of higher socio economic status, there is no evidence that the selection was of the specific type posited in Galor and Moav (2002). In particular, the evidence is that the quality-quantity trade-off that is central to Galor and Moav (2002), while present, was relatively weak in all periods in England before 1920.

As with the other endogenous growth models, Galor and Moav (2002) would also imply a much slower transition between a world of slow technological advance and the modern era than is observed in practice in the total factor productivity data for England.

5.4. TECHNOLOGICAL CHANGE BEFORE THE INDUSTRIAL REVOLUTION

We have been following the traditional assumption, so far, represented by Figures 5.2 and 5.3, that the Industrial Revolution was a relatively abrupt transition to modern productivity growth rates, around 1780. As Figure 5.2 illustrates, for England as a whole, the efficiency of the economy showed no expansion during 1250–1780. The measured productivity growth rate before the Industrial Revolution is effectively 0. This, as discussed above, makes it seem dauntingly difficult to discern reasons for the transition to rapid economic growth. The underlying institutional, political, and social variables were changing slowly if at all in England in the years 1700–1870 when this transition was accomplished.

The conclusion, from the aggregate productivity level of the economy, that the transition to modern growth was rapid, does however, seem at variance with the general historical picture of England between 1200–1780 as a society that was, over time, advancing in education, in scientific knowledge, in technical abilities in navigation, in warfare, and in technical abilities in music, painting, sculpture, and architecture. England in 1780 was a very different place from England in 1250, even if the standard of living of the average consumer measured mainly in terms of their consumption of food, clothing, housing, heat, and lighting had changed little.

The reason for this mismatch is that, as noted above in Equation (5.3), national productivity growth will be related to productivity advance in individual sectors through:

$$g_A = \sum \theta_j g_{Aj}, \qquad (5.3)$$

where g_{Aj} is the growth rate of productivity by sector, and θ_j is the share of j in total value added in the economy. National efficiency advance is measured by weighting gains by sector with the value of output in that sector. The effects of innovation on national productivity measures are thus crucially dependent on the pattern of consumption.

Much of the technological advance of the period 1250–1780 had minimal impact on measured productivity at the national level because the share of expenditure on these goods was so small in the pre-industrial economy. The printing press, for example, led to

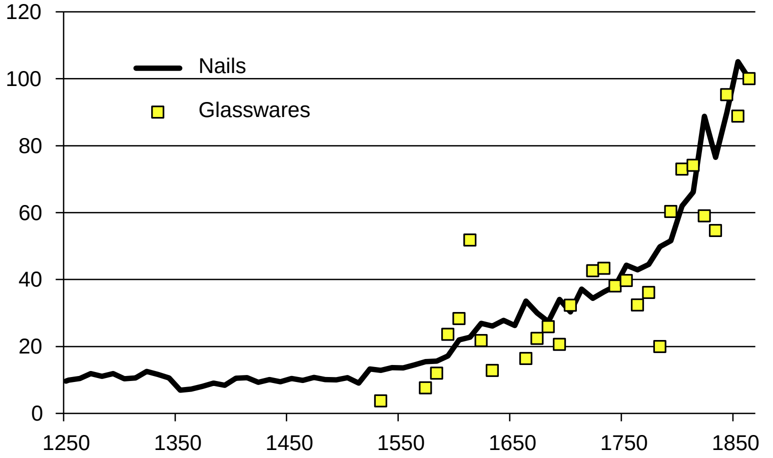


Figure 5.21 Efficiency of production of nails and glassware, by decade, 1250–1869. *Source:* *Clark (2010).*

an approximately 25-fold increase in the production of written material between 1450 and 1600 in England. But since the share of income spent on printed materials in 1600 was only about 0.0005, the productivity gains from this innovation at the national level were miniscule (Clark and Levin, 2001).

We can see also in Figure 5.21 that the production of such manufactured items as iron nails and glassware saw significant productivity advances before 1780. But this efficiency advance would be a negligible contribution to national productivity advance because of the small share of total production value these goods represented in a pre-industrial England where iron nails had limited use, and glasswares were enjoyed only by the richest groups in the society.

Further, for many goods whose production was becoming more efficient through technological advances, no consistent series of prices can be calculated. There was, for example, a great advance in military technologies in European countries such as England over the years 1250–1780. The infantry of 1780, or a naval ship of that period, would have decimated the equivalent medieval force. English troops of 1780 would have quickly overwhelmed the fortifications of 1250, and the fortifications of 1780 would have been impregnable against medieval armies of major size. But none of this would be reflected in conventional productivity measures. There is no allowance in these measures for the delivery of more effective violence by the English Navy over the years.

There is no allowance also in the national productivity measure for improvements in the quality of literature, music, painting, and newspapers. These sources also do not reflect medical advances such as the one third reduction in maternal childbirth mortality between 1600 and 1750.[22]

[22] Wrigley et al. (1997, p. 313).

This makes it possible that the rate of technological advance in the economy, measured just as a count of innovations and new ideas, was actually increasing long before the breakthrough of the Industrial Revolution. But accidents of where these technological advances came in relation to mass consumer demand in the pre-industrial economy create the appearance of a technological discontinuity circa 1780. Suppose that prior to the Industrial Revolution innovations were occurring randomly across various sectors of the economy—innovations in areas such as guns, gunpowder, spectacles, window glass, books, clocks, painting, new building techniques, improvements in shipping and navigation—but that just by chance, all these innovations occurred in areas of small expenditure. Then the technological dynamism of the economy would not show up in terms of output per capita or in measured productivity in the years leading up to the Industrial Revolution.

To illustrate this, suppose we consider a consumer whose tastes were close to those of the modern university professor. Their consumption is much more heavily geared toward printed material, paper, spices, wine, sugar, manufactured goods, light, soap, and clothing than the average consumer in the pre-industrial English economy. Based on their consumption, how would the efficiency growth rate of the economy 1250–1769 look compared to 1760–1869, and 1860–2009? Figure 5.22 shows the results, where efficiency is measured as an index on a log scale on the vertical axis. Thus, the slope of the lines indicates the rate of efficiency growth, or efficiency decline, in each era. Now in the years 1300–1770 there is an estimated efficiency growth rate of 0.09% per year for the goods consumed by a university professor. This is followed by efficiency growth rates of 0.6%

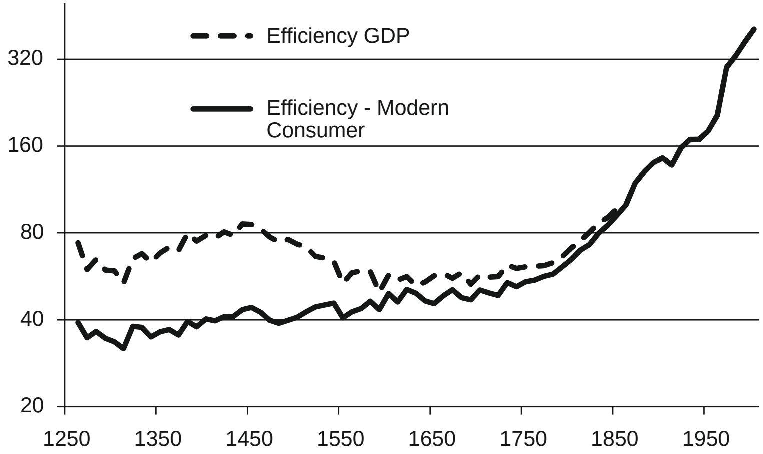


Figure 5.22 Economic efficiency from the perspective of a modern consumer, England, 1250–2009.
Notes: The weights in consumption for the modern consumer are assumed to be half from the consumption basket of the pre-industrial worker. But the other half is composed of books (.1), manufactured goods (.1), clothing (.1), sugar (.03), spices (.03), drink (.05), light (.05), soap (.02), and paper (.02).
Source: Clark (2010).

per year 1760–1870, and 0.9% a year for 1860–2010. Estimated efficiency advance is still very slow for the pre-industrial period, but we can think of the economy in this period as going through a more protracted transition between pre-industrial growth rates and modern growth rates.

Framed in this way, the possibility reopens of some variety of endogenous growth explanation of the Industrial Revolution, with a more gradual transition to higher rates of technological advance starting in the medieval period or earlier. However, the existing endogenous growth models such as Galor and Weil (2000) and Galor and Moav (2002), bring with them a set of assumptions and implications which are difficult to reconcile with empirical reality, as we have discussed above. The key idea in Galor and Moav (2002), however, that in the Malthusian regime preferences might be changed by differential net fertility, does seem to offer some promise. We do see strong differences in fertility by social class in England all the way from 1250 to 1780. And there is evidence that parental characteristics in terms of wealth, occupation, and education were very strongly inherited in pre-industrial England, allowing differential fertility to have significant effects on the characteristics of the population even after relatively few generations.[23] While we do not see a sign in the data of the specific selection for a preference for small family sizes and high child quality, there is a sign of a more generalized selection for characteristics associated with economic success.

5.5. CONCLUSION

The Industrial Revolution remains one of histories great mysteries. We have seen in this survey that the attempts by economists to model this transition have been so far largely unsuccessful. The first approach emphasizing an exogenous switch in property rights stemming from political changes, despite its continuing popularity, fails in terms of the timing of political changes, and their observed effects on the incentives for innovation. The second approach, which looks for a shift between self-reinforcing equilibria, again fails because there is little sign of any major changes in the underlying parameters of the economy circa 1780 which would lead to changed behavior by individuals. The most promising class of models are those based on endogenous growth. The problem here is finding some kind of "driver" that is changing over time that will induce changes in the rate of innovation. Previously, these models seemed to face insuperable difficulties in that they find it very hard to model the kind of one-time upward shift in productivity growth rates that the Industrial Revolution seemed to involve. But as we gather more information on the empirics of the Industrial Revolution, and the years before, the discontinuity in technological innovation rates seems less than has been imagined, and the transition between

[23] As evidenced by the persistence of status of surnames in England 1300–2012, the correlation of underlying social status between fathers and sons seems always to have been of the order of 0.75, which is very high. See Clark et al. (2014).

the old world of zero productivity growth rates and the new world of rapid productivity growth seems much more gradual. This bodes well for endogenous growth models.

REFERENCES

Acemoglu, Daron, Robinson, James A., Johnson, Simon, 2001. The colonial origins of comparative economic development: an empirical investigation. American Economic Review 91, 1369–1401.

Acemoglu, Daron, Robinson, James A., Johnson, Simon, 2002. Reversal of fortune: geography and institutions in the making of the modern world. Quarterly Journal of Economics 117, 1231–1294.

Acemoglu, Daron, Robinson, James A., 2012. Why Nations Fail: The Origins of Power, Prosperity, and Poverty. Crown Business, New York.

Angrist, Joshua, Lavy, Victor, Schlosser, Analia, 2010. Multiple experiments for the causal link between the quantity and quality of children. Journal of Labor Economics 28 (4), 773–824.

Becker, Gary, Murphy, Kevin, Tamura, Robert, 1990. Human capital, fertility and economic growth. Journal of Political Economy 98, S12–37.

Boberg-Fazlic, Nina, Sharp, Paul, Weisdorf, Jacob, 2011. Survival of the richest? Testing the Clark hypothesis using English pre-industrial data from family reconstitution records. European Review of Economic History 15 (3), 365–392.

Bresnahan, Timothy F., Trajtenberg, Manuel, 1996. General purpose technologies: engines of growth? Journal of Econometrics, Annals of Econometrics 65, 83–108.

Clark, Gregory, 1996. The political foundations of modern economic growth: England, 1540–1800. Journal of Interdisciplinary History 26 (4), 563–588.

Clark, Gregory, 2005. The condition of the working-class in England, 1209–2004. Journal of Political Economy 113 (6), 1307–1340.

Clark, Gregory, Hamilton, Gillian, 2006. Survival of the richest. The Malthusian mechanism in pre-industrial England. Journal of Economic History 66 (3), 707–736.

Clark, Gregory, Jacks, David, 2007. Coal and the industrial revolution, 1700–1869. European Review of Economic History 11 (1), 39–72.

Clark, Gregory, 2007. A Farewell to Alms: A Brief Economic History of the World. Princeton University Press, Princeton.

Clark, Gregory, 2010. The macroeconomic aggregates for England, 1209–2008. Research in Economic History 27, 51–140.

Clark, Gregory, Levin, Patricia. 2001. How Different Was the Industrial Revolution? The Revolution in Printing, Working Paper, University of California, Davis, pp. 1350–1869.

Clark, Gregory, Cummins, Neil et. al. 2014. The Son Also Rises: Surnames and the History of Social Mobility, Princeton University Press, Princeton.

Clark, Gregory, Cummins, Neil, 2013a. Malthus to Modernity: England's First Fertility Transition, 1500–1880. Working Paper, UC Davis.

Clark, Gregory, Cummins, Neil, 2013b. The Beckerian Family and the English Demographic Revolution of 1800. Working Paper, UC Davis.

Cressy, David, 1977. Levels of illiteracy in England, 1530–1730. Historical Journal 20, 1–23.

Diamond, Jared M., 1997. Guns, Germs, and Steel: The Fates of Human Societies. W.W. Norton, New York.

Duncan-Jones, Richard, 1990. Structure and Scale in the Roman Economy. Cambridge University Press, Cambridge.

Fitton, 1989. The Arkwrights: Spinners of Fortune. Manchester University Press, Manchester.

Galor, Oded, Weil, David N., 2000. Population, technology and growth: from Malthusian stagnation to the demographic transition and beyond. American Economic Review 90, 806–828.

Galor, Oded, Moav, Omer, 2002. Natural selection and the origin of economic growth. Quarterly Journal of Economics 117, 1133–1191.

Galor, Oded, 2011. Unified Growth Theory. Princeton University Press, Princeton.

Greif, Avner, 2006. Institutions and the Path to the Modern Economy: Lessons from Medieval Trade. Cambridge University Press, Cambridge.

Greif, Avner, Iyigun, Murat, Sasson, Diego L., 2012. Social Institutions and Economic Growth: Why England and Not China Became the First Modern Economy. Working Paper.

Hansen, G., Prescott, Edward C., 2002. Malthus to solow. American Economic Review 92 (4), 1205–1217.

Harley, Knick, 1998. Cotton textile prices and the industrial revolution. Economic History Review 51 (1), 49–83.

Harley, C. Knick, 2010. Prices and Profits in Cotton Textiles during the Industrial Revolution. University of Oxford Discussion Papers in Economic History, #81.

Hopkins, Keith, 1966. On the probable age structure of the Roman population. Population Studies 20 (2), 245–264.

Houston, R.A., 1982. The development of literacy: Northern England, 1640–1750. Economic History Review, New Series, 35 (2), 199–216.

Jones, Charles I., 2002. Introduction to Economic Growth, second ed. W.W. Norton, New York.

Kremer, Michael, 1993. Population growth and technological change: one million B.C. to 1990. Quarterly Journal of Economics 107, 681–716.

Lagerlöf, Nils-Petter, 2006. The Galor–Weil model revisited: a quantitative exercise. Review of Economic Dynamics 9 (1), 116–142.

Li, Hongbin, Zhang, Junsen, Zhu, Yi, 2008. The quantity-quality trade-off of children in a developing country: identification using Chinese twins. Demography 45 (1), 223–243.

Lindberg, Erik, 2009. Club goods and inefficient institutions: why Danzig and Lübeck failed in the early modern period. Economic History Review, New Series 62(3), 604–628.

Long, Pamela, 1991. Invention, authorship, intellectual property, and the origin of patents: notes towards a conceptual history. Technology and Culture 32, 846–884.

Lucas, Robert, 1988. On the mechanics of economic development. Journal of Monetary Economics 22, 3–42.

Lucas, Robert E., 2002. The industrial revolution: past and future. In: Lucas, Robert E. (Ed.), Lectures on Economic Growth. Harvard University Press, Cambridge.

Mitchell, Brian R., 1988. British Historical Statistics. Cambridge University Press, Cambridge.

Mitchell, Brian R., Deane, Phyllis, 1971. Abstract of British Historical Statistics. Cambridge University Press, Cambridge.

North, Douglass C., 1994. Economic performance through time. American Economic Review 84 (3), 359–368.

North, Douglass, Thomas, Robert P., 1973. The Rise of the Western World. Cambridge University Press, Cambridge.

North, Douglass, Weingast, Barry, 1989. Constitutions and commitment: evolution of institutions governing public choice in 17th century England. Journal of Economic History 49, 803–832.

Romer, Paul M., 1986. Increasing returns and long-run growth. Journal of Political Economy 94, 1002–1037.

Romer, Paul M. 1987. Crazy Explanations for the Productivity Slowdown. In: Fischer, Stanley (Ed.), NBER Macroeconomics Annual 1987, MIT Press, Cambridge, Mass.

Romer, Paul M., 1990. Endogenous technological change. Journal of Political Economy 98, S71–102.

Rosenzweig, Mark R., Wolpin, Kenneth I. 1980. Testing the quantity-quality fertility model: the use of twins as a natural experiment. Econometrica 48 (1), 227–240.

Rubinstein, William D., 1981. Men of Property: The Very Wealthy in Britain Since the Industrial Revolution. Croom Helm, London.

Schofield, Roger, 1973. Dimensions of illiteracy, 1750–1850. Explorations in Economic History 10, 437–454.

Schultz, T. Paul. 2007. Population Policies, Fertility, Women's Human Capital, and Child Quality. Economic Growth Center Yale University, Discussion Paper No. 954.

Wrigley, E.A., Davies, R.S., Oeppen, J.E., Schofield, R.S., 1997. English Population History from Family Reconstruction: 1580–1837. Cambridge University Press, Cambridge, New York.

CHAPTER SIX

Twentieth Century Growth

Nicholas Crafts[*] and Kevin Hjortshøj O'Rourke[†]
[*]Department of Economics, University of Warwick, Coventry CV4 7AL, United Kingdom
[†]All Souls College,Oxford OX1 4AL, United Kingdom

Abstract

This paper surveys the experience of economic growth in the 20th century with a focus on technological change at the frontier together with issues related to success and failure in catch-up growth. A detailed account of growth performance based on historical national accounts data is given and is accompanied by a review of growth accounting evidence on the sources of economic growth. The key features of our analysis of divergence in growth outcomes are an emphasis on the importance of "directed" technical change, of institutional quality, and of geography. We provide brief case studies of the experience of individual countries to illustrate these points.

Keywords

Catch-up growth, Divergence, Growth accounting, Technical change

JEL Classification Codes

N10, O33, O43, O47

6.1. INTRODUCTION

This chapter does not pretend to provide a comprehensive survey of the vast literature that has been written on economic growth during the 20th century: for such a task, not even a book would suffice. Rather, it is a brief interpretative essay, which aims to place the 20th century growth experience into a broader historical context, and highlight some of the ways in which the field of economic history can contribute to the study of economic growth.

A theme of the chapter is that the 20th century saw the gradual working out of several long-run implications of the Industrial Revolution: the latter was a massive asymmetric shock to the world economy, which set in train a variety of long-run adjustment processes which are still ongoing, and which seem set to define the economic history of the 21st century as well.

We will be emphasizing two key features of the economic history literature. The first is a focus on institutions, following the insights of North (1990) and others. While institutions have certainly become a central focus of mainstream empirical work on

ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53538-2.00006-X

economic growth (e.g. Acemoglu et al. 2001), economic historians tend to be quite nuanced in their view of how institutions matter, recognizing that different institutional environments may be appropriate at different points in time and in different countries.

The second is a detailed interest in the mechanics of technological change. The endogenous nature of technological change, and the consequences which this has for economic growth in both leader and follower countries, will be a constant theme of the chapter: while theorists like Acemoglu (2002) have recently brought the issue to the forefront of growth theory, economic historians such as Habakkuk (1962) have been emphasizing such themes for many decades.

6.2. SETTING THE STAGE

In this section, we look at the legacy of the Industrial Revolution and its 19th century aftermath. This period saw the advent of modern economic growth (Kuznets, 1966) in what came to be the advanced economies of the 20th century, along with a big shift in the center of gravity of the world economy away from Asia and toward Europe and North America. The world economy of 1900 was hugely different from that of 1700 in terms of its technological capabilities, the income levels in leading economies, the extent of globalization, and the degree of international specialization in production.

6.2.1 The Beginnings of Modern Economic Growth

Recent research has made considerable progress in quantifying growth in the world economy prior to the Industrial Revolution. Table 6.1 reports estimates of income levels measured in purchasing-power-parity adjusted to 1990 international dollars for selected countries. In this metric, it is generally agreed that a bare-bones subsistence income is about $400 per year. The estimates indicate that European countries had incomes well

Table 6.1 GDP per capita, 1086–1850, adjusted to 1990 international dollars

	England/Great Britain	Holland/ Netherlands	Italy	Spain	China	India	Japan
1086	754				1244		
1348	777	876	1376	1030			
1400	1090	1245	1601	885	948		
1500	1114	1483	1403	889	909		
1600	1123	2372	1244	944	852	682	791
1650	1100	2171	1271	820		638	838
1700	1630/1563	2403	1350	880	843	622	879
1750	1710	2440	1403	910	737	573	818
1800	2080	2617/1752	1244	962	639	569	876
1850	2997	2397	1350	1144	600	556	933

Source: Broadberry (2013).

above this level long before the Industrial Revolution, and the same was true of China in medieval times. The implication is that the pre-industrial era should not be seen as one in which people were in a very low income Malthusian Trap equilibrium.

Nevertheless, the overall picture of Table 6.1 is that growth was at best very slow in these pre-industrial centuries. Growth of real income per person averaged 0.2% per year in England, a relative success story, between 1270 and 1700 (Broadberry et al. 2010) while at the other extreme, Chinese income levels almost halved between 1086 and 1800. These estimates re-assert the traditional story of the "Great Divergence", namely, that the most successful parts of Europe overtook China and pulled significantly ahead in the run-up to the Industrial Revolution. They also reflect a "Little Divergence" within Europe between North and South, with Italy and Spain losing out relative to England and Holland.

What were the underpinnings of this modest pre-industrial growth in England? The answer seems to be a combination of increases in hours worked per person and Smithian growth, rather than any major contribution from technological change. The length of the work year may have roughly doubled between the mid-14th and late-18th century (Allen and Weisdorf, 2011). This largely accounts for the long-run tendency for income per person to grow slowly, despite the fact that the null hypothesis that real wage rates were stationary until 1800 cannot be rejected (Crafts and Mills, 2009). Growth in the successful parts of Europe was also strongly correlated with trade expansion. This improved productivity and sustained wage levels in the face of demographic pressure (Allen, 2009).

The term "Industrial Revolution" is commonly used to characterize the unprecedented experience of the British economy during the later decades of the 18th and early decades of the 19th century. Taken literally, it is a misleading phrase; but carefully deployed, it is a useful metaphor. These years saw a remarkable economic achievement by comparison with earlier times, but it must be recognized that by later standards this was in many ways a modest beginning.

The idea of an industrial revolution conjures up images of spectacular technological breakthroughs; the triumph of the factory system and steam power; the industrialization of an economy hitherto based largely on agriculture, and rapid economic growth. Indeed, these were the directions of travel for the British economy but when they are quantified, the numbers although impressive, once put into context, do not live up to the hyperbole. While the economy withstood formidable demographic pressure much better than could have been imagined in the 17th century, the growth of real income per person was painfully slow for several decades. Not much more than a third of the labor force worked in agriculture even in the mid-18th century. In 1851, more people were employed in domestic service and distribution than in textiles, metals, and machine-making combined. Until around 1830, water power was more important than steam power in British industry.

Nevertheless, the economy of the mid-19th century was established on a different trajectory from that of a hundred years earlier. In particular, sustained labor productivity growth based on steady technological progress and higher levels of investment had become the basis of significant growth in real income per person, notwithstanding rapid population growth. This was modern economic growth, as distinct from real income increases based on Smithian growth and working harder. That said, growth potential was still quite limited by 20th century standards: education and scientific capabilities were still quite primitive, the scope to import technological advances from the rest of the world was modest, and institutions and economic policies suffered from obvious limitations.

Table 6.2 reports that the rate of TFP growth more than doubled from 0.3% per year in 1760–1801 to 0.7% per year in 1831–1873. This can certainly be interpreted as reflecting acceleration in the rate of technological progress but TFP growth captures more than this.

Table 6.2 Growth-accounting estimates (percent per annum)

(a) Output growth

	Capital contribution	Labor contribution	TFP growth	GDP growth
1760–1801	0.4*1.0 = 0.4	0.6*0.8 = 0.5	0.3	1.2
1801–1831	0.4*1.7 = 0.7	0.6*1.4 = 0.8	0.2	1.7
1831–1873	0.4*2.3 = 0.9	0.6*1.3 = 0.8	0.7	2.4

(b) Labor productivity growth

	Capital-deepening contribution	TFP growth	Labor productivity
1760–1801	0.4*0.2 = 0.1	0.3	0.3
1801–1831	0.4*0.3 = 0.1	0.2	0.3
1831–1873	0.4*1.0 = 0.4	0.7	1.1

(c) Contributions to labor productivity growth, 1780–1860

Capital deepening	0.22
Modernized sectors	0.12
Other sectors	0.1
TFP growth	0.42
Modernized sectors	0.34
Other sectors	0.08
Labor productivity growth	0.64
Memorandum items	
Labor force growth	1.22
Capital income share (% of GDP)	40
Modernized sectors	5.9

Notes: Growth accounting imposes the standard neoclassical formula in parts (a) and (b). To allow for embodiment effects in part (c) the standard growth-accounting equation is modified as follows to distinguish between different types of capital and different sectors: $\Delta\ln(Y/L) = \alpha_O \Delta\ln(K_O/L) + \alpha_M \Delta\ln(K_M/L) + \gamma \Delta\ln A_O + \Phi \Delta\ln A_M$, where the subscripts O and M denote capital in the old and modernized sectors, respectively; γ and Φ are the gross output shares of these sectors; and α_O and α_M are the factor shares of the capital used in these sectors.
Sources: Crafts (2004a, 2005) revised to incorporate new output growth estimates from Broadberry et al. (2010).

No explicit allowance has been made for human capital or hours worked in the growth accounting equation. Prior to 1830, it is generally agreed that any contribution from extra schooling or improved literacy was negligible, but in the period 1831–1873 education may have accounted for around 0.3 percentage points per year of the measured TFP growth in Table 6.2 (Mitch, 1999). For 1760–1801 there is good reason to think that average hours worked per worker per year were increasing sufficiently that if the growth in labor inputs were adjusted appropriately TFP growth might be pushed down very close to zero (Voth, 2001). Overall then, a best guess might be that the contribution of technological progress, as reflected in TFP growth, went from about zero to a sustained rate of about 0.4% per year by the time the classic Industrial Revolution period was completed.

Neoclassical growth accounting of this kind is a standard technique and valuable for benchmarking purposes, if nothing else. However, it does potentially underestimate the contribution of new technology to economic growth if technological progress is embodied in new types of capital goods, as was set out in detail by Barro (1999). This was surely the case during the Industrial Revolution; as Feinstein put it, "many forms of technological advance . . . can only take place when "embodied" in new capital goods. The spinning jennies, steam engines, and blast furnaces were the "embodiment" of the Industrial Revolution" (1981, p. 142).

Table 6.2 also shows the results of an exercise that allows for embodiment effects. The "modernized sectors" (cottons, woolens, iron, canals, ships, and railways) are found to have contributed 0.46 out of 0.64% per year growth in labor productivity over the period 1780–1860 with the majority of this, 0.34 compared with 0.12%, coming from TFP growth as opposed to capital deepening. If the contribution of technological change to the growth of labor productivity is taken to be capital deepening in the modernized sectors plus total TFP growth, then this equates to 0.54 out of 0.64% per year. It remains perfectly reasonable, therefore, to regard technological innovation as responsible for the acceleration in labor productivity growth that marked the Industrial Revolution as the historical discontinuity that Kuznets supposed, even though the change was less dramatic than was once thought.

It may seem surprising that the Industrial Revolution delivered such a modest rate of technological progress given the inventions for which it is famous, including most obviously those related to the arrival of steam as a general purpose technology (GPT). It should be noted, however, that the well-known stagnation of real wage rates during this period is strong corroborative evidence that TFP growth, which is equal to the weighted average of growth in factor rewards (Barro, 1999), was modest.

Two points can be made straightaway. First, the impact of technological progress was very uneven as is implied by the estimates in Table 6.2. Most of the service sector other than transport was largely unaffected. Textiles, metals, and machine-making accounted for less than a third of industrial employment—or 13.4% of total employment—even in 1851 and much industrial employment was still in "traditional" sectors. Second, the process of technological advance was characterized by many incremental improvements and learning to realize the potential of the original inventions. This took time in an era where scientific and technological capabilities were still very weak by later standards.

Steam power offers an excellent example. In 1830, only about 165,000 horsepower was in use, the steam engine capital share was 0.4% of GDP and the Domar weight for steam engines was 1.7% (Crafts, 2004b). The cost effectiveness and diffusion of steam power was held back by the high coal consumption of the original low-pressure engines and the move to high pressure—which benefited not only factories but railways and steam ships—was not generally accomplished until the second half of the 19th century. The science of the steam engine was not well understood and the price of steam power fell very slowly, especially before about 1850. The maximum impact of steam power on British productivity growth was delayed until the third quarter of the 19th century—nearly 100 years after James Watt's patent—when it contributed about 0.4% per year to labor productivity growth. It seems reasonable to conclude that subsequently leading economies have become much better at exploiting GPTs. The reasons are likely to be found in a superior level of education and scientific knowledge; improvements in capital markets; government policies that support research and development; and thus a greater volume of and higher expected returns to innovative effort.

Indeed, from an endogenous growth perspective the early 19th century British economy still had many weaknesses. The size of markets was still very small in 1820, when modern globalization was yet to begin (O'Rourke and Williamson, 2002), and real GDP in Britain was only about one twentieth of its size in the United States a century later. The costs of invention were high at a time when scientific knowledge and formal education could still make only a modest contribution. This was clearly not a time of high college enrollment, and the highly educated were to be found in the old professions, not science and engineering. Investment, especially in equipment, was a small proportion of GDP. Intellectual property rights were weak since the legal protection offered by patents was doubtful until the 1830s, and even if Britain had less rent-seeking than France, rent-seeking in the law, the bureaucracy, the church, and the military remained very attractive alternatives to entrepreneurship, as is attested by the evidence on fortunes bequeathed (Rubinstein, 1992). Accordingly, TFP growth was modest, although by the 1830s it was still well ahead of the rate achieved in the United States, which averaged 0.2% per year during 1800–1855 (Abramovitz and David, 2001).

6.2.2 Directed Technical Change and the First Industrial Revolution[1]

If the transition to modern economic growth entails a sustained acceleration in the rate of technological progress, why did this happen first in Britain in the late 18th century? Over time many answers have been suggested, but a recent interpretation by Allen, building

[1] This section draws in part on Crafts (2011).

on Habakkuk (1962) and David (1975), has rapidly gained currency. His conclusion is deceptively simple: "The Industrial Revolution . . . was invented in Britain because it paid to invent it there" (Allen, 2009, p. 2). Allen's argument comes from an endogenous innovation perspective but is based on relative factor prices and market size rather than on the superiority of British institutions and policies, at least compared with its European peer group: that is to say, it focuses on the demand for innovation, rather than on the supply side. In particular, Britain's unique combination of high wages and cheap energy plus a sizeable market for the new technologies, which were profitable to adopt only in these circumstances, is held to be the key.

Allen's analysis emphasizes the importance of expected profitability to justify the substantial fixed costs of the investment required to perfect good ideas and make them commercially viable. The rate of return on adopting inventions in textiles, steam power, and coke smelting was a lot higher in Britain than elsewhere and so the potential market for these inventions was much greater. This is very similar to the model of "directed technical change" proposed by Acemoglu (2002).[2] Allen supports his conclusions by empirical analysis of the profitability of adoption of several famous inventions (Hargreaves' spinning jenny, Arkwright's mill, and coke smelting) at British and French relative factor prices. The conclusion is that in each case, adoption would have been rational at the former but not the latter. Eventually, after several decades, a cumulative process of micro-invention had improved these technologies to the point where adoption became profitable in other countries, and the Industrial Revolution began to spread.

Allen's hypothesis is prima facie plausible and theoretically defensible although more research is required to establish that it stands on really solid empirical foundations. For example, Crafts (2011) presents evidence suggesting that it may have been high machinery costs, rather than low wages, which impeded the adoption of the spinning jenny in France. Strikingly, it also appears that it would have been very profitable to invent and adopt the jenny in the high-wage United States.[3] Perhaps the key disincentive there was small market size relative to the fixed development costs of the invention. There are also a number of other detailed issues about the robustness of Allen's calculations that have arisen in the debate prompted by his book.[4] Allen himself recognizes that the supply side of the market for innovation mattered as well as the demand side: to claim that relative factor prices alone were the key to the Industrial Revolution would be a bit too

[2] Acemoglu (2010) extended this analysis to consider the impact of labor scarcity on the rate, rather than the bias, of technological progress and showed that this is positive if technological change is strongly labor saving, i.e. reduces the marginal product of labor. This might be when machines replace tasks previously undertaken by workers, as in Zeira (1998).

[3] This also seems to be true of the Arkwright mill where the prospective rate of return to adoption was 32.5% (Crafts, 2011).

[4] See the further discussion in Crafts (2011) and the interchanges between Gragnolati et al. (2011) and Allen (2011); and between Humphries (2013) and Allen (2013).

bold. Even so, Allen's contribution has been extremely valuable in focusing attention on the incentives facing innovators. In the context of subsequent British relative economic decline and, especially, American overtaking, his suggestion that the key to getting ahead in the Industrial Revolution was relative factor prices together with large market size has the clear implication that British leadership would be highly vulnerable. Insofar as high wages, cheap energy, and a market sufficient to allow fixed costs of research and development continued to be conducive to faster technological progress, the United States would be a more favored location later in the 19th century, as has become abundantly clear in the literature on the Habakkuk (1962) hypothesis.

6.2.3 Catch-Up and Overtaking: The Transition to American Leadership

By the late 19th century, as Table 6.3 reports, modern economic growth had spread to most of Western Europe. Rates of growth of real GDP per person, although modest by later standards, were generally well above those achieved by Britain during the Industrial Revolution (0.4% per year) and during the second and third quarters of the 19th century (1% per year). Faster growth often went hand in hand with industrialization, and there was a clear but not perfect correlation in 1913 between industrial output and GDP per head. The United Kingdom remained the European leader in 1913 but the rest of Europe was slowly catching up, and by the end of the 19th century Britain had been overtaken by the United States. The hypothesis of unconditional convergence across Europe during 1870–1913 is rejected, however (Crafts and Toniolo, 2008). Southern Europe clearly lagged behind northern Europe while nevertheless opening up a substantial gap with China.

Table 6.4 shows that crude TFP growth remained quite slow until it increased appreciably in several countries at the end of the 19th century, around the time of the so-called Second Industrial Revolution. Even so, nowhere in Europe was there a growth experience that resembled the picture famously drawn by Solow (1957) for the United States in the first half of the 20th century in which the residual accounted for seven eighths of labor productivity growth. For almost all countries, technical change came primarily from the diffusion of advances made elsewhere, but technological diffusion was still relatively slow.[5]

It is not possible to implement a full analysis of conditional convergence given data limitations but Table 6.5 offers some clues. Years of schooling increased everywhere but were generally much higher in northern Europe and, by 1913, were way ahead of the 2.3 years of the cohort born before 1805 in England and Wales (Matthews et al. 1982). In the period before World War I, industry was attracted to market potential and cheap

[5] Germany and the UK together accounted for 53% of all foreign patents taken out in the United States in 1883 and 57% in 1913 (Pavitt and Soete, 1982). The diffusion rate of inventions made before 1925 was less than a third of those made subsequently (Comin et al. 2006).

Table 6.3 Growth in late nineteenth-century Western Europe

	1870 GDP/capita ($1990GK)	1913 GDP/capita ($1990GK)	Growth, 1870–1913 (% p.a.)	Industrialization level, 1870	Industrialization level, 1913
Austria	1863	3465	1.46	13	32
Belgium	2692	4220	1.05	36	88
Denmark	2003	3912	1.58	11	33
Finland	1140	2111	1.45	13	21
France	1876	3485	1.46	24	59
Germany	1839	3648	1.61	20	85
Greece	880	1592	1.39	6	10
Ireland	1775	2736	1.01		
Italy	1499	2564	1.26	11	26
Netherlands	2757	4049	0.91	12	28
Norway	1360	2447	1.38	14	31
Portugal	975	1250	0.59	9	14
Spain	1207	2056	1.25	12	22
Sweden	1359	3073	1.92	20	67
Switzerland	2102	4266	1.67	32	87
UK	3190	4921	1.01	76	115
Europe	1971	3437	1.31	20	45
Aide Memoire					
United States	2445	5301	1.83	30	126
China	530	552	0.1	4	3

Note: Industrialization level is defined as an index of the volume of industrial output/person relative to a base of UK in 1900 = 100.
Sources: Maddison (2010) and Bairoch (1982).

coal (Crafts and Mulatu, 2006; Klein and Crafts, 2012) which again favored the north over the south. Institutions improved with regard to underpinning the appropriability of returns to investment, especially in northern Europe, as reflected in the Political Constraint Index which Henisz (2002) shows was positively related to private sector investment in infrastructure. There was a widespread improvement in legislation enabling capital markets to function (Bogart et al. 2010). Even so, a recent study (Kishtainy, 2011) suggests that only Switzerland (after 1848) and Norway (after 1899) could be classified as "open-access" societies with the political and economic competition that is regarded as essential to becoming an advanced economy by North et al. (2009). Nevertheless, much of Europe was on the verge of attaining that open-access status and this contrasts starkly with the continuation of a closed-access society dominated by a coalition of rent-seekers that stifled innovation in China (Brandt et al. forthcoming).

In growth accounting terms, as Table 6.4 shows, American overtaking was associated with a late 19th century acceleration in the rate of TFP growth to a pace far in excess

Table 6.4 Accounting for labor productivity growth (percent per annum)

	Labor productivity growth	**Capital deepening contribution**	**TFP growth**
Austria			
1870–1890	0.9	0.64	0.26
1890–1910	1.69	0.66	1.03
Germany			
1871–1891	1.1	0.39	0.71
1891–1911	1.76	0.58	1.18
Netherlands			
1850–1870	1.02	0.5	0.52
1870–1890	0.94	0.61	0.33
1890–1913	1.35	0.46	0.89
Spain			
1850–1883	1.2	1	0.2
1884–1920	1	0.7	0.3
Sweden			
1850–1890	1.18	1.12	0.06
1890–1913	2.77	0.94	1.83
United Kingdom			
1873–1899	1.2	0.4	0.8
1899–1913	0.5	0.4	0.1
United States			
1855–1890	1.1	0.7	0.4
1890–1905	1.9	0.5	1.4
1905–1927	2	0.5	1.3

Note: All estimates impose a standard neoclassical growth accounting equation based on $Y = AK^{\alpha}L^{1-\alpha}$, calibrated with $\alpha = 0.35$.

Sources: Derived from data presented in the following original growth accounting studies: Austria: Schulze (2007); Germany: Broadberry (1998); Netherlands: Albers and Groote (1996); Spain: Prados de la Escosura and Roses (2009); Sweden: Krantz and Schön (2007); United Kingdom: Feinstein et al. (1982); United States: Abramovitz and David (2001).

of that achieved during the Industrial Revolution.[6] The United Kingdom did not match this acceleration. The origins of faster technological change in the United States may well be along the lines of Habakkuk (1962). He famously claimed that land abundance and labor scarcity in the United States promoted rapid, labor-saving technological change. New economic historians spent quite a long time trying to pin down these arguments. Eventually, it was found that the US was able to exploit complementarities between capital and natural resources to economize on the use of skilled labor in an important

[6] These estimates take no account of education but this would not make much difference according to Abramovitz and David (2001) who found that adjusting TFP growth on this account would reduce TFP growth by 0.0%, 0.1%, and 0.2% per annum in 1855–1890, 1890–1905 and 1905–1927, respectively.

Table 6.5 Variables relating to conditional convergence

	I/Y, 1870	I/Y, 1913	Years of schooling, 1870	Years of schooling, 1913	Polcon, 1870	Polcon, 1913	Market potential, 1910
Austria			3.48	5.58		0.07	55
Belgium			4.45	5.39	0.4	0.48	28
Denmark	8	12.5	4.74	6.08		0.45	20
Finland	12.4	12	0.51	1.12			
France	10.3	12.2	4.04	7.35		0.56	59
Germany	20.8	23.2	5.25	6.92		0.11	62
Greece			1.45	2.79			7
Ireland			2.15	5.5			
Italy	8.8	17.7	0.88	3.06		0.27	40
Netherlands	12.4	21.2	5.33	6.07	0.45	0.55	30
Norway	12.2	20.7	5.67	6.06		0.39	15
Portugal			0.79	2.03	0	0	11
Spain	5.2	12.2	2.43	4.93	0.17	0	26
Sweden	7.7	12	4.86	6.7		0.45	22
Switzerland			6.17	7.65	0.34	0.45	22
UK	7.7	7.5	4.13	6.35	0.33	0.47	89
United States	16.9	19.7	5.57	7.45	0.28	0.39	100

Notes: I/Y is the investment to GDP ratio in percent. "Polcon" is a measure of constraints on the executive; the United States in recent times has scored a little over 0.40. "Market potential" is a measure of proximity to markets which reflects trade costs and the spatial distribution of GDP.
Sources: Investment ratios: Carreras and Josephson (2010) and Rhode (2002);Years of schooling: Morrisson and Murtin (2009); Polcon: database for Henisz (2002); Market potential: Liu and Meissner (2013).

subset of American manufacturing (James and Skinner, 1985), and that scale economies and technological change biased in favor of capital and materials-using were pervasive in manufacturing (Cain and Paterson, 1986). This may partly have been based on localized learning as suggested by David (1975), and partly on directed technical change as in Acemoglu (2010).

Either way, looking at late Victorian Britain, the flip side of this story is that innovations that were made in the United States were frequently "inappropriate" on the other side of the Atlantic because they were not cost-effective at British relative factor prices and/or market size; had they been profit-maximizing, competition in product markets would have ensured rapid adoption (Crafts, 2012). The implication is that lower TFP in British industry was largely unavoidable. Unlike the inappropriate technology literature in development economics, however, this episode concerns the development of north-north rather than north-south technological differences.

Although American overtaking has usually been thought of as centering on industry, this is only part of the story. During the years 1871–1911, the gap between British and

American labor productivity growth was a bit larger in services than in industry, while at the same time employment in both economies shifted strongly toward services. In the services sector, American technological advance was founded on new hierarchical forms of organization based on large volumes and reduced costs of monitoring workers due to falling communication costs (Broadberry, 2006). More generally, US productivity across much of the economy during this period was driven by the organizational innovations that permitted the development of the modern business enterprise and moves toward mass production and mass distribution (Chandler, 1977).

6.2.4 Divergence, Big Time

The fact that the transition to modern economic growth happened first in Britain, and then in Continental Europe and North America, had obvious implications for the international distribution of income. True, buoyant markets in the industrial economies offered new export opportunities for the rest of the world, but this was not sufficient to prevent a large increase in the gap between the industrial rich and the non-industrial poor.

Table 6.6 provides data on per capita incomes in the major regions of the world. The data are mostly taken from Bolt and Van Zanden's (2013) revision of Maddison (2010), although in the case of Africa we have preferred Maddison's original data.[7] We distinguish between Western Europe and Eastern Europe, since industrialization first took hold in the former region, while the English-speaking settler economies of North America and Oceania are considered jointly under the heading "British offshoots." What stands out from the table is the explosive growth in incomes in the British offshoots, where they quadrupled between 1820 and 1913. As a result, this was by far the richest region in the world on the eve of the Great War. Incomes increased by two-and-a-half times in Eastern Europe and Latin America, another European offshoot, during the period, and by slightly less in Western Europe, the richest region in 1820. They increased by much less in Asia and, especially, Africa. Since these two regions had already been the poorest in the world in 1820, and since the British offshoots had been one of the richest, the result was a substantial divergence in living standards—"divergence, big time," as Pritchett (1997) has termed it.

This divergence was due to the rapid growth of the leaders, not the decline of the followers. Incomes rose everywhere over the course of the century, although in Asia, the data show a slight decline in average incomes between 1820 and 1870, perhaps as a result of deindustrialization (Williamson, 2011). From 1870 onwards all regions were growing,

[7] Bolt and Van Zanden present a weighted average of the available data, but since the only available data are for countries in North Africa, as well as Ghana and South Africa, this almost certainly leads to an overstatement of average African incomes. We prefer Maddison's data, which involved making ad hoc judgments about incomes in the rest of the continent, and have adjusted the world per capita figures accordingly.

Table 6.6 GDP per capita, 1820–1913, 1990 international dollars

	1820	1870	1913
Western Europe	1455	2006	3488
Eastern Europe	683	953	1726
British offshoots	1302	2419	5233
Latin America	628	776	1552
Asia	591	548	691
Africa	420	500	637
World	707	874	1524

Sources: Bolt and Van Zanden (2013) and Maddison (2010).

and ended the century more prosperously than they had begun it. Between 1820 and 1913, average incomes rose by 52% in Africa, but by just 17% in Asia.

The net effect was a dramatic increase in international income differentials. In 1820, the then richest region, Western Europe, had an average income twice the world average, and three and a half times the African average. By 1913, Western European incomes were 129% higher than the world average, a small increase, but five and a half times the African average, a sizeable one. Over the same period incomes in the British offshoots rose from being 84% higher than the world average to being 243% higher. By 1913, they were more than eight times those in Africa. Bourguignon and Morrisson (2002, p.734) found that the Theil between-country inequality coefficient almost quintupled between 1820 and 1910.[8]

It is clear, then, that the 19th century saw a large increase in global inequality, driven above all by the rapid income growth of some countries but not others. It is also clear that the primary cause of this rapid income growth was industrialization in Europe and North America. Strikingly, however, some countries, such as Australia and Argentina, had among the highest incomes in the world while remaining largely specialized in primary production. To explain this apparent paradox, we would follow Arthur Lewis (1978), and point to the immigration policies of these resource-abundant countries. While countries such as Burma saw the large-scale immigration of workers from China or India, the temperate settler economies restricted immigration to Europeans only. Racism was undoubtedly a factor here, but the policy also helped maintain living standards. As Lewis (1978, p. 188) put it, "The temperate settlements could attract and hold European emigrants, in competition with the United States, only by offering income levels higher than prevailed in north-west Europe." By appropriately regulating immigration flows, and by absorbing the capital and new technologies of the core, resource-abundant settler economies could thus import rising British living standards.

[8] Based on the data in Maddison (1995).

6.2.5 The Great Specialization

The fact that the Industrial Revolution reduced manufacturing costs so substantially during the 19th century, but only in a small portion of the world, created the potential for a stark international division of labor. Falling transportation costs and relatively liberal trade policy allowed this potential to be realized. North-west Europe and, especially, the United Kingdom exported manufactured goods and imported primary products, while the exports of Oceania, Latin America, and Africa consisted almost entirely of primary products. North America was an intermediary case: its vast natural resources implied net exports of primary products, but rapid industrialization meant that the United States switched to being a net exporter of manufactures just before World War I. Asia was another intermediary case: while it conformed to the peripheral pattern of net primary exports and net manufactured imports, its manufactured exports were non-negligible.

The "great specialization", as Dennis Robertson (1938) called it, between an industrial north and a primary-exporting south, thus dates from the 19th century. Its causes were straightforward enough: geographically unbalanced technological change, and a dramatic reduction in transport costs. Its consequences, especially for the south, were less so. On the one hand, booming northern markets and falling transport costs implied rising terms of trade, especially prior to the 1870s (Williamson, 2011), and this benefited commodity exporters. On the other hand, insofar as this further hastened deindustrialization, it potentially imposed dynamic costs on southern economies, by depriving them of the growth-enhancing externalities associated with manufacturing, by leading to rent-seeking behavior associated with an over-reliance on resource-based production, or by exposing them to greater terms of trade volatility (*ibid.*). Many of the great policy debates of the 20th century thus have their roots in this period. Should developing countries rely on exports of primary commodities to generate growth (a strategy which worked for several countries in the late 19th century [Lewis, 1969, 1970])? Or did such an outward-oriented strategy give rise to Dutch disease problems, suggesting (on the assumption that there are growth-promoting externalities in industry) the need for policy interventions (such as import-substitution strategies) to increase industrial production? The way in which these debates influenced policy decisions would have a major impact on regional growth experiences once the developing world regained policy independence in the 20th century.

It should be noted, however, that by the end of this period several parts of the periphery were reindustrializing. The best-known example is Japan, but there was also rapid industrial growth, albeit from a low base, in several Asian economies, e.g. in Korea, the Philippines, Taiwan, and parts of China. There was also rapid industrial growth in Mexico, Brazil, and the Latin American Southern Cone (Gómez Galvarriato and Williamson, 2009). The spread of industrialization across the developing world would become one of the main features of 20th century economic growth.

Table 6.7 GDP per capita, 1870–2007, 1990 international dollars

	1870	1913	1950	1973	1990	2007
Western Europe	2006	3488	4517	11,346	15,905	21,607
British offshoots	2419	5233	9268	16,179	22,346	30,548
Japan	737	1387	1921	11,434	18,789	22,410
"West"	**1914**	**3690**	**5614**	**13,044**	**18,748**	**25,338**
Asia minus Japan	539	652	639	1223	2120	4830
Latin America	776	1552	2505	4517	5065	6842
Eastern Europe and former USSR		1519	2594	5741	6458	7731
Africa	500	637	889	1387	1425	1872
"Rest"		**853**	**1091**	**2068**	**2711**	**4744**
World	**874**	**1524**	**2104**	**4081**	**5149**	**7504**

Sources: Bolt and Van Zanden (2013), Maddison (2010). This is a revised version of Table 4 in Maddison (2005).

6.3. TWENTIETH CENTURY GROWTH: WHAT HAPPENED?

In this section we briefly set out some of the major facts concerning aggregate growth in the major regions of the world.

6.3.1 World Growth and Its Decomposition

Table 6.7 presents data on the level of per capita GDP between 1870 and 2007, based on Bolt and Van Zanden's (2013) updating of Maddison (2010). As before, we have preferred Maddison's original figures for Africa up to and including 1913, and have revised the world figures accordingly. We follow Maddison in distinguishing Japan from the rest of Asia (which we will, for the sake of brevity, refer to henceforth as "Asia"), since Japan was a precocious industrializer.[9] We also follow Maddison in grouping Western Europe, Japan, and the British offshoots (the United States, Canada, Australia, and New Zealand) together (the "West"), and in considering separately the other four regions (the "Rest"), which we will refer to jointly as the developing world.

Table 6.8 gives per capita GDP growth rates in five successive periods; the late 19th century (1870–1913); the turbulent years between 1913 and 1950; the "Golden Age" which lasted from 1950 to 1973; the period following the first oil crisis, from 1973 to 1990; and the period since 1990.[10] Whereas Maddison treated the entire period since 1973 as one, we have preferred to split it into two, since the years after 1990 were marked

[9] We use Maddison's population data to derive the average figures for the West, the Rest, Asia minus Japan, and Eastern Europe and the former USSR.

[10] The figure for Eastern Europe and the former USSR is a population-weighted average of the growth rates of the two regions, where the latter growth rate is (in the case of 1870–1913) calculated for the period 1885–1913 only, since data for 1870 are lacking.

Table 6.8 Per capita GDP growth, 1870–2007 (percent per annum)

	1870–1913	1913–1950	1950–1973	1973–1990	1990–2007	1913–2007
Western Europe	1.29	0.70	4.09	2.01	1.82	1.96
British offshoots	1.81	1.56	2.45	1.92	1.86	1.89
Japan	1.48	0.88	8.06	2.96	1.04	3.00
"West"	**1.54**	**1.14**	**3.73**	**2.16**	**1.79**	**2.07**
Asia minus Japan	0.45	−0.06	2.87	3.29	4.96	2.15
Latin America	1.63	1.30	2.60	0.68	1.78	1.59
Eastern Europe and former USSR	1.64	1.46	3.51	0.69	1.06	1.75
Africa	0.57	0.90	1.95	0.16	1.62	1.15
"Rest"	**0.73**	**0.67**	**2.82**	**1.61**	**3.35**	**1.84**
World	**1.30**	**0.87**	**2.92**	**1.38**	**2.24**	**1.71**

Sources: Based on Bolt and Van Zanden (2013), Maddison (2010). This is a revised version of Table 6 in Maddison (2005).

by the collapse of the Soviet Union, the rapid spread of globalization, and a succession of international financial crises.

Table 6.8 shows that world economic growth was higher in the 20th century (1913–2007) than in the late 19th (1870–1913), at 1.7% per annum as opposed to 1.3%. Latin America is the only exception to the rule that 20th century growth was faster, and even there growth rates in the two periods were very similar. Growth was also very similar in the two periods in the British offshoots, reflecting the relative constancy of the long-run United States growth rate (Jones, 1995).

These aggregate 20th century figures disguise a considerable amount of variation between periods. The period between 1913 and 1950, marked by two world wars and the Great Depression, saw world growth fall to just 0.9%. It declined everywhere with the exception of Africa, although it fell by less in the British offshoots, which saw strong wartime growth (helping to offset an especially severe depression after 1929), and in Eastern Europe and the former USSR, where Stalin embarked on a major industrialization drive during the interwar period. The period between 1950 and 1973 clearly deserves the "Golden Age" label, with world growth rates higher than at any other time in history. All regions saw their highest ever growth rates during this quarter century, with just one exception: Asian growth accelerated after 1973, and again after 1990.

With the aforementioned exception of Asia, growth rates declined everywhere after 1973. After 1990, they continued to decline in the "West," but they increased in all four developing regions. The result was that, for the first time since 1870, per capita growth rates after 1990 were higher in the "Rest" than in the "West."

Table 6.9 World shares of GDP, 1870–2007 (percent)

	1913	1950	1973	1990	2007
Western Europe	33.3	26.0	25.5	22.2	17.4
British offshoots	21.3	30.8	25.4	24.6	22.1
Japan	2.6	3.0	7.8	8.6	5.8
"West"	**57.3**	**59.7**	**58.6**	**55.4**	**45.2**
Asia minus Japan	22.1	15.6	16.4	23.3	37.0
Latin America	4.6	7.8	8.7	8.3	7.9
Eastern Europe and former USSR	13.1	13.0	12.9	9.8	6.3
Africa	2.9	3.8	3.4	3.3	3.6
"Rest"	**42.7**	**40.3**	**41.4**	**44.6**	**54.8**
World	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**

Sources: Bolt and Van Zanden (2013), Maddison (2010).

Table 6.9 presents data on regional shares of world GDP. This requires information on not only per capita GDP levels, but population sizes, the latter being taken from Maddison (2010). As can be seen, the West's share of world GDP peaked in 1950, at almost 60%, before declining slowly before 1990, and more rapidly thereafter: in 2007 it was just 45%. This overall trend masks considerable variation within the "West". The share of the British offshoots was slightly higher in 2007 than in 1913, at 22%, although it was over 30% in the immediate aftermath of World War II, declining slowly thereafter. In contrast, Western Europe's share fell by almost a half, from 33% to 17%; while Japan's share rose from 2.6% in 1913 to 8.6% in 1990, before falling sharply afterward. Within the developing world Asia's share fell substantially between 1913 and 1950, had recovered by 1990, and increased rapidly since then. It was over a third in 2007. The Latin American share rose in the early 20th century, and has held steady since 1950; while Africa's share rose between 1913 and 1950 and has been stable since then. One of the most striking features of the table is the share of Eastern Europe and the former USSR, which was steady until 1973 and then collapsed, falling not just during the last two decades of communism, but after 1990 as well. The share was just 6% in 2007, less than half the 1973 level.

6.3.2 Catching Up, Forging Ahead, and Falling Behind

Since the publication of Moses Abramovitz's (1986) presidential address to the Economic History Association, it has become commonplace to distinguish between economic growth in the leading economy or economies, at the frontier of technological knowledge, and in follower countries which may or may not be catching up on that frontier. Growth in the leading economy is determined by those forces pushing back the frontier; growth in the followers is determined by the extent to which they can import technologies from the leading economies, and embody them in their own capital stock.

Abramovitz pointed out that such catching up is inherently self-limiting, an insight that has been subsequently formalized by growth theorists such as Robert Lucas (2000, 2009). His argument that catching-up was dependent on adequate "social capability" anticipated the enormous literature on conditional convergence. Abramovitz also argued that, given social capability, circumstances had to be conducive to the international diffusion of knowledge. Subsequent research has followed Abramovitz's lead, focusing both on the diffusion of technologies (Comin et al. 2006; Comin and Hobijn, 2010) and on the role of trade in stimulating or hindering the process. While there are disagreements on many details of the international growth process, the broad distinction between growth in the leaders and in the followers tends to be taken as given.

A common theme in economic history is the story of how economic leadership has passed from nation to nation over the course of the last millennium. How to explain this remains unclear (for one attempt to do so, see Brezis et al. 1993). Fortunately, for our purposes the issue is moot, since it is commonly accepted that the economic leader throughout the 20th century was the United States, although it was not until after World War II that the US was willing to translate its technological superiority into economic policy leadership. Figure 6.1 shows the evolution of per capita GDP in the United States between 1800 and 2007, allowing the 20th century performance to be compared with what came before. As is well known (Jones, 1995), and has already been noted, per capita growth rates have been remarkably stable in the United States over time. The heavy straight line is a linear projection backwards and forwards in time of trend growth during the late 19th century (1870–1913). The shaded areas represent the US Civil War (1861–1865), World War I (1917–1918), and World War II (1941–1945), while the dashed vertical lines represent the onset of the Great Depression (1929) and the first oil crisis (1973).

As can be seen, per capita growth accelerated in the United States after 1870. It averaged 1.8% per annum between 1870 and 1913, as opposed to 1.2% between 1820 and 1870.[11] As can also be seen, the long-run trend was very similar in the 20th century, despite the remarkable collapse in incomes during the Great Depression, and the equally remarkable increase in per capita output during World War II. Growth averaged 2.1% per annum between 1913 and 2007, with a slight acceleration evident from the early 1980s. Consistent with Lucas (2000, 2009), per capita growth in the frontier economy has been around 2% per annum for a very long time.

For variety and drama, we need to turn to the followers. There, the 20th century has thrown up growth miracles, reversals of fortune, and sorry tales of steady decline (Pritchett, 2000). Figure 6.2 plots per capita GDPs in the major economies and regions of the world, as a percentage of US GDP, thus indicating whether or not these countries were converging on the technological frontier, keeping pace, or falling further behind. For

[11] Here and elsewhere, reported growth rates are based on regressions of the log of per capita output on time.

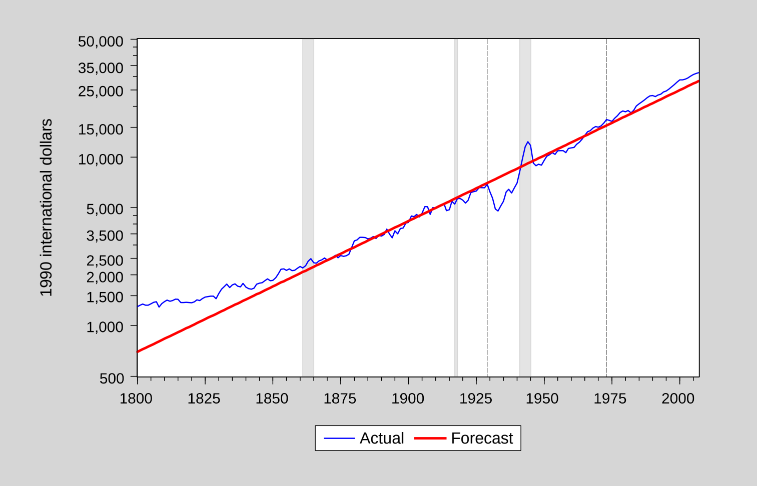


Figure 6.1 US GDP per capita, 1800–2007 (1990 international dollars). *Sources:* *Bolt and Van Zanden (2013)*.

the sake of brevity we will henceforth refer to these percentages as countries' or regions' relative GDP, or relative income. While our interest is in the 20th century experience, we also provide the backstory by plotting the trends beginning in 1870. The figures are, for the most part, regional averages, and therefore average out individual country experiences.

The major point that emerges from the late 19th century data is that as US growth accelerated after the Civil War, other regions, with three exceptions, saw their relative incomes decline. For the purposes of Figure 6.2 we have dated the two world wars, in Eurocentric fashion, between 1914–1918 and 1939–1945.

The first exception is Japan, which managed to keep pace with the United States after 1870. Like all regions it saw its relative GDP increase during the catastrophic interwar period, and then decline during the Second World War. It then caught up on the technological frontier in impressive fashion, experiencing per capita growth of 8% per annum during the Golden Age (Table 6.8), and overtaking Western Europe in the late 1970s. Its subsequent relative decline has been quite astonishing: Japan's relative GDP peaked at almost 85% in 1991, but it was only around 70% in 2007, back to the level of 1979.

The second exception is Latin America, whose relative GDP, like that of Japan, remained constant at just under 30% between 1870 and 1913. Unlike Japan, it stayed at this level until 1940, avoiding both the catch-up of the Depression years and the collapse that followed during World War II: one interpretation might be that the continent's economies were closely linked with that of the US during this period. Indeed, Latin America's relative income remained fairly constant during the next four decades, dipping to around 25–26% during the 1950s and 1960s, and recovering its 19th century level of

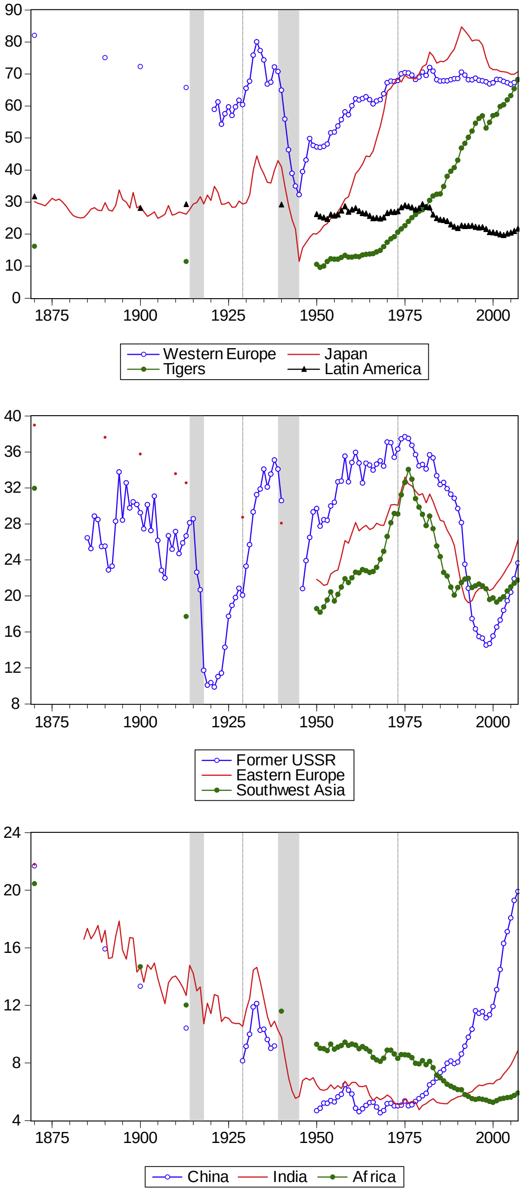


Figure 6.2 Regional GDP per capita, 1870–2007 (percentage of US level). *Sources:* *Bolt and Van Zanden (2013) and Maddison (2010).*

29% in 1980. The next three decades saw Latin America's relative income steadily decline, and it stood at just 20% at the end of the 20th century.

The East Asian Tigers (Hong Kong, South Korea, Singapore, and Taiwan) are, with Japan, the main success story emerging from Figure 6.2. Their relative GDP fell from 16% in 1870 to around 11% by 1913, and there it stood until 1950. It then started to rise, accelerating dramatically in the late 1960s, until by 2007 it stood at just under 70%, on a par with both Western Europe and Japan.

The European growth miracle of the Golden Age was real enough, with growth rates of 4% per annum, but in a longer run perspective this quarter century episode stands out as an exception to what was a generally disappointing performance. Like most other regions, Western Europe's relative GDP fell between 1870 and 1913, from 82% to 66%, and it collapsed during World War II to a low point of 32% in 1945. The Golden Age saw the region's relative GDP recover to its 1913 level, and even surpass it slightly, so that it stood at around 70% in the early to mid-1970s. Since then there has been absolutely no convergence on the technological frontier.

The most dramatic experience, in this catching-up perspective, was probably that of the former USSR. This region was the third to keep pace with the United States during the late 19th century (although we only have data from 1885), but its relative GDP was highly volatile during the period. It then collapsed during the First World War, recovered dramatically during the interwar period to the point where it surpassed its previous peak, reaching 35% in 1938. It collapsed again during the Second World War, and recovered in equally dramatic fashion, peaking at 38% in 1975. There followed a spectacular decline, to a nadir of 14.5% in 1998. It then rose sharply, reaching 24% in 2007.

Given the extent to which the Soviet and Eastern European economies were connected after 1945, it is not surprising to see Eastern Europe's relative GDP tracing out the same rise and fall as its imperial master before and after 1975. It arrested its decline earlier than the former USSR, in 1993, and has been richer ever since. More surprising is the fact that Southwest Asia, which essentially comprises oil rich states in the Middle East and the Gulf, along with Israel, Lebanon, and Turkey, followed a very similar trajectory as well, with its post-1976 decline extending all the way to 2001.

Finally, Africa, China, and India all saw their relative incomes decline steadily until the late 20th century. China's relative GDP fell more sharply early on, and then stagnated at a very low level from 1950 onwards, about 5%, before starting a remarkable rise in the late 1970s. It stood at 20% in 2007. India's relative decline was slower, and its catch-up began around a decade after China's, again from a level of around 5%. Africa's relative decline was the slowest of all, with its relative GDP only hitting 5% in the mid-1990s. From 2000 onwards it started to slowly recover, reaching 6% in 2007.

6.4. THE PROXIMATE SOURCES OF GROWTH

This section explores the proximate sources of growth, as revealed by growth accounting techniques. We provide a broad overview of results relating to 20th century economic growth. We also review a number of issues relating to the use of these methods and, in particular, the interpretation of results obtained by using them. Handled with care, we believe that growth accounting can provide an important benchmarking or diagnostic tool but there is also considerable scope to make misleading comparisons or inferences.

6.4.1 Conventional Growth Accounting Results

The conventional growth accounting approach assumes that GDP is given by:

$$Y = AK^{\alpha}L^{1-\alpha},$$

where Y is output, K is capital, L is labor, and A is TFP, while α and $(1 - \alpha)$ are the elasticities of output with respect to capital and labor, respectively. The level of TFP is usually measured as a residual after the other items in the expression have been measured.

This can be converted into the basic growth-accounting formula:

$$\Delta \ln(Y/L) = \alpha\, \Delta \ln(K/L) + \Delta \ln A,$$

which gives a decomposition of the percentage rate of growth of labor productivity into a contribution from the percentage rate of growth of capital per unit of labor input (capital deepening) and a term based on the percentage growth rate of TFP. For benchmarking purposes, it is convenient to adopt a standardized value for α.[12]

It is tempting but misleading to assume that residual TFP growth in this formula captures the contribution of technological progress to labor productivity growth. Technological change may be less than TFP growth if there are scale economies or improvements in the efficiency with which factor inputs are used. On the other hand, if technological progress is partly embodied in new forms of capital (rather than "manna from heaven") then some of its contribution will seem to accrue to capital when this approach is used.

A more general approach seeks to take account of human capital and modifies the production function to be:

$$Y = AK^{\alpha}(L^{*}(HK/L))^{1-\alpha},$$

where HK/L is the average educational quality of the labor force, typically approximated by years of schooling. The growth-accounting formula then becomes:

$$\Delta \ln(Y/L) = \alpha\, \Delta \ln(K/L) + (1 - \alpha)\Delta \ln(HK/L) + \Delta \ln A,$$

so that the decomposition now includes a contribution from the rate of growth of the quality of the labor force, which in practice is based on the additional earnings from years

[12] It is common to use $\alpha = 0.35$ which is similar to the share of profits in GDP for many countries. The profits share is potentially a misleading estimate of the output elasticity of capital, for example in the presence of significant externalities or market power, but in practice it is probably acceptable (Aiyar and Dalgaard, 2005; Bosworth and Collins, 2003).

Table 6.10 Accounting for labor productivity growth in OECD Countries, 1913–1950 (percent per annum)

	K/L	HK/L	TFP	Y/L
France	0.59	0.36	1.06	2.01
Germany	0.19	0.22	0.74	1.05
Japan	0.62	0.61	0.49	1.72
Netherlands	0.43	0.27	0.88	1.58
UK	0.42	0.32	0.83	1.57

Source: Maddison (1987).

of schooling. The estimates of the TFP growth contribution are less crude and, of course, tend to be smaller once education is taken into account.

Tables 6.10 and 6.11 report growth-accounting estimates on this basis where the methods used allow international comparisons to be made. Taken at face value, several interesting points stand out from these estimates. First, even after allowing for education, TFP growth in the advanced economies compares very favorably with the 19th century until the end of the Golden Age in the 1970s. Second, the rise of East Asian countries after 1960 is notable for a very strong capital deepening contribution to labor productivity growth, which was much greater than had been observed in the European transition to modern economic growth in the 19th century. Third, TFP growth in sub-Saharan Africa was disastrous in the last 30 years of the 20th century and most disappointing in Latin America post-1980, and for both these regions there was virtually no capital deepening contribution after 1980.

Table 6.12 provides an account of productivity gaps based on an application of growth accounting to levels pioneered in a classic article by Hall and Jones (1999). Its results are quite similar to those given in that paper although for a later study that took place in 2005. The results are striking: by far the most important reason for differences in labor productivity (and income per head) is differences across countries in levels of TFP.[13] This is a striking rejection of the basic set-up of the pure Solow growth model which assumes that technology is the basis of TFP and is both exogenous and universal—an assumption which underpins the neoclassical predictions of β- and σ-convergence.

In principle, there are two reasons why TFP levels may differ, namely technology and efficiency. The most obvious reason why technology might differ is that technological progress has been uneven and has improved the production function at some factor intensities (high capital-labor or human capital-labor ratios) but not others. The evidence

[13] There are alternative ways to specify the "development accounting" equation and measurement issues, in particular with regard to human capital. Nevertheless, there seems to be general agreement that residual TFP is the biggest part of the story, accounting for 50–70% of cross-country income differences (Hsieh and Klenow, 2010).

Table 6.11 Proximate sources of labor productivity growth, 1960–2003 (percent per annum)

	K/L	HK/L	TFP	Y/L
Industrial countries				
1960–1970	1.4	0.3	2.3	4
1970–1980	1	0.5	0.4	1.9
1980–1990	0.6	0.2	0.9	1.7
1990–2003	0.8	0.2	0.6	1.6
East Asia				
1960–1970	1.7	0.4	1.6	3.7
1970–1980	2.7	0.6	1	4.3
1980–1990	2.5	0.6	1.3	4.4
1990–2003	2	0.5	0.6	3.1
Latin America				
1960–1970	0.8	0.3	1.7	2.8
1970–1980	1.3	0.3	1.1	2.7
1980–1990	0	0.5	−2.3	−1.8
1990–2000	0.1	0.3	−0.1	0.3
Sub-Saharan Africa				
1960–1970	0.8	0.2	1.9	2.9
1970–1980	1.3	0.1	−0.4	1
1980–1990	−0.1	0.4	−1.5	−1.2
1990–2000	0	0.4	−0.5	−0.1

Sources: Bosworth and Collins (2003) and website update.

Table 6.12 Decomposition of cross-country differences in GDP per capita, 2005 (USA = 100)

	Y/P	K/Y	HK/L	L/P	TFP
United States	100	100	100	100	100
Japan	72.6	130.7	100.4	105.1	52.6
EU27 + EFTA	64.7	114.1	91.2	91.3	67.8
Russia	28.6	97.4	84.9	99.3	31.5
Brazil	20.5	103.1	70.1	96.8	29.3
China	9.8	105.2	57.3	119.5	13.6
India	5.2	98.3	47.7	87.1	12.7
World	22.8	104.2	64.2	95.8	27.9

Notes: GDP per capita (Y/P) is measured at PPP. Estimates derived by imposing the production function $Y = K^{\alpha}(AhL)^{1-\alpha}$ where h is human capital per worker (HK/L). This can be re-written as $Y/L = (K/Y)^{\alpha/(1-\alpha)}$ Ah so that $Y/P = (K/Y)^{\alpha/(1-\alpha)}$ Ah(L/P) which is the formula used for the decomposition.
Source: Duval and De la Maisonneuve (2010).

suggests that this was the case during the 20th century (Allen, 2012), as might be expected, in a world of directed technical change where research and development is oriented primarily to the incentives provided by the economic environment of advanced economies. In other words, there could be an "inappropriate technology" explanation for the TFP gap. An inefficiency explanation for TFP gaps might relate to differences in institutional quality which impact on allocative and/or productive inefficiency. Again, there is evidence that points in this direction, notably the finding by Hsieh and Klenow (2009) that if capital and labor were used as efficiently in Chinese and Indian manufacturing as in the United States, TFP would increase by 30–50% and 40–60%, respectively.

Table 6.13 reports results from one attempt to discriminate between these two hypotheses. The overall conclusion in Jerzmanowski (2007) is that in 1995 (1960) factor inputs accounted for 31 (45)% of the variation in output per worker while of the 69 (55)% attributable to TFP, 43 (28)% came from efficiency and 26 (27)% from technology differences. These estimates imply that, while both efficiency and technology are important in explaining TFP gaps, on average, efficiency matters more, and increasingly so over time. These results suggest that episodes of rapid catch-up growth are likely to be based

Table 6.13 Decomposing TFP levels relative to the United States (USA = 1.00)

	1960			1995		
	TFP	**E**	**T**	**TFP**	**E**	**T**
France	0.72	0.71	1.01	0.77	0.87	0.89
Greece	0.49	0.57	0.86	0.56	0.58	0.97
Spain	0.64	0.74	0.86	0.76	0.85	0.9
Italy	0.67	0.71	0.94	0.84	0.88	0.96
UK	0.85	0.89	0.95	0.82	0.85	0.97
India	0.3	0.41	0.74	0.29	0.44	0.67
Indonesia	0.31	0.55	0.57	0.37	0.54	0.69
Japan	0.48	0.56	0.86	0.68	0.79	0.86
Korea	0.33	0.37	0.88	0.49	0.49	0.99
Singapore	0.47	0.54	0.87	0.85	1	0.85
Argentina	0.76	0.79	0.96	0.57	0.65	0.88
Brazil	0.42	0.49	0.86	0.5	0.6	0.84
Chile	0.51	0.57	0.89	0.58	0.73	0.8
Mexico	0.65	0.72	0.9	0.49	0.58	0.84
DR Congo	0.38	0.58	0.65	0.23	0.35	0.67
Malawi	0.23	0.39	0.6	0.16	0.27	0.61
Mauritius	0.62	0.71	0.88	0.8	1	0.8
Tanzania	0.15	0.22	0.69	0.11	0.17	0.64

Note: $\text{TFP}_\text{i} = \text{Efficiency}_i{*}\text{Technology}_i = \text{E}_\text{i}\text{T}_\text{i}$ where E_i is obtained by estimating an efficient production frontier, TFP is obtained by growth accounting in levels and T is then inferred.
Source: Jerzmanowski (2007).

on major improvements in both efficiency and technology. They also suggest that the lengthy period of negative TFP growth in Africa reported in Table 6.11 is a sign of deteriorating efficiency of factor use, rather than of technological retrogression.

6.4.2 Some Issues of Measurement and Interpretation[14]

There are a number of important issues that have to be addressed when trying to compare growth-accounting exercises for the late 19th and early 20th centuries with similar exercises for the late 20th century. For example, to obtain estimates of real GDP an accurate GDP price deflator is required. Boskin et al. (1996) thought that, for a variety of reasons, inflation had been overestimated (and thus real GDP growth and TFP growth had been underestimated by a similar amount) in the national accounts during the period of the productivity slowdown in the 1970s and 1980s, and that the correction required was of the order of 0.6% per year. On the other hand, the Boskin bias in inflation measurement does not appear to generalize to other periods (Costa, 2001).

Perhaps a more serious concern in many cases is a potential index number problem regarding the measurement of capital inputs. The standard approach, used in virtually all historical studies, relies on estimates of the perpetual inventory capital stock that are weighted using asset prices. A theoretically more appropriate (but much more data demanding) method is to estimate flows of capital services using rental prices as weights. This requires estimates of the user cost of capital for different assets. The difference between the two methods will be especially important when investment switches toward short-lived assets (like computers) and away from long-lived assets (like structures), since the user cost of the former is much higher relative to the asset price. Not surprisingly, this issue has come to prominence since the ICT revolution.[15] Generally speaking, using the capital services methodology raises the growth contribution of capital and lowers that of TFP. However, this probably makes relatively little difference, even in the United States, before the second half of the 20th century, as Table 6.14 reports; it is, however, very important for analyses of recent growth.[16]

It has long been recognized that research and development is an intangible investment and that the R & D knowledge stock could be introduced as an input in growth-accounting estimates. More recently, it has been argued that intangible investments generally (including design and product development; investments in branding; firm-specific human and organizational capital formation including training and consultancy; and computerized information, especially software) should be treated in this way.

[14] This section draws in part on Crafts (2009a, 2010).

[15] Estimates for the UK show that the volume of capital inputs for the period 1950–2006 grew by 3.1% per annum measured by the traditional capital stock data but by 3.5% when measured by the capital services method. The difference relates entirely to the post-1980 and, in particular, the post-1990 period (Wallis, 2009).

[16] The EUKLEMS database which covers recent decades and permits international comparisons is constructed using a capital services methodology: see O'Mahony and Timmer (2009).

Table 6.14 Sources of labor productivity growth in the United States (percent per annum)

	K/L	Crude TFP	Labor quality	Capital quality	Refined TFP	Y/L
1800–1855	0.19	0.2	0	0	0.2	0.39
1855–1871	0.53	−0.39	0	0	−0.39	0.14
1871–1890	0.84	1	0	0	1	1.84
1890–1905	0.55	1.38	0.1	0	1.28	1.93
1905–1927	0.48	1.57	0.19	0	1.38	2.05
1929–1948	0.07	1.89	0.38	0.08	1.43	1.96
1948–1966	0.81	2.3	0.43	0.4	1.47	3.11
1966–1989	0.57	0.66	0.31	0.31	0.04	1.23

Note: capital quality reflects the adjustment required to move from a capital stock to a capital services basis.
Source: Abramovitz and David (2001, Table 1: IVA).

Expenditure on these items has been growing rapidly in the context of the "knowledge economy" and in both the UK and the USA, is of similar magnitude to investment in tangible capital. If these expenditures are treated as final investment rather than intermediates in growth-accounting exercises, this will imply that there is more output, more input, and revised factor-share weights.

In principle, the impact of switching to accounting with intangibles on TFP growth is ambiguous. In practice, at least in the ICT era, the impact is to increase estimated labor productivity growth a bit, to raise the contribution of capital deepening considerably, and to reduce measured TFP growth appreciably, as Table 6.15 reports. Growth accounting with intangibles for earlier periods has not yet been attempted but the impact would surely be much less dramatic since, in the 1950s, intangible investment added only about 4% to US GDP compared with about three times that amount 50 years later.

Neoclassical growth accounting is normally carried out by imposing a Cobb-Douglas production function. In some circumstances, it may be that a CES specification is more appropriate with the elasticity of substitution between capital and labor, σ, being set to a value less than 1. In that case, especially when the capital-labor ratio is growing rapidly and technical change exhibits capital-using bias, TFP growth will be underestimated by the conventional method. For example, taken at face value, the estimates in Table 6.14 (which assume that $\sigma = 1$) invite the conclusion that technical change was insignificant in the American economy for much of the 19th century, and only became significant with the rise of the science-based industries and R & D in the so-called Second Industrial Revolution. This runs counter to standard historical discussions, however, and is certainly not the interpretation in Abramovitz and David (2001). If, as they argue, the 19th century US economy was characterized by a low elasticity of substitution between factors and capital-using technical change, then TFP growth was considerably stronger than shown in Table 6.14. If estimates are obtained assuming $\sigma = 0.3$, as Abramovitz and David believe is appropriate, then TFP growth turns out to have been 0.9% per year between

Table 6.15 Sources of labor productivity growth, United States non-farm business sector, 1973–2003 (percent per annum)

	1973–1995	1995–2003
Traditional growth accounting		
Labor productivity growth	1.36	2.78
Capital deepening	0.6	0.98
IT capital	0.33	0.7
Other tangible capital	0.27	0.28
Labor quality	0.28	0.38
TFP	0.48	1.42
Accounting with intangibles		
Labor productivity growth	1.63	3.09
Capital deepening	0.97	1.68
Tangible capital deepening	0.55	0.85
IT capital	0.3	0.6
Other tangible capital	0.25	0.24
Intangible capital deepening	0.43	0.84
Software	0.12	0.27
Other intangible capital	0.31	0.57
Labor quality	0.25	0.33
TFP	0.41	1.08

Note: accounting with intangibles is based on the formula $\Delta\ln(Y^*/L) = s^*_{TK}\Delta\ln(TK/L) + s^*_{IK}\Delta\ln(IK/L) + \Delta A/A$, where Y^* includes expenditure on intangible investments and s^*_{TK} and s^*_{IK} are the factor shares of tangible and intangible capital in Y^*.
Source: Corrado et al. (2009).

1835 and 1890, much higher than a crude estimate of 0.24% per year, assuming $\sigma = 1$.[17] Other cases where a similar issue arises, and which are discussed below, include the "East Asian Miracle" (Rodrik, 1997) and the 1970s growth slowdown in the USSR (Allen, 2003).

A further problem with conventional growth accounting that matters in some circumstances is that it assumes no costs of adjustment, fixed factors of production, or economies of scale. Morrison (1993) proposed an econometric procedure to address these problems and her results indicated that the 1970s slowdown in TFP growth in American manufacturing was very largely a weakening of economies of scale rather than of technological progress. Using Morrison's methodology, Crafts and Mills (2005) found that adjustment costs meant that technological progress was about 2 percentage points faster than conventional TFP growth in both British and German manufacturing during 1950–1973

[17] This calculation applies the correction to TFP growth in Rodrik (1997). The correction is given by $0.5\,\alpha((1-\sigma)/\sigma)(1-\alpha)(\Delta K/K - \Delta L/L)(\Delta A_L/A_L - \Delta A_K/A_K)$ where the last term captures the degree of factor-saving bias in technological progress measured as the difference between the rate of labor augmentation and the rate of capital augmentation.

but not much different thereafter. Once again, as with the previous examples, the point is that intertemporal comparisons of conventional TFP growth may be hazardous because the degree of measurement bias appears to have varied considerably over time.

Underpinning growth accounting in the neoclassical tradition is, of course, the neoclassical growth model.[18] The later development of endogenous growth models could potentially call for alternative growth-accounting formulae or a different interpretation of the standard results (Barro, 1999). The most obvious implication might be to recognize the importance of the embodiment of technical change in new varieties of capital, as in the voluminous literature that has applied growth accounting to the impact of ICT (e.g. Oliner et al. 2007). The growth-accounting formula that has been applied in the ICT literature is:

$$\Delta \ln(Y/L) = \alpha_{KO} \Delta \ln(KO/L) + \alpha_{KICT} \Delta \ln(KICT/L) + \phi\, \Delta \ln A_{ICT} + \eta\, \Delta \ln A_{O},$$

where ϕ and η are gross output weights, KICT is capital used in ICT production, KO is the rest of the capital stock, A_{ICT} is TFP in ICT production, and A_O is TFP in the rest of the economy. The contribution of the new ICT technology to labor productivity growth is taken to be the sum of the second and third terms. Given that ϕ and α_{KICT} are very small initially, it is easy to see why a new GPT initially adds very little to overall labor productivity growth. By including the ICT capital deepening term, however, the implication is that TFP growth underestimates the contribution of technological progress to growth.

It should be noted that this approach seeks only to benchmark the direct ex-post ICT component of productivity growth. It does not answer the (much harder) question, "How much faster was productivity growth as a result of ICT?" This hinges on the counterfactual rate of growth of other capital in the absence of ICT, estimation of which would be a complex modeling exercise taking account of both crowding out and crowding in effects. Fogel (1964) took the view that no capital deepening component should be included because in the absence of the new technology similar returns would have been earned on alternative investments. However, this is not a position that everyone would accept, especially in the case of GPTs.[19]

This links to a deeper concern regarding the use of growth accounting to identify the sources of growth, which was very clearly articulated by Abramovitz (1993). The issue is two-way interdependence between the trajectories of technological change on the one hand, and physical and human capital formation on the other. While some endogenous growth models stress the latter interdependence, it is actually the former

[18] As Griliches (1996) underlined, the big contribution of Solow (1957) was to put the economics into growth accounting by making this connection.

[19] Fogel (1964) measured the contribution of railways to American economic growth in terms of "social savings," essentially a measure of user benefits arising from the impact of technical change on the transport supply curve. It is easy to show that this is equivalent to $\phi\, \Delta \ln A_{ICT}$ (Foreman-Peck, 1991).

which is highlighted by a comparison of the American growth process in the 19th and 20th centuries.[20]

Three key points should be noted. First, using conventional growth accounting to estimate TFP growth is not always a good guide to underlying technological change. As we have seen, TFP growth can be either an under- or an overestimate of the contribution of technological progress to economic growth. Second, the size and direction of the bias in neoclassical growth accounting varies considerably in different periods or types of economy; this can make historical comparisons quite difficult. Third, while growth accounting invites its users to treat the growth of capital and technological change as independent and additive, this assumption is potentially quite misleading and may detract from a deeper understanding of the sources of growth.

6.4.3 Economic Miracles are Not All the Same[21]

All of that having been said, an interesting application of growth accounting is to compare episodes of rapid catch-up growth, which exhibit some striking differences when viewed through this lens.[22] Table 6.16 reports estimates relating to the Golden Age of Western European growth, the East Asian Miracle, the Celtic Tiger, the rise of the BRICs, and the Soviet Union. This last case ended in failure but back in the 1960s it was conventional wisdom that the USSR was on track to overtake the United States before the beginning of the 21st century (Levy and Peart, 2011).

The European Golden Age saw strong contributions to labor productivity growth from both capital deepening and TFP growth, but it is the latter that was typically larger in the poorer countries which exhibited the fastest growth. This was not based to any significant extent on domestic R & D, but rather on a combination of technology transfer, structural shift away from agriculture, economies of scale, and more efficient utilization of factors of production. The transfer of "surplus labor" from small-scale family farms was an important part of the process (Crafts and Toniolo, 2008). External trade liberalization and the increased integration of the European market were factors that speeded up technology transfer and helped Europe to reduce the technology gap with the United States (Badinger, 2005; Madsen, 2007). Nelson and Wright (1992) also stressed the increased cost-effectiveness of American technology in Europe, the greater codification of

[20] Of course, in the neoclassical growth model an increase in exogenous TFP growth raises the growth rate of the capital stock; some implications of this point for growth accounting are explored by Hulten (1979). However, Abramovitz has in mind a richer story about 19th century American growth in which, *inter alia*, the great expansion of the domestic market resulting from technological change in transport leads to larger-scale and more capital-intensive methods of production.

[21] This section draws in part on Crafts and Toniolo (2008).

[22] Our examples are all taken from the late 20th century, so hopefully, the problems of inter-temporal comparability highlighted earlier will not be too severe. See however the caveats in the succeeding two footnotes.

Table 6.16 Accounting for growth during "economic miracles" (percent per annum)

(a) Sources of labor productivity growth

	K/L	HK/L	TFP	Y/L
Western Europe 1960–1970				
France	2.02	0.29	2.62	4.93
Germany	2.1	0.23	2.03	4.36
Italy	2.39	0.36	3.5	6.25
Spain	2.45	0.38	3.73	6.56
East Asia 1960–2003				
Korea	2.7	0.7	1.28	4.68
Singapore	2.86	0.46	1.2	4.52
Taiwan	3.04	0.54	2.16	5.74
Ireland				
1990–2003	0.49	0.26	2.24	2.99
USSR				
1928–1940	2		0.5	2.5
1940–1950	−0.1		1.6	1.5
1950–1970	2.6		1.4	4
1970–1985	2		-0.4	1.6
China				
1978–1993	2.1	0.4	3.9	6.4
1993–2004	3.7	0.3	4.5	8.5
India				
1978–1993	0.8	0.3	1.3	2.4
1993–2004	1.6	0.4	2.6	4.6

(b) Sources of output growth

	K	L = Employment + Education	TFP	Y
Western Europe 1960–1970				
France	2.24	0.42 + 0.29	2.62	5.57
Germany	2.13	0.06 + 0.23	2.03	4.45
Italy	2.2	−0.35 + 0.36	3.5	5.71
Spain	2.74	0.55 + 0.38	3.73	7.4
East Asia 1960–2003				
Korea	3.64	1.75 + 0.70	1.28	7.37
Singapore	4.03	2.18 + 0.46	1.2	7.87
Taiwan	3.97	1.74 + 0.54	2.16	8.41
Ireland				
1990–2003	1.7	2.24 + 0.26	2.24	6.44
USSR				
1928–1940	3.2	2.1	0.5	5.8
1940–1950	0.1	0.5	1.6	2.2
1950–1970	3.1	0.9	1.4	5.4
1970–1985	2.4	0.8	−0.4	2.8

(*Continued*)

Table 6.16 (*Continued*)

China				
1978–1993	3	1.6 + 0.4	3.9	8.9
1993–2004	4.1	0.8 + 0.3	4.5	9.7
India				
1978–1993	1.5	1.4 + 0.3	1.3	4.5
1993–2004	2.3	1.2 + 0.4	2.6	6.5

Notes: Ireland and USSR are GNP not GDP. Education is included in TFP growth for USSR.
Sources: Bosworth and Collins (2003, and web update); for USSR derived from Ofer (1987); and for China and India derived from Bosworth and Collins (2008), in each case assuming $\alpha = 0.35$.

technological knowledge, and increases in European technological competence based on increased investments in human capital and R & D. Overall though, this is clearly a case where TFP growth involved much more than technological progress.

The East Asian Miracle was quite different. Table 6.16 shows that TFP growth contributed relatively less, and capital deepening more, than in Golden Age Europe. Rapid growth of the capital stock was underpinned by increasingly high investment rates which reached around 35% of GDP in Korea and Singapore, around 10 percentage points higher than the average in 1960s Europe. East Asian growth was also notable for a very strong growth of labor inputs, underpinned by a "bonus" from the age structure effects of the demographic transition which (unlike in Western Europe) coincided with the growth spurt. Although East Asian countries were successful in importing technology, overall the developmental states of the region were better at mobilizing factor inputs than at achieving outstanding TFP growth (Young, 1995; Crafts, 1999).[23]

The Celtic Tiger was a very different animal from its Asian counterpart and contrasts quite strongly with Golden Age European growth (Crafts, 2009). Ireland's labor productivity growth was a good deal lower, mainly because of a small capital deepening component in an economy where investment was about 20% of GDP. TFP growth was strong but relied on ICT production which accounted for nearly two thirds of TFP growth during the 1990s (van Ark et al. 2003). In turn, this was based on Ireland's exceptional ability to attract FDI, especially from the United States: domestic R & D was only about 1.4% of GNP. Apart from ICT production, as Table 6.16 reports, the other outstanding feature of the Celtic Tiger was employment growth which far outstripped population growth. As unemployment fell, female participation rose, and emigration turned into immigration. Irish growth thus benefited from a very elastic labor supply (Barry, 2002).

The striking feature of catch-up growth in the Soviet Union is that, if standard growth accounting assumptions are adopted, it relied much more on "extensive growth." While

[23] Rodrik (1997) argues that TFP growth may be underestimated by standard techniques because σ was less than 1, given biased technical change and strong capital deepening. It is unclear how big this effect may have been.

the capital deepening contribution to growth in the Golden Age was similar to that in Western Europe, or a bit lower, TFP growth was decidedly inferior. Its contribution was very weak compared with countries like Italy with similar catch-up potential.[24] The problem with the Soviet growth model is that it ran into a rapidly rising marginal capital to output ratio, implying that the rate of capital stock growth delivered by a constant investment/GDP ratio fell steadily over time. The problem became acute when TFP growth ceased in the 1970s and further increases in the investment rate (which had doubled between 1950 and 1970 to 30%) became infeasible given the commitment to high defense spending.

Finally, we consider the "growth miracles" in two of the BRICs with rather different growth trajectories. China has experienced very rapid growth of real GDP per person since reforms began at the end of the 1970s. This has been based on impressive contributions from both capital deepening and TFP growth. The former has resulted from investment rates which are massive by historical standards, reaching well over 40% of GDP by the early 2000s. The latter has two components, technology transfer linked closely to FDI (Whalley and Xin, 2010), and increases in efficiency starting from the very low base of the Maoist economy. Here, de-collectivization of agriculture which led to a surge in TFP in the 1980s (McMillan et al. 1989) played a big part, initially. The rapid reduction in the state-owned enterprise share of GDP has also been a key component, and it is TFP growth in industry which has been most impressive. India experienced a productivity surge after the disappointing period of the so-called Hindu growth rate (Rodrik and Subramanian, 2005). Even so, capital deepening and investment rates have been well below those in China. So has TFP growth, although this strengthened appreciably after the Indian reforms of the early 1990s. The detailed comparisons in Bosworth and Collins (2008) show that TFP growth in the industrial sector in India has been very disappointing (averaging 0.6% per annum from 1978–2004 compared with 4.3% in China), while TFP growth in services has been strong—in 1993–2004 averaging 3.9% per annum compared with 0.9% in China.

6.5. GROWTH IN THE LEADER: THE UNITED STATES

The United States overtook Britain at the start of the 20th century in terms of real GDP per person and maintained its leading position throughout the "American century." By mid-century, the United States had become the clear technological leader

[24] It has been suggested that this may be an artifact of the methodology and that the USSR is better described in terms of a production function with a very low elasticity of substitution between capital and labor and thus severely diminishing returns to capital (Weitzman, 1970). Allen (2003) provides a convincing rebuttal of this claim, noting that the technological possibilities were similar in West and East and that there is clear evidence of massive waste of capital in the Soviet system, which implies that standard benchmarking is appropriate.

and had developed a very different "national innovation system" from that which had prevailed in 1900. Although other OECD countries, notably Japan, reduced the gap from the 1960s to the 1980s, the United States reasserted its leadership in the context of the ICT revolution. This section examines the foundations of this exceptional performance and considers American technological prowess using an endogenous innovation lens.

6.5.1 Technological Leadership

During the 20th century, the United States was in the forefront of the development of the most important new technologies, including: the internal combustion engine, electricity, petrochemicals, aviation, and ICT. In this era, technological progress increasingly became the result of systematic research and development based on formal science and engineering, and was associated much more with corporate research laboratories and public investment than with independent invention.

That said, there is a clear difference between the pre- and post-World War II eras (Mowery and Rosenberg, 2000). In the former period, the United States developed a formidable record in the commercial development of technologies which had typically originated from Europe. Already, by the interwar period, revealed comparative advantage in American exports was strongly correlated with research intensity (Crafts, 1989). In the latter period, United States' science and invention played a much bigger role as American universities became world leaders in academic research and federal funding for research soared in the context of the Cold War. These points are epitomized by the automobile, where the American contribution was the development of mass production, and the computer, where the transistor and integrated circuit were American inventions. Federal funding accounted for less than 20% of R & D in the 1930s but well over half on average from the 1950s through the 1970s. Germany had 41 (44) Nobel Prize winners prior to 1950 (1950 to present) compared with 27 (229) for the United States (excluding Economics and Peace).

American industrial research was built up during the first half of the 20th century by corporate investment in laboratories. Although independent inventors still accounted for 50% of patents in the late 1920s, down from about 80% at the start of the century, their share had fallen to only 25% by the 1950s (Nicholas, 2010). About three-quarters of industry-funded R & D was performed by firms with more than 10,000 employees in the early 1980s, when defense-related expenditure still accounted for about a quarter of all R & D. This picture had changed quite significantly by 2001 when the large firms' share had fallen to just over a half, the defense-related share was below 15%, and R & D was increasingly outsourced to specialist, smaller firms resembling—to some extent—an early 20th century landscape rather than the classic post-war American national innovation system (Mowery, 2009).

The transition to an economy with substantial investments in R & D and higher education is reflected in Table 6.17. This was clearly a very different technological leader than

Table 6.17 The knowledge economy in the United States

	R & D expenditure/ GDP (%)		R & D stock/ GDP (%)		Tertiary education/ person (years)
1920	0.2	1900–1910	0.03	1913	0.200
1935	1.8	1929	4.5		
1953	1.4	1948	13.0	1950	0.420
1964	2.9	1973	38.2	1970	0.674
1990	2.7	1990	47.7	1995	1.474
2007	2.7			2005	1.682

Sources: R & D expenditure: Edgerton and Horrocks (1994), Nelson and Wright (1992), and National Science Board (2012) R & D stock: Abramovitz and David (2001) Tertiary education: Barro and Lee (2012) and Maddison (1987).

Table 6.18 Productivity growth arising from US research, 1980s (percent of total in each country)

France	42
Germany	42
Japan	36
UK	33
USA	60

Source: Eaton and Kortum (1999, p. 558).

Industrial Revolution Britain. The size of these investments also marks the United States out from the rest of the OECD, especially in the third quarter of the century. Not only was R & D spending relative to GDP higher than anywhere else, but its absolute size loomed very large: as late as 1969, US R & D expenditure was more than twice the combined total of France, Germany, Japan, and the UK (Nelson and Wright, 1992). Similarly, the educational attainment of the American population far outstripped OECD rivals. In 1970, the next highest country (Denmark) had only about half the American tertiary education years per person. The dominant role of American (relative to all other countries') R & D as a source of productivity growth across the OECD is clearly shown in Table 6.18.

6.5.2 Explaining Technological Progress

There is a rich analytical narrative literature on the underpinnings of 20th century American technological progress, seeking to explain both its strength and its factor-saving bias. It is generally agreed that the geography of the United States in terms of the scale of the domestic market, the distances between major population centers, and the natural resource endowment, was an important influence, especially in the early part of the century. These features of American geography are seen as favorable to key

Table 6.19 Resource abundance

(a) US share of world totals (%)

	1913 output	1989 reserves	1989 + cumulative 1913–1989 production
Petroleum	65	3.0	19.8
Copper	56	16.4	19.9
Phosphate	43	9.8	36.3
Coal	39	23.0	23.3
Bauxite	37	0.2	0.5
Zinc	37	13.9	14.0
Iron Ore	36	10.5	11.6
Lead	34	15.7	18.1
Gold	20	11.5	8.6
Silver	30	11.7	16.3

Source: David and Wright (1997).

(b) Ratio of labor cost/hour to electricity cost/hour

	United Kingdom	United States
1909		8.8
1919		31.8
1929	14.8	44.6
1938	20.3	57.0
1950	35.6	157.5

Source: Melman (1956).

technological clusters such as those based on the internal combustion engine and the chemical industry (Mowery and Rosenberg, 2000). The rise of mass production in the later railroad era can be seen as "the confluence of two technological streams: the ongoing advance of mechanical and metal-working skills … focused on high-volume production of standardized commodities"; and the exploration and utilization of the mineral resource base (Nelson and Wright, 1992, p. 1938). Table 6.19 reports the concentration of world minerals output in the United States in 1913. It also implies that the country had been very efficient in discovering and developing minerals relatively early on. A relatively low price of electricity (Table 6.19) was conducive to the electrification of factories, which led to a surge in manufacturing productivity growth in the 1920s (David and Wright, 1999).

Over time, these influences became somewhat less important and the accumulation of human capital mattered more. The United States led the way in the expansion both of secondary and tertiary education. High school enrollment among 14–17 year olds rose from 10.6% in 1900 to 51.1% in 1930 and 86.9% in 1960, a time when only 17.5% of British 15–18 year olds were enrolled (Goldin and Katz, 2008). While about 5% of Americans born in 1880 went to college, nearly 60% of the cohort born in the 1960s did so. Throughout the third quarter of the century, average years of college education in the American adult population were a long way ahead of leading European countries (Barro and Lee, 2012). Even so, the rate of return to a year of college education in 1990 was only slightly below what it had been in 1915. This is symptomatic of the change in factor-using bias, from tangible-capital to intangible-capital using, between the 19th and the 20th centuries that Abramovitz and David (2001) detected.[25] Notable also, is that American leadership in electronics technology after World War II owed a great deal to the abundance of scientific and engineering human capital and federal research funding, rather than to the natural resource endowment.

Before World War II, relatively rapid American technological progress primarily reflected the capabilities of firms and thus the incentive structures that they faced. Endogenous innovation models point to several features of the American economy which were more favorable than in Europe at the time, and much more favorable than in Industrial Revolution Britain. These include a better system of intellectual property rights (Nicholas, 2010), a stricter anti-trust policy (Mowery and Rosenberg, 2000), a larger market potential (Liu and Meissner, 2013), and a significant fall in the costs of research as experimental science improved and the supply of specialized human capital expanded rapidly (Abramovitz and David, 2001).

This may be sufficient to explain the acceleration in technological progress but there is more to be said in terms of its direction. The experience of the American economy during the 20th century has been described as a "race" between education and technology (Goldin and Katz, 2008). Goldin and Katz highlight the development of a complementarity between advances in technology and the use of human capital that is visible from the early 20th century. The outcome of the race between increased demand for human capital as technology evolved, and increasing supply as the education system expanded, is captured in the behavior of the college wage premium (Table 6.20). Over the long-run the outcome was a photo-finish, but relative demand grew more strongly after 1960 and eventually outstripped supply after 1980.

The "directed technical change" model proposed by Acemoglu (2002) might be a suitable framework within which to analyze these trends. The key element of this model is its incorporation of a market size effect, as well as a relative price effect in the incentives that inform innovative effort. If the market size effect dominates, technological progress will be biased toward complementarity with a factor whose relative supply expands, rather than the opposite as would be expected on the basis of the ceteris paribus fall in its relative price. This induced innovation will in turn underpin the factor's rate of return through outward shifts in its demand curve.

[25] Abramovitz and David use a composite notion of intangible capital which includes both R & D and human capital; this is different from the definition in the recent growth accounting with intangibles literature reviewed in Section 6.4.

Table 6.20 Supply and demand for college educated workers and changes in the college wage premium, 1915–2005 (100 × annual log changes)

	Relative wage	**Relative supply**	**Relative demand**
1915–1940	−0.56	3.19	2.41
1940–1960	−0.51	2.63	1.92
1960–1980	−0.02	3.77	3.74
1980–2005	0.90	2.00	3.27
1915–2005	−0.02	2.87	2.83

Note: estimates assume the elasticity of substitution between college and high school graduates = 1.64.
Source: Goldin and Katz (2008).

6.5.3 Lessons from the ICT Revolution[26]

The Solow Productivity Paradox was announced in 1987 with the comment that "You can see the computer age everywhere except in the productivity statistics." A great deal of effort was subsequently devoted to explaining this (Triplett, 1999) and it was an important trigger for the literature on General Purpose Technologies. This developed models that had negligible or even negative impacts on productivity performance in their first phase but substantial positive effects later on. Indeed, a GPT can be defined as "a technology that initially has much scope for improvement and eventually comes to be widely used, to have many uses and to have many Hicksian and technological complementarities" (Lipsey et al. 1998, p. 43).

Table 6.21 compares ICT with the two other GPTs, electricity and steam, which are commonly placed in the pantheon on account of their impact on productivity growth in the leading economy of the time. The comparison reveals that the impact of ICT has been relatively big, and that it has come through very quickly. This new GPT is unprecedented in its rate of technological progress, reflected in the speed and magnitude of the price falls in ICT equipment reported in Table 6.21. The impact of ICT on the rate of productivity growth throughout 1973–2006 exceeded that of steam in any period and was already close to twice the maximum impact of steam by the late 1980s. Indeed, these estimates suggest that the cumulative impact of ICT on labor productivity by 2006 was about the same as that of steam over the whole 150-year period, 1760–1910.

A plausible inference seems to be that society is getting better at exploiting the opportunities presented by new GPTs. This may reflect a number of factors including more investment in human capital, superior scientific knowledge, improved capital markets, and greater support for R & D by public policy. Taking an historical perspective, the true paradox is that Solow's ICT paradox was regarded as such, given that by earlier standards the contribution of ICT to productivity performance in the American economy in the late 1980s was already stunning.

[26] This section draws in part on Crafts (2013a).

Table 6.21 GPTs: contributions to labor productivity growth (percent per annum)

Steam (UK)	
1760–1830	0.01
1830–1870	0.30
Electricity (USA)	
1899–1919	0.40
1919–1929	0.98
ICT (USA)	
1973–1995	0.74
1995–2006	1.45
Memorandum item: real price falls (%)	
Steam horsepower	
1760–1830	39.1
1830–1870	60.8
Electric motors (Sweden)	
1901–1925	38.5
ICT equipment	
1970–1989	80.6
1989–2007	77.5

Notes: Growth-accounting contributions include both capital deepening from use and TFP from production. Price fall for ICT equipment includes computer, software, and telecoms; the price of computers alone fell much faster (22.2% per year in the first period and 18.3% per year in the second period).
Sources: Growth accounting: Crafts (2002, 2004b) and Oliner et al. (2007). Price falls: Crafts (2004b), Edquist (2010), and Oulton (2012).

Table 6.22 Sources of labor productivity growth in the market sector, 1995–2005 (percent per annum)

	Labor quality	ICT K/hour worked	Non-ICT K/hour worked	TFP	Labor productivity growth
EU	0.2	0.5	0.4	0.4	1.5
France	0.4	0.4	0.4	0.9	2.1
Germany	0.1	0.5	0.6	0.4	1.6
UK	0.5	0.9	0.4	0.8	2.6
USA	0.3	1.0	0.3	1.3	2.9

Source: Timmer et al. (2010).

A very noticeable feature of the ICT revolution is that the United States exploited the opportunities much better than did European countries, generally speaking (Oulton, 2012). Table 6.22 shows that the ICT capital deepening contribution in the United States was about twice that in the European Union between 1995 and 2005. Indeed, this episode saw an ending of the long period of productivity catch-up achieved by Western Europe since the early 1950s.

A lens through which to examine this experience is to think about varieties of capitalism (Hall and Soskice, 2001). The core of this approach is based on a comparison between two ideal types, the co-ordinated market economy (CME) and the liberal market economy (LME), which comprise different environments in which firms operate. The purest cases of the CME and the LME are Germany and the United States, respectively (Schneider and Paunescu, 2012). Each of these economies can be thought of as having a different set of complementary institutions and, as a corollary of this, different comparative advantages in production, trade, human capital formation, and crucially, innovation. The LME is characterized by extensive equity markets and flexible labor markets, while the CME offers high employment protection and corporate governance that is based on monitoring by banks and an absence of hostile takeovers. LMEs place more emphasis on university education and less on vocational training, and are also more lightly regulated in terms of the standard indices calculated by the OECD.

Hall and Soskice (2001, pp. 38–39) argued that CMEs would be relatively strong at "incremental innovation, marked by continuous but small-scale improvements to existing product lines and production processes," while LMEs would be more successful at "radical innovation, which entails substantial shifts in product lines, the development of entirely new goods, or major changes to the production process." Empirical testing of claims about radical and incremental innovation poses considerable problems, but Akkermans et al. (2009) developed an approach based on patent citations, basically taking radical innovations to be those which are more highly cited. They found that the United States is indeed strongly specialized in radical innovation.

With regard to ICT, CMEs and LMEs might also be expected to differ in their abilities to exploit its opportunities since investment in ICT capital is much more profitable and has a much bigger productivity payoff if it is accompanied by organizational change in working and management practices and is therefore encouraged by low adjustment costs (Brynjolfsson and Hitt, 2003). The empirical evidence is that the diffusion of ICT has been aided by complementary investments in intangible capital and high-quality human capital, but weakened by relatively strong regulation in terms of employment protection and regulations that restrict competition, especially in the distribution sector (Conway et al. 2006).

ICT is a technology that is very well suited both to management practices in American-owned companies (Bloom et al. 2012) and the economic environment in the United States. Perhaps the more general message is that, when a disruptive GPT appears, American institutions are at an advantage.

6.6. THE ECONOMIC HISTORIAN'S VIEW OF CATCH-UP

In Section 6.3 we saw that while some regions—notably Japan and the East Asian Tigers—caught up on the world technological frontier in spectacular fashion after 1945, others—notably Latin America and Africa—did not. The growth miracles of the 20th century, including not only the Japanese and Tiger experiences, but Western Europe during the Golden Age, China from the late 1970s onwards, or Ireland during the 1990s, were above all convergence miracles. Economic historians have known, since the work of Gerschenkron (1962) and even before, that backwardness can sometimes lead to rapid growth. The further behind the technological frontier a country is, the faster is its potential growth, since by importing the latest technologies and machinery it can improve its total factor productivity much more rapidly than an economy closer to the frontier. As Gerschenkron (1962, p. 8) put it, "Borrowed technology, so much and so rightly stressed by Veblen, was one of the primary factors assuring a high speed of development in a backward country entering the stage of industrialization." And indeed, industrialization, or the modernization of existing industries, was at the heart of the best-known 20th century growth miracles.

The problem is that while being economically backward implied a potential for rapid catch-up growth, it also implied obstacles to realizing that growth—since otherwise the country or region concerned would not have been backward in the first place. Economic historians have thus also always stressed that there is nothing inevitable or automatic about catch-up. This section will present some general insights from economic history relating to the question of whether countries are able to exploit catch-up opportunities or not. Sections 6.7–6.9 will then go on to apply these insights to well-known episodes of success followed by disappointment, success up to now, and failure, respectively.

6.6.1 Catch-Up is Not Automatic

The logic that backward countries should be able to grow more rapidly than the rich, by importing best-practice technologies, is powerful, but we know that in practice poor countries do not always grow more rapidly than the rich. If they did, then we would not regard those instances where convergence has most visibly been at work as growth miracles. We have seen that some groups of countries have managed to converge on the US technological frontier, while others have not. We have also seen that convergence was widespread in some periods, particularly the 1950–1973 Golden Age, while in other periods there was little or no convergence. Indeed in some periods divergence was more the rule, for example, during the late 19th century when the United States pulled further ahead of most of the rest of the world. Looking at variations in growth among those countries chasing the United States frontier, there have been contrasting experiences of convergence and divergence, depending on the groups of countries and time periods being considered. Absolute convergence characterized the rich economies as a group in the four decades since World War II, but there was no worldwide tendency during these years for poorer countries to grow more rapidly than the rich (Abramovitz, 1986; De Long, 1988; Barro, 1991).

Why does convergence happen sometimes but not always, and in some countries but not others? And why does it sometimes cease altogether, after promising beginnings? The logic of convergence suggests that it should be self-limiting: as countries catch up on the technological frontier, the scope for further catching up diminishes. As workers leave low-productivity agriculture for high-productivity service and manufacturing jobs, the pool of workers who can be similarly redeployed diminishes. One would thus expect converging economies to continue catching up on the lead economy, but at a diminishing rate over time, as in Lucas (2000, 2009). Yet, Western European convergence ceased after the first oil crisis, at a relative GDP level of only 70%, while in the former Soviet empire, and Southwest Asia, convergence not only halted at the same time, but was replaced by two decades or more of sharp divergence. Japan's convergence was also succeeded by divergence, beginning in the 1990s. More generally, there is evidence that countries often experience growth slowdowns after phases of rapid growth that are much sharper than would be expected on the basis of convergence logic alone. Eichengreen et al. (2012) found that the probability of such rapid slowdowns peaks at per capita GDP levels of about $17,000 in 2005 international prices, and that the probability is higher after periods of rapid economic growth.

Many economic historians have written about why convergence may not take place, the advantages of backwardness notwithstanding. Gerschenkron (1962, p. 8), whose major focus was Europe, argued that the major obstacles were "formidable institutional obstacles (such as the serfdom of the peasantry or the far-reaching absence of political unification," as well as (in some countries) a lack of natural resources. True, backward countries also lacked the prerequisites for growth that had been built up in Britain over the course of many decades and even centuries, but for Gerschenkron this handicap could be surmounted by means of institutional substitutes such as universal banks or a developmental state.

Gerschenkron was writing in the early 1960s, at a time when the Soviet Union and its allies were still converging rapidly on the United States, and decolonization with its ensuing policy experimentation was still in its infancy. By the 1980s, greater scepticism regarding the ability of backward states to engineer convergence seemed in order. Abramovitz (1986, pp. 387, 390, 393, 397) lists several reasons why convergence may not take place. Countries may lack the "social capability" required to realize their catch-up potential; the global economy may not be operating in a way that facilitates technological transfer; there may be obstacles to structural change within the backward economies; short-run macroeconomic policies may not encourage investment, with long-run consequences; best-practice technologies may not be appropriate for developing economies' size or factor endowments; and major shocks such as war may disrupt the convergence process. We briefly review each of these arguments in subsequent sections.

6.6.2 The Consequences of Directed Technological Change

Technological change is not exogenous, but an endogenous response to economic conditions. This can make it difficult for countries to catch up on the technological frontier, irrespective of whatever institutions they may have or which policies they adopt. Frontier technologies may have been invented with conditions in the leading economy in mind, and may therefore not be easy or profitable to adopt in poorer countries.

This possibility has been raised by growth economists such as Basu and Weil (1998), and Acemoglu and Zilibotti (2001), in debates about appropriate technology, but it is also a long-standing theme of economic historians. The argument that technologies are invented so as to take advantage of local factor endowments is most often associated with Habakkuk (1962), who as we have seen, argued that it was high American wages, due to an extensive land endowment, that explained the relatively labor-saving nature of US mass production technology. The evidence suggests that since the late 19th century, improvements in the production function have been concentrated at the capital to labor ratios at which rich countries operate. For example, there was a big increase in output per worker at capital to labor ratios between $15,000 and $20,000 (1985 prices) between 1939 and 1965, but no further improvement in recent decades. Indeed, at very low capital to labor ratios, output per worker in 1990 appears to have been no higher than in 1820 (Allen, 2012). This is symptomatic of a pattern of directed technical change where advances are made in accordance with the incentives provided by market conditions in rich countries, especially the United States.

The possibility then arises that relative factor prices in less developed economies made the new technologies unprofitable. Gerschenkron (1962, pp. 8–9) raised the issue, noting that "The industrialization prospects of an underdeveloped country are frequently . . . judged aversely, in terms of cheapness of labor as against capital goods and of the resultant difficulty in substituting scarce capital for abundant labor." He also noted that this argument flies in the face of the opposite argument that low wages give developing countries a powerful competitive advantage. But he went on to dismiss the argument in the context of 19th century Europe, on the grounds that "a stable, reliable and disciplined" labor force was scarce, rather than abundant, in backward economies where people were still close to the land. Indeed, he claimed that this fact gave entrepreneurs in countries like Russia an incentive to import technologies that were as modern, efficient, and labor-saving as possible.

In contrast, as we have seen, Allen (2009) argued that it was rational for other countries to not immediately adopt the new technologies of the British Industrial Revolution, implying that Britain initially forged ahead while others fell behind. It was only with time, as the new technologies became more productive, that they became profitable to adopt elsewhere. By the late 19th century, however, Britain was no longer the leading innovator, and the question was whether American inventions were suitable for British conditions or not. British entrepreneurs have often been criticized for failing to adopt the latest technologies—for example, cotton manufacturers were slow to adopt ring spinning, preferring to stick with mule spinning, while soda manufacturers were slow to abandon the Leblanc process for the superior Solvay process. Magee (2004) surveys an abundant literature that argues that British entrepreneurs were in fact responding rationally, not only to British relative factor prices (skilled workmen were cheaper than in America, and natural resources were dearer), but also to different (and in particular less homogenous) demand conditions. If Lancashire cotton manufacturers used mules, this was because they produced more fine yarns, and more yarn for export, than their American counterparts, and mule spinning was superior on both counts (Leunig, 2001). More generally, fragmented demand and skilled labor made it rational for British manufacturers to eschew resource-intensive and labor-saving mass production techniques, and adopt "a more flexible form of production, based on general purpose machinery, skilled labor and customized demand" (Magee, 2004, p. 95). Similar considerations can explain why British firms did not adopt Chandlerian organizational forms during the same period (Harley, 1991).

We will consider the British case in more detail below, merely noting that if frontier technologies do not correspond to the needs of developing countries, then those countries may fall further behind the leaders for perfectly rational reasons, with no "failure" being necessarily involved. What might reverse such a trend? Educational policies are one obvious candidate. Another is late 20th century globalization, which Wright (1990) argues was a major turning point, in that it transformed mineral resources from being endowments to commodities, available to all countries at roughly equal prices. The implication is that resource-intensive American technologies now became potentially easier to implement around the world. Similarly, the opening of the rest of the OECD to international trade meant that American mass production techniques could more easily be adopted elsewhere, and the process of convergence itself strengthened this tendency by further increasing the size of overseas markets. Finally, postwar US technological strength in sectors like semiconductors were based on the expansion of scientific education and research; and research and development, which could be replicated abroad, especially given the inherently international nature of scientific activity (Nelson and Wright, 1992; Abramovitz and David, 1996). As Alice Amsden (1989, p. 7) put it, "Although technology remained … idiosyncratic even in basic industries, higher scientific content increased its codifiedness or explicitness, making it more of a commodity and hence more technically and commercially accessible and diffusible from country to country." Multinational corporations made technology even more diffusible. For all these reasons, US frontier technologies could now in principle be more easily implemented abroad, at least in the relatively advanced economies of Europe and Japan, than had been the case before. The extent to which they were actually implemented presumably depended on a variety of other factors, some of which will be considered below.

6.6.3 Catch-Up and Social Capability

According to Abramovitz (1986), tenacious social characteristics could inhibit countries from importing best-practice technology, and one would therefore only expect poorer countries to grow more rapidly than richer ones if these social characteristics, which he termed "social capability," were roughly similar. Rapid growth was thus most likely when countries were "technologically backward but socially advanced" (p. 388). Abramovitz (1989, pp. 200–201) considered both these conditions to have been present in Europe and Japan after World War II. Both regions had generally well-educated populations, and were well endowed with scientists and engineers, who were increasingly influential within industry. This helped in implementing new technologies invented abroad. Both firms and governments promoted research and development. Large corporations were becoming increasingly well managed. The resumption of international trade, air travel, the press, and American cooperation facilitated the importation of technical knowledge. Such attributes of backwardness as a large agricultural population could be turned into an advantage, since agricultural productivity growth facilitated the release of labor to new and growing sectors of the economy. Other aspects of social capability included openness to change and competition, which were necessary as rapid structural change was part and parcel of the catch-up process. Abramovitz cites Olson's (1982) view that the war itself, by sweeping aside existing vested interests, helped create a *tabula rasa* that facilitated such change.

Social capability can be thought of as being equivalent to the parameter μ_m in Schumpeterian growth theory (Aghion and Howitt, 2006). This refers to the extent to which countries' growth rates are boosted by virtue of their distance from the technological frontier. Abramovitz largely discusses education when referring to social capability, but he also mentions institutions, and these have been a major focus of economic historians seeking to understand different countries' growth experiences. A standard list of institutions that might matter for growth includes "the security of property rights, prevalence of corruption, structures of the financial sector, investment in public infrastructure and social capital, and the inclination to work hard or be entrepreneurial" (Sokoloff and Engerman, 2000, p. 218). The degree and nature of unionization; attitudes toward cartels and competition; social welfare and taxation systems; and the general nature of government involvement in the economy, could also be added to this list. What matters from the perspective of convergence is the incentive structures shaped by policy and institutions which influence the diffusion and assimilation of new technology in follower countries by, for example, determining the expected profitability of innovation, or by mitigating or exacerbating agency problems in the firms which have to invest in the new technologies. Economic historians emphasize that we do not inhabit a "one-size-fits-all" world, and that optimal institutional design may therefore vary according to the degree of backwardness, the technological era, etc.[27]

[27] This is a key point made in Aghion and Howitt (2006).

Gerschenkron believed that the institutional mix could adapt to meet the needs of the backward country seeking to catch-up. Where capital markets were not as well functioning as in the mature British economy, and where entrepreneurship was scarce, universal banks mobilizing large amounts of saving and providing not only capital but also entrepreneurial guidance for heavy industries, could fill the void. Where the economy was so backward that this was not an option, as in Russia, the state could step in instead. Gerschenkron believed that in the boom years immediately prior to World War I, universal banks played a more important role in Russia than they had done during the boom of the 1890s, reflecting the fact that Russia was no longer as backward as she had been a quarter of a century earlier.[28] If institutions can adapt in this manner, then although they may be crucial for economic performance, they are also endogenous; and endogenous variables do not make convincing explanatory variables. The view that historical institutions were efficient solutions to economic problems characterized much early cliometric work on the subject, including that of Douglass North (e.g. North and Thomas, 1973). However, institutions can arise for other reasons as well: for example, they could be the result of accident, followed by path dependence; or of cultural belief systems; or of distributional conflicts (Ogilvie, 2007). A frequent theme in modern economic history is that particular institutions may have originated as efficient solutions to context-specific problems, but that they can also be politically hard to reform and subject to path dependence (North 1990). Thus, when the context changed, the institutions stayed the same, and turned from being a help to a hindrance. We will see examples of this kind of logic at work in the case studies below.

Of particular interest is the possibility that institutions which help countries catch-up on the technological frontier may no longer be appropriate once countries have converged. For example, Rosenstein-Rodan (1943) argued that intersectoral complementarities could mean that modern industrialization might only get going if it happened across a broad front. For Gerschenkron (1962, pp. 10–11) this was one explanation for why, in his view, the transition to industrialization tended to happen in a dramatic and even discontinuous fashion (a claim which subsequent quantitative research has however cast doubt on—see Sylla and Toniolo, 1991). Such "big push" arguments naturally suggest a potentially important coordinating role for the state (Murphy et al. 1989) in the early phases of industrialization, but it is far from clear that such state involvement would make sense when countries have reached the frontier, and the question is no longer how to import and implement existing technologies, but to develop new ones. More radically, Baldwin (forthcoming) argues that modern multinational-led globalization and what he refers to as the "second unbundling" has destroyed big push arguments for state-led

[28] Gerschenkron's account is controversial. Sylla (1991, pp. 52–53) reviews evidence which suggests that banks played a larger role in the 1890s industrialization than Gerschenkron had allowed, while Gregory (1991) argues strongly that state involvement was by no means as beneficial as Gerschenkron had believed it to be.

industrialization that were valid not so long ago: small developing countries can now begin to industrialize by colonizing individual niches in global supply chains. Of course, this argument relies on globalization being sustained in the future, which is something that can never be taken for granted (Findlay and O'Rourke, 2007).

Such arguments suggest that institutional reform may be needed as countries progress economically. Unfortunately, if institutions are path dependent then such reform may not always proceed as smoothly as Gerschenkron believed had been the case in pre-1914 Russia.

6.6.4 Geography

It is striking that income levels around the world are highly spatially correlated. Since these income levels are the result of long-standing historical processes, it is not surprising that there was a degree of regional clustering in the timing of the shift to modern industrialization. Signs of rapid industrialization can be found in Eastern Europe and Latin America from the 1870s onwards, and in parts of Asia by the end of the 19th century (Bénétrix et al. 2013). Why do we observe such geographical correlations in the data, and what do they imply for convergence?

One possibility is that countries with similar resource endowments tend to be located close to each other, and thus end up with similar growth experiences and incomes in the long run. This may be because geographical conditions and resource endowments matter directly for growth (Sachs and Warner, 1997), or because they matter indirectly via their impact on institutions (Easterly and Levine, 2003). Economic historians have long argued that institutions may respond to endowments: for example, Domar (1970) argued that forced labor systems such as serfdom and slavery were a predictable outcome in labor-scarce and land-abundant societies, since in the absence of such exploitation the return to owning land was zero, which was not in the interests of would-be aristocrats. Domar is cited by Engerman and Sokoloff (1997), who argue that institutional differences based on underlying differences in geography, rather than superior culture, were the main reason why the United States and Canada eventually became so much richer than other countries in the Americas.[29] Brazil and the Caribbean were ideally suited to producing crops such as sugar, and thus developed slave-based economies, societies, and political institutions, irrespective of what European powers colonized them. British, French, Dutch, Portuguese, and Scandinavian sugar colonies all developed highly unequal, slave-based societies, and remained highly unequal even after the suppression of slavery. Spanish American colonies developed by exploiting Native American workforces in both agriculture and mining, and were also highly unequal. The result was political institutions and economic policies designed to maintain elite privileges: restricted franchise, barriers

[29] Acemoglu et al. (2001) suggest an alternative mechanism through which geography may have influenced development via its impact on institutions. For a discussion, see Albouy (2012).

to European immigration, limited investment in education, conservative taxation systems, and expensive access to patent protection, to name but a few (Sokoloff and Engerman, 2000; Sokoloff and Zolt, 2007).

It is important not to romanticize the US or Canadian experiences, to ignore the treatment meted out to their own native American populations, or to forget that universal suffrage was only attained in the United States as late as the 1960s. This last fact is a shocking reminder of the corrosive effect of inequality and racial segregation on the quality of political institutions. But inequality was relatively low among US and Canadian whites, and because whites made up a relatively large share of those two countries' populations (since sugar was not an important crop even in the US, and since native Americans were so few in number by the 19th century), the net result was societies that were relatively egalitarian in the aggregate. This in turn encouraged not only inclusive institutions for whites, but directly stimulated economic growth by encouraging commercial activity and the development of mass marketing. Canada and the US also invested heavily in public education, funded out of local taxes on income and wealth, made the patent system cheaply accessible to a broad range of people, and promoted economic growth in a variety of other ways. Engerman and Sokoloff (1997) sketch a story in which a rapidly growing and relatively equal population boosted 19th century US growth, by promoting a Smithian process of division of labor and exploitation of scale economies, and by encouraging market-oriented innovation. This route to prosperity was barred to Latin American economies whose institutions perpetuated historical patterns of inequality.

Such arguments explain geographical correlations in GDP by pointing to geographical correlations in resource endowments. They would work even if each country were isolated from its neighbors. But it is also possible that geographical correlations in economic outcomes are shaped by interactions between countries located closer or further together. One possibility, emphasized by the new economic geography, is that market access matters for income levels and, in particular, for the location of industry across the world (Krugman and Venables, 1995). Redding and Venables (2004) find that GDP per capita is strongly related to market access and proximity to suppliers, and argue that this can retard convergence in per capita incomes and wages. Another possibility is that the diffusion of technology itself is a decreasing function of distance (Comin et al. 2013). Economic historians have increasingly been adopting such a geographical perspective in recent years (Crafts and Venables, 2003).

6.6.5 Events, Dear Boy

Easterly et al. (1993) show that shocks are as important as fundamentals in explaining countries' decadal growth performances. Nor can we ignore the impact of shocks over the longer run, as the disastrous performance of the interwar period shows.

Economic theorists such as Lucas (2009) understandably tend to construct models in which certain patterns—such as the gradual diffusion of economic growth across the globe—can be expected to apply in the absence of such non-economic forces as "wars, breakdowns of internal order, and misguided ventures into centralized economic planning" (p. 23).

Economic history, by contrast, focuses heavily on such events, for a number of reasons. First, while economic historians seek general explanations, like other economists, they also want to understand what happened in specific countries and at specific times, like other historians. For example, while the relative decline of the Caribbean may have been in part due to the institutional legacies mentioned above, it was also surely due to the British-led suppression of slavery and the development of beet sugar production in Europe. This tension between the specific and the general is one of the defining features of the field. Second, economic historians are trained to think in terms of path dependence (David, 1985), and sufficiently major crises can have very long-term effects, for example because of their impact on subsequent policy choices (Buera et al. 2011). Third, economic history is an inherently interdisciplinary subject: as Hicks (1969, p. 2) put it, "A major function of economic history … is to be a forum where economists and political scientists, lawyers, sociologists, and historians—historians of events and of ideas and of technologies—can meet and talk to one another." As such, economic historians are more likely than other economists to try to understand the causes of non-economic shocks, and to integrate them into their analyses.

Although it is beyond the scope of a chapter such as this, it would be impossible to tell a convincing story of 20th century growth without describing the major shocks that defined the century, and tracing out their consequences. The two world wars, the Great Depression, decolonization, the oil shocks of the 1970s, and the Cold War and its ending, all had major effects on regional growth patterns during our period. This is evident from Figure 6.2, which shows major breaks in regional performance relative to the United States coinciding with the world wars, the onset of the Depression in 1929, and the first oil crisis of 1973.

World War I not only brought to an end the period of globalization that preceded it, but changed the economic and geopolitical landscape in ways that defined the rest of the 20th century. It led to the collapse of the German, Austro-Hungarian, and Russian empires, spawning a host of new nation states in Europe; it led to the Russian revolution of 1917, which had an enormous impact on the economies of not only the USSR, but (after 1945) of Eastern Europe and China as well; it permanently weakened the British economy, leaving the interwar world without a hegemon able and willing to provide global public goods (Kindleberger, 1973); it led to major imbalances in the structure of international trade, and to war debts and reparations, which would cast a dark shadow over the interwar period's flawed attempt to recreate a globalized world based on the gold standard. Most accounts of the Great Depression begin with these and other legacies of the conflict (Eichengreen, 1992), while the Depression and German post-war resentments combined to produce the election of Hitler, and ultimately the outbreak of World War II.

That conflict in turn cemented the relative decline of Europe, and paved the way for both the Soviet-US duopoly which lasted until the end of the 1980s, and decolonization throughout Asia and Africa. The historical association between globalization and European imperialism, the distrust of markets which naturally flowed from the disastrous experience of the interwar period, and the spread of communism, all predisposed the leaders of newly independent countries to pursue state-led growth policies, often based on import-substituting industrialization. Western Europe and North America developed a variety of more or less social democratic economies and societies, by and large (but by no means exclusively) using markets to generate wealth, and using the state to redistribute it, provide safety nets, and correct market failures. Elsewhere, the reaction against markets was far more severe. It was only reversed after the poor economic performance of the 1970s, which in turn had at least something to do with the oil shock which followed the Yom Kippur War of 1973—yet another event with important long-term consequences. The policy transition accelerated after the collapse of the Soviet Union, and is still ongoing. As historians, we would not want to bet that there will not be another policy reversal in the future, as a result of further unexpected shocks to the system.

The reason for dwelling on such major events and their consequences is to make the point that economic historians do not just focus on deep historical legacies and institutional path dependence. If these were all that mattered, then one would not expect to see more or less simultaneous reversals in both economic policy regimes and growth experiences across countries with very different histories and institutional legacies. The interwar growth experience was bad across all major regions of the world, while the Golden Age was good. This had a lot to do with the specific historical circumstances at work in both periods, and circumstances change. Change as well as continuity has always concerned historians, since both matter in the real world.

6.6.6 Openness and Other Economic Policies

Previous subsections have looked at some of the difficulties that countries can face in their attempts to join the convergence club—difficulties which may be difficult to overcome, since individual countries cannot easily change the appropriateness of foreign technology, or their geography, or the international geopolitical environment, or even (perhaps) their own institutions. However, countries can change their economic policies for better or worse. The question is whether such policy transitions can produce better growth performances, and if so, which policies are good for growth.

Most attention in the literature has focused on the impact of market-friendly economic policies in general, and trade policy in particular. A key reference is Sachs and Warner (1995), who produce an index of trade openness (subsequently updated in Wacziarg and Welch, 2008). Sachs and Warner used this index to study the impact of trade policy between 1970 and 1989. They found that openness was associated with higher growth, and that unconditional convergence characterized the experience of open economies but not of closed economies. Following discussion of the way in which this index was constructed (e.g. Rodríguez and Rodrik, 2001), several subsequent researchers (e.g. Buera et al. 2011) have preferred to interpret this index as indicating whether a country has adopted generally market-friendly policies or not.

This is how the index is used by Hausmann et al. (2005), who study the characteristics and determinants of growth accelerations from the 1950s to the 1990s. They find that while market-friendly economic reforms are a statistically significant predictor of sustained growth accelerations, they are not a quantitatively reliable predictor. Most pro-market reforms do not lead to such accelerations, and most accelerations are not preceded by such reforms. While their study finds that growth accelerations are difficult to predict, it also finds that they tend to have certain characteristics in common. In particular, growth accelerations are associated with higher investment rates, increases in trade, and real exchange rate depreciations. We will see examples of this below. We also note that the two growth accelerations that have mattered most for human welfare in recent decades—those in China and India—clearly seem to be related to market-friendly policy reforms.

There has been a vigorous debate about whether openness to trade is associated with faster growth or not, with Rodríguez and Rodrik (2001) among others strongly questioning the Sachs and Warner result. A recent contribution (Estevadeordal and Taylor, forthcoming) finds that lower tariffs on imported capital goods were associated with higher growth between 1975 and 2004, and it is probably fair to say that most economists assume that openness and growth go hand in hand today. Economic historians, however, tend to emphasize that the "right" policies may be context-specific, and may have varied over time. Clemens and Williamson (2004) find that tariffs were positively correlated with economic growth during the interwar period: perhaps the benefits to individual countries of maintaining open trade policies were lower in an environment where demand was depressed, and other countries had closed their own markets. Policies that were collectively costly may have been individually rational in such a context.

O'Rourke (2000) finds that tariffs and growth were positively correlated in the late 19th century as well, controlling for country fixed effects, in a sample of 10 relatively well-developed economies. A lack of aggregate demand was not a problem in this period, so unless the correlation is spurious we need another explanation. The growth-promoting externalities associated with industry would seem to offer one such explanation: as is well known, the United States industrialized behind very high tariff barriers during this period, and Germany and other continental European countries similarly protected their heavy industry. The fact that it was industrial tariffs that were associated with high growth, rather than agricultural tariffs, adds weight to this interpretation (Lehmann and O'Rourke, 2011). But even if the argument is correct, it does not follow that such policies would have worked in even less developed countries at the same time, or in the same countries in later periods. There is thus an important potential role for country histories in elucidating the impact of economic policies on growth, since panel growth regressions which estimate effects that are consistent across countries or over time may be seriously misleading.

6.7. CASE STUDIES I: INITIAL SUCCESS, SUBSEQUENT DISAPPOINTMENT

In the following three sections we explore several case studies that illustrate some of the themes of this chapter in slightly greater depth. We begin by looking at two cases where initial growth successes were succeeded by disappointment. The first is Western Europe, which converged strongly on the US during the Golden Age. Since the 1970s, however, convergence in GDP per capita has come to a halt. The second is the United Kingdom, which pioneered the transition to modern economic growth, but whose 20th century performance was much more disappointing, especially during the Golden Age.

6.7.1 The European Golden Age and the Subsequent Slowdown[30]

We have seen that Western Europe achieved its highest ever growth rates, roughly 4% per annum, during the Golden Age which lasted from 1950 to 1973. The period between the Second World War and the first oil crisis has subsequently passed into folk memory as the *Trente Glorieuses* (glorious thirty) or the *Wirtschaftswunder* (economic miracle). Eastern Europe and the former USSR also grew rapidly, although somewhat less so than Western Europe, when in a convergence perspective they should have grown more quickly. Relative to other miracles, for example in East Asia, a relatively large share of Western Europe's growth was due to TFP improvements, suggesting a large role for technological catch-up and structural change (Crafts and Toniolo, 2008). What explains the European growth miracle and the subsequent slowdown?

Western Europe's per capita GDP stood at just 31% of the American level in 1945. Austria's GDP had regressed to its 1886 level, France's to its 1891 level, and Germany's to its 1908 level (Crafts and Toniolo, 1996, p. 4). Rapid growth as a result of post-war reconstruction is hardly surprising. However, pre-war levels of GDP were restored by 1951 at the latest. Strikingly, in that year, Western Europe's relative GDP stood at only 47%.

The potential for catch-up growth seems obvious, and it seems even more obvious when a number of supplementary factors are taken into account. First, American technology and European conditions were more technologically congruent than they had been in earlier decades, as natural resources and larger markets became more easily available to European firms (Abramovitz and David, 1996). European economic integration would make both even more easily available as the 1950s progressed. Second, Western Europe possessed a high level of social capability: a generally well-educated population, and a history of well-functioning political and market institutions. According to Abramovitz

[30] This section draws in part on Crafts (2013a).

and David the war further strengthened Western Europe's social capability, by sweeping aside lingering *Ancien Régime* attitudes toward such things as mass education, mass production, industry, and economic growth. Finally, the disastrous experience of Depression and war gave a powerful impetus to European integration, and thus to the reversal of the protectionist policies of the interwar period.

High levels of social capability in an economically backward society—impoverished sophistication, in Sandberg's (1979) memorable phrase—should be optimal for achieving economic growth, especially if that society is engaged in a process of economically integrating disparate national economies into a continental common market. A further factor emphasized by many economic historians (Kindleberger, 1967; Broadberry, 1997; Temin, 2002) is the large agricultural workforces in most European countries that could be redeployed to higher productivity non-agricultural occupations. Such structural change accounted for a large share of Golden Age labor productivity growth (Crafts and Toniolo, 2008). And so the European growth miracle can be comparatively easily explained in terms of the convergence framework outlined earlier—which is hardly surprising, since it was this European experience that largely gave rise to the convergence paradigm in the first place. Not only did Western Europe as a whole grow more rapidly than the United States, but there was strong unconditional income convergence within Western Europe as well. And yet there is more to be said about this episode, for at least two reasons. First, economic growth in some countries was a lot faster than would be expected on the basis of post-war reconstruction and convergence alone (Crafts, 1992a, Table 6, p. 401). Second, some countries did a lot better during the Golden Age than others, even once their initial incomes have been taken into account.

Eichengreen (1996) shows that growth was positively correlated across Western European countries with both investment and export growth, consistent with Hausmann et al. (2005). According to Eichengreen, who further develops his argument in Eichengreen (2007), high levels of investment and trade were sustained by a variety of domestic and international institutions. Domestic institutions, which can be collectively described as corporatist, ensured that workers moderated their wage demands so that profits were high, and as a quid pro quo ensured that profits were reinvested rather than being paid out as dividends, thus ensuring higher wage growth in the future. Worker representation on firm's boards helped ensure that employers did not defect from this mutually beneficial equilibrium; centralized wage bargaining overseen by government, which had both sticks and carrots at its disposal, ensured that workers did not defect. The welfare state was one way in which workers were compensated for wage moderation in the short run. The result was high investment, capital deepening, high rates of TFP growth, and an economic miracle.

The international institutions that mattered were those associated with European integration: the European Payments Union; the European Coal and Steel Community, the European Economic Community; and EFTA. These facilitated the resumption of multilateral trade in Europe, which was necessary both for standard efficiency reasons, and so that firms could be ensured of foreign markets when making their investment decisions. European international integration was one of the demands of the US government, which used Marshall Aid as a lever to obtain this and other market-friendly structural reforms (DeLong and Eichengreen, 1993). Crafts and Toniolo (2008) portray the Golden Age as a period in which there was scope for growth simply by undoing the policy mistakes of the interwar period.

If Eichengreen is right, then investment, trade, and growth should have been higher in countries which adopted appropriate domestic institutions, and liberalized their trade earlier. Ireland is one example of a country which only liberalized its trade in the late 1950s and early 1960s, and which, like the UK, had a more fragmented and less corporatist trade union structure. Both countries performed relatively disappointingly during the Golden Age. On the other hand, Belgium, West Germany, the Netherlands, and Scandinavia all had relatively corporatist systems of industrial relations, and liberalized their trade policy relatively early. The Eichengreen argument is thus a priori plausible, although econometric testing of the hypothesis is difficult (Crafts, 1992b).

What explains the post-1973 growth slowdown? The arguments outlined above suggest that to a large extent this was inevitable, as Europe caught up to the technological frontier, and the pool of agricultural workers who could be redeployed gradually vanished. While this is surely a large part of the story, it is not the whole story, for several reasons (Crafts and Toniolo, 2008). First, while GDP per capita convergence on the US ceased in the 1970s, labor productivity convergence continued until some point in the 1990s. The difference is due to diverging trends in hours worked in the two continents: how to interpret this remains unclear (Blanchard, 2004; Prescott, 2004; Alesina et al. 2006). Second, the distributional conflicts associated with the oil crises of the 1970s may have undermined the viability of the cooperative political institutions which Eichengreen believed had promoted growth during the Golden Age. Third, even if these institutions had remained as viable as they had been before, it is unclear that they were well adapted to a new era in which growth based on importing best-practice technology from abroad was no longer as easy as it had been when Europe had been more backward (Eichengreen, 2007; Aghion and Howitt, 2006). Rather than mobilizing large amounts of capital to mass produce well-understood technologies that had been developed elsewhere, the problem was now how to innovate: the argument is that this required more competitive product markets, different methods of finance, and alternative training systems.

The growth rate of real GDP per hour worked increased in the United States between 1973–1995 and 1995–2007 from 1.28% per year to 2.05% per year. In contrast, in the EU15 it fell from 2.69% per year to 1.17% per year. The rate of labor productivity growth fell in most European countries: in Italy and Spain it was below 1% per year after 1995. By contrast, Sweden saw a productivity revival while for part of the period Ireland continued to be a Celtic Tiger, and both countries exceeded the American productivity growth rate.

So while there was falling behind in productivity performance on average, there was also considerable diversity in European performance.

The acceleration in American productivity growth was underpinned by ICT. As we have seen, historical comparisons reveal that the impact of ICT has been relatively large and that it has come through very quickly. The main impact of ICT on economic growth comes through its diffusion as a new form of capital equipment rather than through TFP growth in the production of ICT equipment. This is because users get the benefit of technological progress through lower prices, and as prices fall more of this type of capital is installed.[31] The implication is that ICT has offered Europe a great opportunity to increase its productivity growth. However, as we saw in Table 6.22, European countries have been less successful than the United States in seizing this opportunity.

The empirical evidence is that the diffusion of ICT has been aided by complementary investments in intangible capital and high-quality human capital, but weakened by relatively strong regulation in terms of employment protection and restrictions to competition, especially in the distribution sector (Conway et al. 2006). Since these forms of regulation have weakened over time, the story is not that European regulation has become more stringent, but rather that existing regulation became more costly in the context of a new technological era. Of course, European countries have varied considerably in these respects; for example, the UK and Sweden have been better placed than Italy and Spain.

The example of ICT prompts some more general comments on European supply-side policies in the decades before the crisis. In some respects, these provided conditions more favorable to growth. European countries became more open to trade, with positive effects on productivity, partly as a result of the European single market. Years of schooling were steadily increased and product market regulation inhibiting competition was reduced. Corporate tax rates have fallen since the early 1980s. Nevertheless, supply-side policies are in need of further reform if the issue of disappointing growth performance is to be adequately addressed and catch-up resumed. Aghion and Howitt (2006) stress that as countries get closer to the frontier it becomes more important to have high-quality education and strong competition in product markets. These are areas where European countries generally have room for significant improvement.

There have been serious question marks about the quality of schooling in many European countries, which recent research suggests exacts a growth penalty. A measure of cognitive skills, based on test scores, correlates strongly with growth performance (Hanushek and Wössmann, 2012) and it is striking that even the top European countries such as Finland have fallen behind Japan and South Korea, with some countries such as Germany and, especially, Italy deteriorating. These authors estimate that, if cognitive

[31] In a country with no ICT production, a neoclassical growth model whose Cobb-Douglas production function has two types of capital (ICT and other) shows that the steady-state rate of growth will be TFP growth plus a term denoting the rate of real price decline for ICT capital multiplied by the share of ICT capital in national income, all divided by labor's share of national income (Oulton, 2012).

skills in Italy were at the standard of South Korea, its long-run growth would be raised by about 0.75 percentage points per year. Wössmann et al. (2007) show that the variance in outcomes in terms of cognitive skills is explained by the way the schooling system is organized rather than by educational spending.

Competition and competition policy has tended to be weaker than in the United States. This has raised mark-ups and lowered competitive pressure on managers to invest and to innovate with adverse effects on TFP growth (Buccirossi et al. forthcoming; Griffith et al. 2010). Productivity growth in market services has been very disappointing in many European countries (Timmer et al. 2010). One reason is continued weakness of competition reflected in high price-cost mark-ups which have survived the introduction of the Single Market (Høj et al. 2007). Addressing these issues by reducing the barriers to entry maintained by member states would have raised productivity performance significantly but governments still have considerable discretion to maintain these barriers notwithstanding the Services Directive (Badinger and Maydell, 2009).

Western Europe remains a tremendously rich and successful economy, despite the slowdown in its relative growth rate. The major problems facing it at the time of writing have to do with its broken banking system and dysfunctional monetary union, a reminder that growth experiences even over quite lengthy periods of time can be influenced by what are often thought of as short-run monetary factors. Once these issues have been sorted out, one way or another, a longer run issue will remain: how to reshape European economies so as to make them more dynamic without abandoning those elements of the postwar settlement that are most valued by Europe's citizens.

6.7.2 The UK in the Golden Age and After[32]

After being the undisputed economic leader for much of the 19th century, Britain entered a prolonged phase of relative economic decline. This became so pronounced during the Golden Age that by the end of the period Britain had been overtaken by seven other European countries in terms of real GDP per person, and by nine others in terms of labor productivity. Growth was at least 0.7 percentage points per year slower in the UK than in any other country, including those which started the period with similar or higher income levels. The proximate reasons for relatively slow labor productivity growth were weak growth in capital per worker and TFP compared with more successful economies like West Germany. Although slower growth was partly due to convergence forces, being overtaken is a clear indicator of failure.

What is particularly interesting about this episode is the way in which long-standing institutions interacted with changes in the political and economic environment in a way that not only rendered them toxic, but also precluded reform for several decades. The key changes in the economic and political environment were a serious erosion of competition

[32] This section draws in part on Crafts (2012, 2013b).

in product markets, and the need to maintain very low levels of unemployment in order for governments to be re-elected. There were two distinctive institutional legacies that turned out to be costly when the Golden Age opportunity for rapid growth came along. First, corporate governance exhibited an unusual degree of separation of ownership and control in large companies without dominant shareholders (Foreman-Peck and Hannah, 2012). Given that the market for corporate control through takeovers did not work effectively as a constraint (Cosh et al. 2008), weak competition allowed considerable scope for managerial underperformance. Second, the system of industrial relations was characterized by craft control, multi-unionism, and legal immunities for industrial action (Crouch, 1993).

Britain did not achieve the transformation of industrial relations—Eichengreen's cooperative equilibrium—that happened elsewhere in Europe and this implied a considerable growth penalty (Gilmore, 2009).[33] When it is not possible to write binding contracts, either the absence of unions or strong corporatist trade unionism would have been preferable to the idiosyncratic British system. In Britain it was generally not possible to make corporatist deals to underpin investment and innovation, because bargaining took place with multiple unions or with shop stewards representing subsets of a firm's workforce. These unions had considerable bargaining power as a result of full employment and weak competition, but no incentive to internalize the benefits of wage restraint. This exposed sunk cost investments to a hold-up problem, with knock-on implications for investment and growth.[34]

Failure to successfully reform industrial relations was a major shortcoming of British governments from the 1950s through the 1970s. However, throughout this period there were continual efforts to persuade organized labor to accept wage moderation, not only to encourage investment, but even more to allow low levels of unemployment without inflation at a time when politicians believed that this was crucial to electoral success after the interwar trauma. At worst this was tantamount to allowing a de facto trade union veto on economic reforms. In any event, British supply-side policy, which was shaped by the postwar settlement, was unhelpful toward growth in several respects. Problems included a tax system characterized by very high marginal rates, described by Tanzi (1969) as the least conducive to growth of any of the OECD countries in his study; missing out on benefits from trade liberalization by retaining 1930s protectionism into the 1960s (Oulton, 1976); a misdirected technology policy that focused on invention rather than diffusion (Ergas, 1987); and an industrial policy that ineffectively subsidized physical investment (Sumner, 1999) and slowed down structural change by protecting ailing industries through subsidies (Wren, 1996).

[33] Gilmore (2009) finds that coordinated wage bargaining was positive for investment and growth prior to 1975 but not subsequently. This fits with the suggestion in Cameron and Wallace (2002) that the key to the Eichengreen equilibrium is that both sides be patient, and that this was no longer the case when the macroeconomic turbulence of the 1970s erupted.

[34] This can readily be understood in terms of the Eichengreen (1996) model or an extension of it to incorporate endogenous innovation.

A key feature of the Golden Age British economy was the weakness of competition in product markets that had developed in the 1930s and intensified subsequently. Competition policy was largely ineffective while market power was substantial and entrenched politically (Crafts, 2012). The lack of competition had an adverse effect on British productivity performance during the Golden Age working at least partly through industrial relations and managerial failure. Broadberry and Crafts (1996) found that cartelization was strongly negatively related to productivity growth in a cross-section of manufacturing industries for 1954–1963, a result which is confirmed by the difference-in-differences analysis in Symeonidis (2008). In the 1970s and 1980s, greater competition increased innovation (Blundell et al. 1999) and raised productivity growth significantly in companies where there was no dominant external shareholder (Nickell et al. 1997). Both these results underline the role of weak competition in permitting agency cost problems to undermine productivity performance.

Case studies strongly implicate bad management, and restrictive labor practices resulting from bargaining with unions, in poor productivity outcomes. Pratten and Atkinson (1976) reviewed 25 such studies and found evidence of either or both of these problems in 23 of them. Prais (1982) reported similar findings in 8 out of 10 industry case studies and in each case noted that competition was significantly impaired. Multiple unionism, unenforceable contracts, and plant bargaining with shop stewards created an environment in which, unlike West Germany, workers and firms could not commit to "good behavior." This weakened incentives to invest and innovate (Bean and Crafts, 1996; Denny and Nickell, 1992).

The competitive environment that had largely precluded failure in the pre-1914 period had disappeared. This allowed the problems of poor management and dysfunctional industrial relations, often seen as the Achilles' heel of the British economy in the Golden Age, to persist. The politics of economic policy operated to prevent supply-side reforms that could have prevented relative economic decline by enhancing social capability. This period only ended with the election of a maverick prime minister in 1979.

The post-Golden Age period is helpful as a test of this interpretation, since government policy moved in the direction of increasing competition in product markets. In particular, protectionism was discarded. Trade liberalization in its various guises reduced price-cost margins (Hitiris, 1978; Griffith, 2001). The average effective rate of protection fell from 9.3% in 1968 to 4.7% in 1979, and 1.2% in 1986 (Ennew et al. 1990). Industrial policy was downsized as subsidies were cut, and privatization of state-owned businesses was embraced while de-regulation was promoted. In addition, legal reforms of industrial relations reduced trade union bargaining power, which had initially been undermined by rising unemployment. Reforms of fiscal policy were made including the re-structuring of taxation by increasing VAT while reducing income tax rates. The Thatcher government saw itself as ending the trade unions' veto on economic policy reform. Many of the changes of the 1980s would have been regarded as inconceivable by informed opinion in the 1960s and 1970s.

European productivity growth slowed down markedly after the Golden Age, but less so in the UK than in most other countries. Increased competition and openness in the later 20th century was associated with better productivity performance. Proudman and Redding (1998), exploring differing experiences across British industry between 1970 and 1990, found that openness raised the rate of productivity convergence with the technological leader. In a study looking at catch-up across European industries, Nicoletti and Scarpetta (2003) found that TFP growth was inversely related to product market regulation (PMR). The implication of a lower PMR score as compared with France and Germany was a TFP growth advantage for the UK of about 0.5 percentage points per year in the 1990s. At the sectoral level, when concentration ratios fell in the UK in the 1980s, there was a strong positive impact on labor productivity growth (Haskel, 1991). Entry and exit accounted for an increasing proportion of manufacturing productivity growth, rising from 25% in 1980–1985 to 40% in 1995–2000 (Criscuolo et al. 2004).

The impact was felt at least partly through greater pressure on management to perform and through firm-worker bargains that raised effort and improved working practices. Increases in competition resulting from the European Single Market raised both the level and growth rate of TFP in plants which were part of multi-plant firms, and thus most prone to agency problems (Griffith, 2001). Liberalization of capital market rules allowed more effective challenges to incumbent management. A notable feature of the period after 1980 was divestment and restructuring in large firms and, in particular, management buyouts (often financed by private equity) which typically generated large increases in TFP levels in the 1988-1998 period (Harris et al. 2005).

The 1980s and 1990s saw major changes in the conduct and structure of British industrial relations. Trade union membership and bargaining power were seriously eroded. This was prompted partly by high unemployment and anti-union legislation in the 1980s but also owed a good deal to increased competition (Brown et al. 2008). Increased competition may have been the more important factor in boosting British performance, since the 1980s saw a surge in organizational change in those unionized firms exposed to increased competition (Machin and Wadhwani, 1991). De-recognition of unions in the context of increases in foreign competition had a strong effect on productivity growth in the late 1980s (Gregg et al. 1993). The negative impact of multi-unionism on TFP growth, apparent from the 1950s through the 1970s, evaporated after 1979 (Bean and Crafts, 1996). The productivity payoff was boosted by the interaction between reforms to industrial relations and product market competition.

6.8. CASE STUDIES II: SUCCESS, AT LEAST FOR NOW

In this section we briefly examine three more postwar growth miracles, in East Asia, China, and Ireland. While the East Asian Miracle was called into question after the crisis of 1997, growth there soon resumed. It remains to be seen whether Irish growth will go the way of Japan in the 1990s, but income levels there are massively higher than in the 1980s; the same is true of China, despite concerns about the future of economic growth in that country.

6.8.1 The East Asian Miracle

We have seen that the four East Asian Tiger economies enjoyed one of the most impressive growth experiences of the 20th century, although they had to wait until the end of World War II for it to begin. The four countries concerned differed in terms of size, history, and political system. In 1950, South Korea had the largest population, some 20 million, while the Taiwanese population was almost 7.5 million. Hong Kong and Singapore were both city states, with populations of just 2 million and 1 million, respectively. Korea and Taiwan had both been Japanese colonies, and were on the front line of the struggle between communism and the West. Singapore and Hong Kong were both British colonies, but whereas Singapore sought and eventually obtained independence, first as part of Malaysia in 1963, and then as an independent state in 1965, Hong Kong remained British until the handover to China in 1997. GDP per capita in South Korea and Taiwan was around $900 in 1990 international prices, at or below the level of several countries in sub-Saharan Africa; it was slightly over $2000 in Hong Kong and Singapore. Korea, Singapore, and Taiwan all developed more or less authoritarian systems of government, while Hong Kong was ruled as a Crown Colony until 1997. And yet all four countries achieved spectacular growth, with the result that per capita GDP in 2007 was as high in Korea and Taiwan as in Western Europe, and substantially higher in Hong Kong and Singapore. This is a sufficiently impressive performance deserving of the label miraculous, irrespective of the relative contributions of factor accumulation and TFP growth in producing it.

Several features of this growth experience fit neatly into the general convergence framework outlined above. The four countries were relatively poor in 1950, and had high levels of social capability. The famous index of Adelman and Morris (1967) showed that both Korea and Taiwan having extremely high levels of social and economic development in the late 1950s and early 1960s, and Temple and Johnson (1998) have found that this indicator was strongly correlated with subsequent growth. As elsewhere, growth in the Tiger economies was characterized by high rates of investment—in human as well as physical capital—and rapidly growing trade. Investment stood at around 30% of GDP in both Korea and Taiwan in 1980 (Rodrik, 1995, p. 59). Technology transfer was actively encouraged via licensing, technical assistance, inward investment by multinationals, and joint ventures (with the mix differing between countries: for example, Korea discouraged inward direct investment, while Singapore actively encouraged it). Another feature of East Asian growth was that it was based on industrialization, with countries specializing first in textiles, then in heavy industry, and then in electronics and high-tech industries. This pattern of switching over time from labor-intensive, to capital-intensive, and finally to technology-intensive industries, promoted spillover effects on neighboring countries in Southeast Asia, as later industrializers started manufacturing goods which earlier industrializers were no longer producing: the so-called flying geese phenomenon (Ito, 2001).

A substantial literature emerged in the late 1980s and early 1990s which argued that these economies had developed on the basis of institutions and policies which differed greatly from the hands-off prescriptions of the Washington Consensus, involving among other things public enterprise, active industrial policy, export promotion, and protectionism. For Amsden (1989, p. 6), the institutions that mattered in Korea were "an interventionist state, large diversified business groups, an abundant supply of competent salaried managers, and an abundant supply of low-cost, well-educated labor." The key policies were to provide subsidies to firms that, crucially, were conditional on better performance and export market share; and to establish what were effectively multiple prices for capital and foreign exchange, via subsidies or other policies. According to Amsden, these multiple prices were needed in order to accommodate conflicting objectives, for example to encourage savings while keeping the cost of capital to firms low; or to encourage exports while keeping the cost of imports low. For Wade (1990) activist East Asian governments not only influenced the growth rate, but the sectoral composition of output, with beneficial consequences. Some of these conclusions, but not all, were taken on board in a 1993 World Bank study (World Bank, 1993), before the TFP literature reviewed above stimulated a debate about whether the East Asian Miracle was really as impressive as all that: if TFP growth was low, after all, then interventionist arguments based on learning by doing or other growth-promoting externalities seemed less convincing (Krugman, 1994, p.78).

While Amsden and Wade emphasized export performance, subsequent work has downplayed the role of exports. Rodrik (1995) argues that exports cannot have been a prime mover of industrialization, since if they had been, this would have been manifested in a rising relative price of exports as world demand rose. Since there was no such price rise, the ultimate sources of growth must have been internal, with exports arising as a consequence. For Rodrik the key to growth was investment, consistent with the growth-accounting evidence. State intervention was required, not just to boost savings and investment rates via subsidies, public investment, and other measures, but to coordinate investment across a range of complementary sectors for "big push" reasons. Rodrik shows a striking positive correlation over time in both Korea and Taiwan between investment and imports, which can be explained by the fact that investment required large imports of machinery and other capital equipment. Exports were needed to pay for these imports, and were thus necessary for the investment drive, but they were not the ultimate cause of rapid growth. This does not mean that exports were not essential: they were, both to finance required imports of capital goods, and to sell the output that the high investment rates were designed to produce. An implication is that while countries like Korea and Taiwan were not liberal free traders themselves, they did benefit from the generally open international trade policies of the major Western economies during this period.

Some of the same features which had been seen as aiding rapid East Asian convergence on the frontier—in particular, close relationships between the state and big business, and heavy reliance on bank finance—were blamed for the East Asian financial crisis which erupted in 1997. Given the debate that had recently taken place about East Asian TFP, it is not surprising that the crisis emboldened some to argue that these institutional features of East Asian growth had been a hindrance rather than a help, all along. It is difficult to see how one could establish such counterfactual claims convincingly. For example, studies failing to find correlations across sectors between subsidies or other policy interventions, on the one hand, and productivity or other outcomes on the other, would presumably not convince an advocate of big push policies, which are based on intersectoral complementarities. The argument that the financial crisis proves anything about the sources of rapid East Asian growth from the 1960s to the 1990s seems somewhat dated today, in the light of the global financial crisis of 2007–2008. This hit some of the richest countries in the world, with very different institutional structures than those found in, say, Korea. The experience of the Eurozone periphery shows the dangers of being exposed to capital inflows in the presence of a currency peg; unlike Korea or Thailand these countries do not have the option of devaluation, and by 2013, have not yet recovered. In sharp contrast, East Asian growth resumed rapidly in 1999. We thus agree with Ito (2001), who argues that the debate about which institutions and policies mattered for East Asian growth during the growth miracle needs to be sharply distinguished from the debate about how to regulate banks and international capital flows. This does not mean, however, that at some stage East Asia may not have to rethink its institutional mix as it moves even closer to the international technological frontier.

6.8.2 China

While Chinese economic statistics are unreliable, it is clear that China's growth since the start of economic reforms in the late 1970s has been extraordinary.[35] Even if Maddison and Wu's (2008) data are accepted, GDP growth averaged 7.85% per annum between 1978 and 2003, as compared with an official growth rate almost 2 percentage points higher. As in the Western European and East Asian cases, very high levels of investment, the importation of technology, and increasing links with the outside world played key roles in China's growth acceleration. As in East Asia, Chinese industry went through successive phases, from exporting labor-intensive toys, clothes, and footwear to producing more capital-intensive, and ultimately high-tech goods. The Chinese savings rate averaged an extraordinary 37% between 1978 and 1995 according to Kraay (2000), although Heston and Sicular (2008) favor a lower (but still large) figure somewhere in the 20–30% range.

[35] Our account draws heavily on the collection of essays in Brandt and Rawski (2008a), the standard reference on the subject.

This has allowed an equally high investment rate to be internally financed (Lee et al. 2012).

What makes the Chinese experience unique is the way in which a gradualist reform program has seen the role of the state in economic life being steadily diminished over time, at the same time as that state has maintained a highly authoritarian political system. China is no poster child for the Washington Consensus. It still maintains exchange controls, state-owned enterprises remain an important drag on the economy (Brandt et al. 2008), and the government intervenes in the economy in myriad other ways. But the facts that the direction of change since 1979 has so clearly been in a market-friendly direction, and that China's economic situation has improved so much since then, mean that the literature on China's economic miracle has tended to focus on how gradual liberalization improved performance, rather than (as in the East Asian Tiger case) on whether some government interventions helped speed China's convergence on the technological frontier.

China's reforms came in two stages (Brandt and Rawski, 2008b), which according to Naughton (2008) corresponded to different configurations of political power. In the first stage, which lasted from 1979 to 1992, political power was fragmented, and reforms were incremental, and concerned with not creating losers. Agricultural households were permitted to engage in cultivation. A growing number of firms, notably township and village enterprises (TVEs), were allowed to enter an increasing range of sectors. Once firms had satisfied their plan targets, they were able to sell additional output at what evolved into market prices. Four special economic zones were set up in the southern coastal provinces, in which Hong Kong and Taiwanese firms produced labor-intensive goods for export. Fourteen additional zones were created in 1984, and regulations regarding foreign direct investment were further relaxed in 1986 (Branstetter and Lardy, 2008, pp. 640-1).

In the second stage, from 1993 onwards, power was consolidated as the first generation of Communist leaders left the political stage. Reforms became more decisive, and capable of producing losers as well as winners. The plan component of the dual pricing system was abandoned; restrictions on mobility between town and countryside were relaxed, setting the stage for a mass migration of rural workers to industrial cities; TVEs were privatized; state-owned enterprises (SOEs) were subjected to more market discipline, and were downsized and occasionally closed. The number of workers in SOEs fell from 76 million in 1992–1993 to 28 million in 2004 (Naughton, 2008, p. 121): Brandt et al. (2008) estimate that the resulting reallocation of workers toward more productive firms elsewhere in the economy made an even more important contribution to GDP growth than rural to urban migration. For Hsieh and Klenow (2009), the reallocation added 2 percentage points to China's TFP growth rate between 1998 and 2005, while for Song et al. (2011), it not only helps explain China's growth since the early 1990s, but its growing external surpluses as well (since the growing private sector relies more on internal financing for its investment needs than the SOEs). The growing liberalization of Chinese trade policy culminated with China's accession to the WTO in 2001. Even prior to that, foreign companies had been enabled to operate in China subject to many fewer restrictions and less interference, leaving China well positioned to benefit from the boom of the last few years of the Great Moderation (Branstetter and Lardy, 2008, p. 645). Despite the clear acceleration in the pace of reforms, Chinese reforms remained gradual compared with the experience of Eastern Europe and the former Soviet Union (Svejnar, 2008).

Given that corruption is a severe problem in China, and that other aspects of its institutional structure remain deeply problematic, it is perhaps not surprising that much analysis has focused on the dismantling of Communist economic controls as being at the heart of China's economic success: merely getting rid of obstacles can lead to significant growth if these were costly enough (Brandt and Rawski, 2008b, p. 9). And yet government intervention may have helped growth as well as hindered it. The size of the Chinese market has allowed national and regional officials to extract concessions from foreign multinationals, notably with respect to the transfer of technology and research and development activities, that could in principle have accelerated China's catching up relative to that of other poor economies (Brandt et al. 2008, p. 623). The fact that China's real exchange rate depreciated by 70% vis à vis the dollar between 1980 and 1995 presumably increased the attraction of China as a manufacturing location. This in turn helped make China's exchange rate policy more politically sustainable, by creating overseas political constituencies favorable to it (Branstetter and Lardy, 2008, pp. 639, 675–676). And underlying everything else has been competition between regional officials, whose promotion prospects depend on their region's economic performance. This "regionally decentralized authoritarianism" (Xu, 2011) has been a major factor in China's economic success.

There were clear signs in 2013, that the Chinese financial sector might be heading for a major crisis with unpredictable consequences for the Chinese economy and political system. Even aside from this risk, it may be that the policies and institutional structures which have underpinned China's economic growth since 1978 are beginning to outlive their usefulness and will have to be changed. Relative prices skewed in favor of exports may be distorting the Chinese economy (Branstetter and Lardy, 2008, p. 676), and are in any rate leaving the economy vulnerable to overseas shocks. Many commentators argue that an investment rate approaching 50% of GDP may no longer be sustainable, and that an increasing focus on consumption is now required. For some China is now approaching its "Lewis (1954) moment," as rural to urban migration slows and wages start to rise. Elastic supplies of labor from agriculture (and, if Song et al. 2011 are right, from SOEs as well) made high Chinese investment rates consistent with high returns on capital; as both pools of labor shrink, diminishing returns to capital will set in (Das and N'Diaye, 2013; Krugman, 2013).

If extensive growth at current rates becomes more difficult, and the need for intensive growth thus increases, deeper institutional changes may be needed. Perhaps, as Naughton (2008, p. 127) speculates, as China moves closer to the technological frontier its economy will need "transparency and recourse to impartial independent regulatory authority that the current system is not yet able to provide." It is not yet clear that China will be able to escape the middle income trap. Eichengreen et al. (2012) find that the probability of a growth slowdown (defined as a decline in growth rates of 2 percentage points or more) increases not only at income levels which China can be expected to attain in the next few years, but in countries which have maintained undervalued exchange rates, have low consumption shares of GDP, and aging populations. On all fronts China appears vulnerable, implying that the probability of a growth slowdown is high there, and that is even before taking the country's financial problems into account. Whether such a slowdown will occur, and how the country's economy, society, and political system would respond, are among the major uncertainties facing the world economy in the early 21st century.

6.8.3 Ireland: The Celtic Tiger

The spectacular growth of the Celtic Tiger period when a small economy rode the globalization wave with massive success attracted enormous attention. Its proximate sources in export platform FDI and ICT production are apparent. Less well understood is the fact that up through the mid-1980s Ireland had been a failure (Ó Gráda and O'Rourke, 1996). We saw earlier that affluent Western economies experienced unconditional convergence after 1950, with poorer countries growing more rapidly than richer ones. Seen in this perspective, Ireland was the great underperformer prior to 1987, as Figure 6.3 shows, with growth rates well below those that would have been expected given its relative poverty in 1950. Ireland's average growth rate between 1950 and 1987, 2.8% per annum, was approximately the same as that in the Benelux countries, despite the fact that its 1950 per capita income lay between those of Austria and Italy. In the context of the Golden Age, this was a spectacular economic failure.

The reasons for this failure are related to the reasons for success in the rest of Western Europe at the same time. The 1950s were particularly unimpressive in Ireland, with per capita growth rates of only 1.7%. Education remained underfunded and underprovided. Instead of corporatist labor market institutions as in continental Europe, Ireland had a fragmented British-style trade union system incapable of delivering wage moderation in return for high investment. Even if such wage moderation had been delivered, Irish firms were small, unproductive, and focused on the home market, while foreign firms were discouraged from investing in the country. This was the legacy of 1930s protectionism, which might have been the correct response to the Great Depression, but should have been abandoned much earlier than it actually was. Such investment as there was too often went to relatively unproductive purposes, with Irish savings being invested in low-yielding projects for political reasons. Not surprisingly, Irish TFP was very low by European standards in 1960 (Crafts, 2009).

Gradually these impediments to growth were done away with. The late 1950s and early 1960s saw the introduction of export tax relief, and measures to attract foreign direct investment, which was to become the key to Ireland's convergence on the technological frontier. Trade was gradually liberalized: Ireland entered an Anglo-Irish free trade area in 1965, and the EEC in 1973. The late 1960s saw belated educational reforms that made secondary schooling available to everyone. Growth was twice as high in the 1960s as in the 1950s, but was slightly less than the Western European average: Ireland was still not converging, and in 1973 was poorer than Greece, Portugal, and Spain. EEC membership helped to modernize the economy in many ways, but the oil crisis that coincided with entry ushered in a period of low growth, large government budget deficits, and a subsequent fiscal crisis, which led to a second post-war lost decade during the 1980s.

After 1987, Ireland's economic performance was transformed out of all recognition. Between 1987 and 2000 its per capita growth rate averaged 5.7% per annum, with the result that by 2000 Ireland lay on the advanced economy "convergence line" (Figure 6.4). So how was Ireland turned around? Figures 6.3 and 6.4 suggest a straightforward explanation: that the Irish miracle was simply a delayed version of the Western European growth miracle of the 1950s and 1960s (Ó Gráda and O'Rourke, 2000; Honohan and Walsh, 2002). What changed was that many of the structural impediments to convergence had been eliminated over the course of the 1960s and 1970s, leaving Ireland well positioned to take advantage of deeper European integration and a buoyant international economy in the 1990s. The catastrophe of the 1980s meant that trade unions were willing to

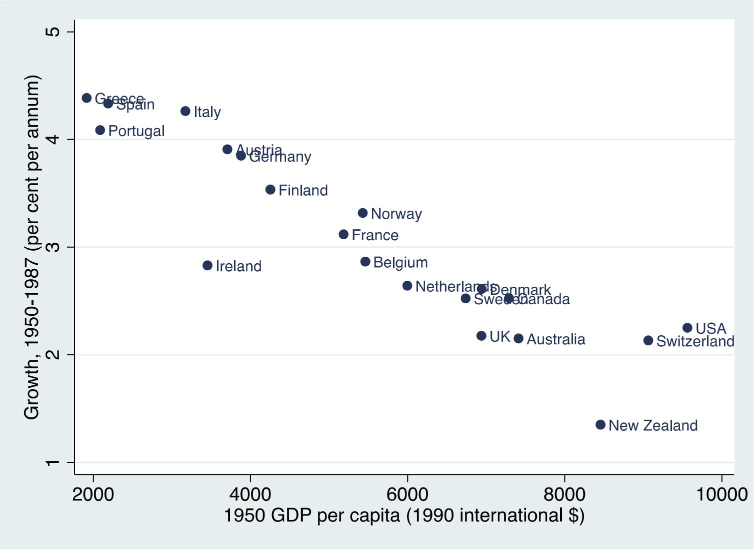


Figure 6.3 GDP per capita growth, 1950–1987. *Source:* Bolt and Van Zanden (2013).

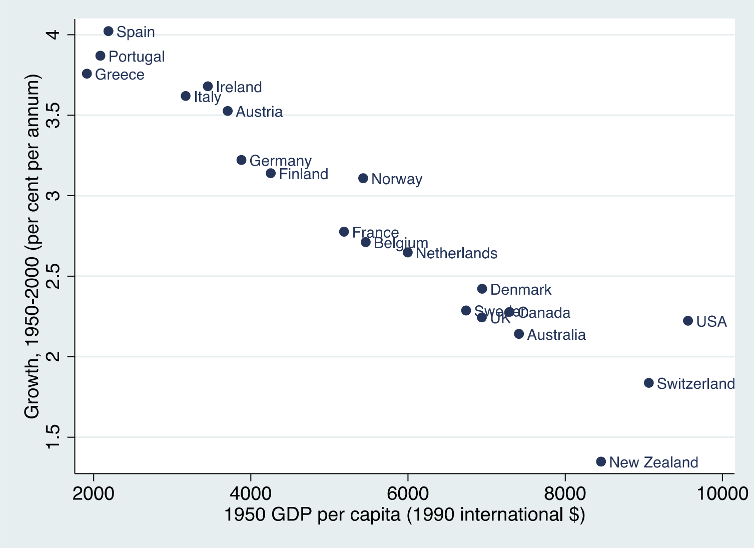


Figure 6.4 GDP per capita growth, 1950–2000. *Source:* *Bolt and Van Zanden (2013)*.

enter into corporatist social partnership agreements, trading wage restraint against the promise of growth and employment. Irish workers were now far better educated than they had been during the 1960s. Devaluations in 1987 and 1993 helped to boost Irish competitiveness. A healthier labor market now interacted with Ireland's long-standing low corporate taxes and produced a surge of inward investment, rising TFP levels, and increases in employment.

Imports of technology, corporatist labor bargains leading to investment, and a reliance on exports are all reminiscent of the Western European miracle of three or four decades previously. The differences were also noteworthy, and reflected the period. Much of the investment occurred via FDI, rather than being financed via retained profits by domestic firms. Workers' wage restraint was compensated more with tax cuts than with an expansion of the welfare state. And Ireland did not go through all the stages of industrialization to the same extent as other countries, specializing far more in ICT and other high-tech sectors than the typical fast grower of the 1950s or 1960s. This specialization did not just reflect the invisible hand of the market, but active Irish government attempts to develop clusters of activity in ICT, pharmaceuticals, and similar sectors (Barry, 2002).

The Celtic Tiger period ended in 2000 or 2001, and was replaced by the Celtic Bubble of 2001–2007, financed by cross-border capital flows which boomed in the aftermath of Ireland's entry into the Euro. Not only the bubble, but also the crash which followed, was reminiscent of the East Asian crisis of 1997, but with the important difference that Ireland was not able to respond to the crisis by adjusting its exchange rate. Foreign observers have not subjected the Irish model to the same sort of scrutiny that the East Asian model faced after 1997, which is perhaps ironic since the Irish economy had still not started to recover by 2013, in stark contrast with the rapid and durable post-crisis recovery in East Asia. The net result is, that at the time of writing, it seemed that Ireland risked facing a third post-war "lost decade," after those of the 1950s and 1980s. It is too soon to say whether and when growth will resume, in Ireland or in the rest of the Eurozone periphery. Nonetheless, the Celtic Tiger was no mirage: Ireland is now one of the richest countries in Western Europe, not one of the poorest.

6.9. CASE STUDIES III: FAILURES

In this section we look briefly at two cases which can fairly be considered failures, the USSR and Africa.

6.9.1 Failed Catch-Up in USSR[36]

The USSR was always a long way below the United States in terms of real GDP per person—about 30% in 1950 and 36% in 1973—and, despite a promising start, only reduced the gap very slowly. A growth rate of 3.37% per year in the Golden Age compares quite unfavorably with the achievements of Western European countries like Italy or Spain which also started out with relatively low income levels. Growth regression evidence confirms that communist countries underperformed in the Golden Age: allowing for initial income levels, their growth rate was about 1.3 percentage points lower than that of their Western European counterparts (Crafts and Toniolo, 2008).

Worrying signs of a serious slowdown in productivity growth did not appear until the 1970s. Golden Age Soviet growth was extensive, in that the investment/GDP ratio roughly doubled between 1950 and the early 1970s to about 30%. The capital stock grew at about 8.5% per year in this period (Ofer, 1987). However, diminishing returns to capital accumulation exacerbated by slow TFP growth implied that the rate of capital stock growth delivered by a given investment rate was falling over time: the capital-stock growth rate fell from 7.4% per year in the 1960s to 3.4% per year in the 1980s. Negative TFP growth post-1970 (Table 6.16) was driven by "waste of capital on a grand scale" (Allen, 2003, p. 191) as old factories were re-equipped and expansion of natural resource industries in Siberia were pursued.

Relatively low TFP growth was not the result of inadequate volumes of R & D, which by the 1970s were very high by world standards at around 3% of GDP. Rather, the problem lay in the incentive structures that informed innovation at the firm level. This was a classic case of a social capability failure. The planning system rewarded managers who achieved production targets in the short term rather than those who found ways to reduce costs or improve the quality of output over the long term. The balance of risk and

[36] This section draws in part on Crafts and Toniolo (2008).

reward was inimical to organizational and technological change, and the "kicking foot" of competition was absent (Berliner, 1976).

The incentive structures used by the Soviet leadership to motivate managers and workers were a complex mixture of rewards, punishments, and monitoring. Each of these became increasingly expensive over time, implying that the viability of the system was threatened. Product innovation drove up monitoring costs, and this inhibited moves from mass to flexible production. A more educated population meant that punishment (incarceration) was more costly in terms of loss of human capital, and that rewards needed to be higher. The slowdown in productivity growth led to a search for reforms that might improve economic performance and lower monitoring costs, but these ended up undermining the regime's reputation for brutality, which could help sustain high effort in circumstances when punishment costs became particularly high. The interesting feature of this system is that it could be tipped from a high coercion, high effort equilibrium to a low coercion, shirk and steal equilibrium if rewards and punishments were no longer credible and workers understood this. Harrison (2002) argues that such a shift accounts for the sudden collapse at the end of the 1980s.

6.9.2 Post-Colonial Sub-Saharan Africa

As we have seen, average growth performance in this region was dismal between the mid-1970s and the late 1990s. There was stagnation in real GDP per person (Table 6.8), TFP growth was actually negative (Table 6.11), and it became commonplace to talk about a chronic growth failure (Collier and Gunning, 1999). However, the first decade of the 21st century saw a revival in growth performance. Taking a long view of African growth, it may be more accurate to see a picture of growth accelerations followed by growth reversals, with the former typically triggered by strong commodity prices, as in the recent growth spurt (Jerven, 2010). Unfortunately, econometric analysis shows that while commodity price booms have raised income levels in the short run, their long-run effect is to lower them somewhat (Collier and Goderis, 2012).

Very low institutional quality is the most obvious explanation for disappointing growth and low income levels at the end of the 20th century. On average, sub-Saharan African countries score badly on the World Bank's *Governance Matters* and *Doing Business* indicators and do so persistently. Thus, ever since 1996, when it was first compiled, the average score on the "rule-of-law" indicator has been about −0.7 (on a scale of −2.5 to +2.5) compared with an average for Western Europe of around +1.6. Similarly, the norm across the region is a closed access society (Kishtainy, 2011). If the fundamental reason for poverty is insecure property rights (Acemoglu and Johnson, 2005) then sub-Saharan Africa is a prime exhibit. Indeed, this is now often taken as a stylized fact, with "absolutist weak states" having "little ability or interest in providing public goods" and operating on a "neopatrimonial" basis (Acemoglu and Robinson, 2010, pp. 23, 40). Of course, there are exceptions to this dismal picture, such as Botswana and Mauritius, but they are the exceptions that

Table 6.23 Growth of real GDP/person, 1960–2000 (percent per annum)

	Resource-scarce & coastal (%)	Resource-scarce & landlocked (%)	Resource-rich (%)
Africa	0.50 (33)	−0.36 (33)	0.29 (33)
Other developing	3.79 (88)	1.40 (1)	2.89 (11)

Note: numbers in parentheses refer to percentages of population by region in each category.
Source: Collier (2007).

prove the rule, and score relatively well when it comes to governance indicators. This is an account that is entirely consistent with the New Institutional Economic History tradition.

However, the intriguing question remaining is what part geography may have played in explaining African failure. On a range of indicators, including climate, coastal access, disease environment, and population density, Africa scores much less well than other regions of the developing world (Sachs et al. 2004). It seems reasonable to suppose that this carries a growth penalty in terms of adverse impacts on investment and productivity. "Naive" growth regressions suggest that this is the case and accord geographic factors nearly as much weight as institutions in accounting for the differential between African and East Asian growth performance in the late 20th century (Bleaney and Nishiyama, 2002). If the focus is switched to "second-nature" geography, then sub-Saharan Africa scores very badly compared with almost all other parts of the world in terms of market potential, which is strongly correlated with income even after controlling for institutional quality (Redding and Venables, 2004).

Table 6.23 offers a simple but powerful summary of growth performance classified by geographic type. The table shows (in parentheses) the percentages of the population in both Africa and other developing regions in each of three categories: resource-scarce and coastal; resource-scarce and landlocked; and resource-rich. It also shows the average growth rates between 1960 and 2000 in each of these six regions. As can be seen, being both landlocked and resource-scarce is a particularly bad combination for growth, and this is unfortunate for Africa since it has a relatively high percentage of its population in this category. It also has a relatively low proportion of its population in resource-scarce and coastal regions, which saw higher growth rates both in Africa and elsewhere. Geography does not favor Africa, therefore, but this is not the whole story since the table also shows that in each geographic category Africa has seriously underperformed relative to the rest of the developing world.

A more satisfactory way to explain post-colonial African growth failure may be to consider interactions between institutions and geography. One aspect of such interactions is the possibility, noted earlier, that first-nature geography may have its strongest effects through its impact on institutions (Easterly and Levine, 2003). But it is also important to recognize that, "on top of its physical geography and remoteness, Africa has been held back by the fragmentation of its political and economic geography" (Venables, 2010, p. 481)—the median country has a population of only 8 million people. This fragmentation implies a number of serious disadvantages with regard to small city size, weak competition in product markets, reduced supply of public goods, greater difficulty in escaping from bad policies, etc. (Venables, 2010). Payoffs to better policies are often highly dependent on the reform efforts of neighbors, which further hinders economic progress.

A final perspective on sub-Saharan failure is both more historical and, perhaps, more optimistic. Bates et al. (2007) point out that economic performance was also very disappointing in Latin America in the first 50 years following independence in the 1820s, and that it was only in the late 19th century that sustained economic growth began. Furthermore, this growth was as high as that experienced in the British offshoots, as we saw in Section 6.3. Bates et al. explain the initial poor performance as being due to the political instability of the time: international and civil wars, foreign military incursions, and a general atmosphere of violence. This is suggestive, since wars and violence have been prevalent in post-independence Africa as well, and it is often suggested that this is one reason for the continent's poor growth performance. Perhaps the transition to post-colonial independence for new states, with arbitrarily drawn borders, is inherently difficult. If so, Africa may yet see a brighter 21st century, as it gradually leaves these transition problems behind.

6.9.3 The Natural Resource Curse

One major reason for long-term growth failure which has received a great deal of attention in the literature is the so-called natural resource curse. This refers to the tendency for countries with large natural resource exports or minerals production relative to GDP to grow relatively slowly at best, and experience prolonged periods of negative growth at worst. A number of hypotheses have been put forward to explain this correlation and there is a substantial body of empirical work that examines issues of robustness and causality.[37] The standard suggested mechanisms explaining the natural resource curse include crowding out of tradable goods sectors with greater productivity growth potential (Dutch Disease); promoting low quality institutions which undermine growth; making civil war more likely; and engendering macroeconomic volatility. There is some empirical support for all these arguments (Van der Ploeg, 2011). It is also clear that there is a wide range of historical experience that needs to be explained. Some countries have indeed been cursed by natural resources, for example, Angola, Congo, Sierra Leone, and Sudan. However, others have been blessed including, for example, Australia, Canada, Chile, and the United States.

[37] For an excellent recent survey article, see Van der Ploeg (2011).

It seems highly plausible that the implications of a resource windfall will differ depending on whether there are good or bad institutions. In the former case, it might be expected that the bonanza leads to an increase in productive activities, while in the latter case, even more resources will be devoted to rent-seeking. The evidence of growth regressions is consistent with this prediction. Mehlum et al. (2006) use a variable interacting institutional quality, as measured by the ICRG index popularized by Knack and Keefer (1995), and resource abundance. They find that values above 0.60 for the ICRG index make mineral resources good for growth. This accords with common sense: oil has been very good for Norway, but bad for Nigeria.

An economic history perspective allows some of these ideas to be taken further. First, the most notable success story in recent African economic history is Botswana, a resource-abundant country in which diamonds are a large share of GDP. Botswanan success is based not only on diamonds, but also on high institutional quality and secure property rights plus good policies. The underpinnings of good institutions were a combination of historical accident and the economic interests of the pre-diamond era elite, the cattle ranchers (Acemoglu et al. 2003). There was thus a bulwark against the pursuit of mineral rents which led to rent-seeking and states which were ineffective modernizers elsewhere in Africa, for example Angola and Nigeria (Isham et al. 2005).

Second, going beyond the argument that good institutions make natural resources more of a blessing than a curse, it should be noted that natural resource endowments actually reflect the amount of effort devoted to their discovery and effective exploitation. This depends inter alia on the quality of institutions and policies. A classic example is the 19th century United States whose status as a leading minerals producer was the product of big investments in exploration and human capital underpinned by a favorable property rights regime (David and Wright, 1997).

Third, the implications of mineral resources seem to have varied over time for reasons which still need to be fully researched but link to ideas familiar from new economic geography. In the 19th and early 20th centuries, industrialization was encouraged by the proximity of coal, whereas in the later 20th century it seems to have been discouraged by the proximity of oil. Regression evidence for the natural resource curse relates to samples drawn only from the recent past. The difference between now and then is likely to relate to much higher transport costs for minerals, especially over land, in the past; and changes in energy sources with electrification (Wright and Czelusta, 2007).

6.10. CONCLUSIONS

The convergence of a succession of countries onto the technological frontier is a process whose roots lie in the great divergence of the 19th century. That divergence was due to new industrial technologies being implemented in some regions of the world but not in others, and was magnified in the short run by the globalization of the period which, given technological asymmetries, created a stark division of labor between an industrializing West and a deindustrializing Rest.

The key to reducing the resulting regional inequalities has been the erosion of these technological asymmetries, via the spread of modern industrialization. The succession of growth miracles briefly surveyed above seems reminiscent of the process of sequential convergence on the frontier modeled by Lucas (2000, 2009). Since industrial technologies are transferable across borders, convergence should not surprise us. But neither should we assume that convergence will be as smooth as simple growth models assume: the economic history of 20th century growth is also a story of the various frictions that can impede this process. In addition to successes, there have been a variety of failures.

As we have seen repeatedly throughout this chapter, innovation tends to reflect the economic circumstances of the leading economy of the time. This was Britain until some time in the late 19th century, and the United States thereafter. Even in the best of all institutional worlds, with no political or other frictions and Scandinavian levels of social capability, directed technological change would be a factor preventing or at least slowing down the process of technological convergence. Nor is this just a story of developing countries finding it uneconomical to adopt best-practice technology, since European economies, and even Britain itself, found themselves at a disadvantage when it came to adopting American techniques that had been developed with American factor prices, and the American market, in mind.

What is more, we do not live in the best of all institutional worlds, frictions of all sorts are prevalent and we are not all Scandinavian. Social capability matters for growth and not all countries have it. Institutions are path dependent, and can be an impediment to growth. And even in countries where they have always been an asset, they can become a liability, since the right institutional set-up may change over time as countries converge on the technological frontier, or as the nature of frontier technologies change. Chasing a moving target can be a tricky business in a world where history matters.

Geography is another reason why convergence is not as smooth in practice as it can seem in theory. First-nature geography matters, although it may matter in different ways at different points in time: resource abundance may be a blessing in some time periods, but a curse in others, depending on the tradability of resources, on their nature, and on the extent to which frontier technologies are resource-using. It may also be a blessing or a curse depending on a country's institutional set-up, which may in turn reflect that country's geography as well as its history. Being far from trade routes, on the other hand, has never been good for growth in the past, and it is hard to see why it should become so in the future.

Finally, economic historians emphasize the importance of wars, ideological revolutions, financial crises, and other events that are typically regarded as exogenous shocks in economic models, but which are part and parcel of the world in which we live. The First World War, the Russian Revolution, or the Great Depression were not mere complications in the story of 20th century economic growth, but a part of its very fabric. Even episodes which are conventionally regarded as shortrun in nature, having to do with macroeconomic or financial policy, can, if handled sufficiently badly, have a significant impact on economic growth over the course of a lifetime, which is what most of us tend to care about. History, and economic history, have not yet ended.

REFERENCES

Abramovitz, M., 1986. Catching up, forging ahead, and falling behind. Journal of Economic History 46, 385–406.

Abramovitz, M., 1989. Thinking About Growth and Other Essays on Economic Growth & Welfare. Cambridge University Press, Cambridge.

Abramovitz, M., 1993. The search for the sources of growth: areas of ignorance, old and new. Journal of Economic History 53, 217–243.

Abramovitz, M., David, P.A., 1996. Convergence and delayed catch-up: productivity leadership and the waning of American exceptionalism. In: Landau, R., Taylor, T., Wright, G. (Eds.), The Mosaic of Economic Growth. Stanford University Press, Stanford, pp. 21–62.

Abramovitz, M., David, P.A., 2001. Two centuries of American macroeconomic growth: from exploitation of resource abundance to knowledge-driven development. Stanford Institute for Economic Policy Research Discussion Paper No. 01–05.

Acemoglu, D., 2002. Directed technical change. Review of Economic Studies 69, 781–809.

Acemoglu, D., 2010. When does labor scarcity encourage innovation? Journal of Political Economy 118, 1037–1078.

Acemoglu, D., Johnson, S., 2005. Unbundling institutions. Journal of Political Economy 113, 949–995.

Acemoglu, D., Johnson, S., Robinson, J.A., 2001. The colonial origins of comparative development: an empirical investigation. American Economic Review 91, 1369–1401.

Acemoglu, D., Johnson, S., Robinson, J., 2003. An African success story: Botswana. In: Rodrik, D. (Ed.), In Search of Prosperity: Analytic Narratives on Economic Growth. Princeton University Press, Princeton, pp. 80–119.

Acemoglu, D., Robinson, J.A., 2010. Why is Africa poor? Economic History of Developing Regions 25, 21–50.

Acemoglu, D., Zilibotti, F., 2001. Productivity differences. Quarterly Journal of Economics 116, 563–606.

Adelman, I., Morris, C.T., 1967. Society, Politics, & Economic Development: A Quantitative Approach. Johns Hopkins University Press, Baltimore.

Aghion, P., Howitt, P., 2006. Appropriate growth policy: a unifying framework. Journal of the European Economic Association 4, 269–314.

Aiyar, S., Dalgaard, C.-J., 2005. Total factor productivity revisited: a dual approach to development accounting. IMF Staff Papers 52, 82–102.

Akkermans, D., Castaldi, C., Los, B., 2009. Do "liberal market economies" really innovate more radically than "coordinated market economies"?: Hall and Soskice reconsidered. Research Policy 38, 181–191.

Albers, R., Groote, P., 1996. The empirics of growth. De Economist 144, 429–444.

Albouy, D.Y., 2012. The colonial origins of comparative development: an empirical investigation: comment. American Economic Review 102, 3059–3076.

Alesina, A.F., Glaeser, E.L., Sacerdote, B., 2006. Work and leisure in the US and Europe: why so different? In: Gertler, M., Rogoff, K. (Eds.), NBER Macroeconomics Annual 2005. MIT Press, Cambridge, Massachusetts, pp. 1–64.

Allen, R.C., 2003. Farm to Factory: A Reinterpretation of the Soviet Industrial Revolution. Princeton University Press, Princeton.

Allen, R.C., 2009. The British Industrial Revolution in Global Perspective. Cambridge University Press, Cambridge.

Allen, R.C., 2011. The spinning jenny: a fresh look. Journal of Economic History 71, 461–464

Allen, R.C., 2012. Technology and the great divergence: global economic development since 1820. Explorations in Economic History 49, 1–16.

Allen, R.C., 2013. The high wage economy and the industrial revolution: a restatement. University of Oxford Discussion Paper in Economic and Social History No. 115.

Allen, R.C., Weisdorf, J.L., 2011. Was there an “industrious revolution” before the industrial revolution? Economic History Review 64, 715–729.

Amsden, A.H., 1989. Asia’s Next Giant: South Korea and Late Industrialization. Oxford University Press, Oxford.

Badinger, H., 2005. Growth effects of economic integration: evidence from the EU member states. Review of World Economics 141, 50–78.

Badinger, H., Maydell, N., 2009. Legal and economic issues in completing the EU internal market for services: an interdisciplinary perspective. Journal of Common Market Studies 47, 693–717.

Bairoch, P., 1982. International industrialization levels from 1750 to 1980. Journal of European Economic History 11, 269–331.

Baldwin, R., forthcoming. Trade and industrialisation after globalisation’s second unbundling: how building and joining a supply chain are different and why it matters. In: Feenstra, R.C. Taylor, A.M. (Eds.), Globalization in an Age of Crisis: Multilateral Economic Cooperation in the Twenty-First Century. University of Chicago Press, Chicago.

Barro, R.J., 1991. Economic growth in a cross section of countries. Quarterly Journal of Economics 106, 407–443.

Barro, R.J., 1999. Notes on growth accounting. Journal of Economic Growth 4, 119–137.

Barro, R.J., Lee, J.-W., 2012. A new data set of educational attainment in the world, 1950–2010. Journal of Development Economics 104, 184–198.

Barry, F., 2002. The Celtic Tiger era: delayed convergence or regional boom? ESRI Quarterly Economic Commentary, Summer, 84–91.

Basu, S., Weil, D.N., 1998. Appropriate technology and growth. Quarterly Journal of Economics 113, 1025–1054.

Bates, R.H., Coatsworth, J.H., Williamson, J.G., 2007. Lost decades: postindependence performance in Latin America and Africa. Journal of Economic History 67, 917–943.

Bean, C., Crafts, N., 1996. British economic growth since 1945: relative economic decline... and renaissance? In: Crafts, N., Toniolo, G. (Eds.), Economic Growth in Europe Since 1945. Cambridge University Press, Cambridge, pp. 131–172.

Bénétrix, A.S., O’Rourke, K.H., Williamson, J.G., 2013. The spread of manufacturing to the poor periphery 1870–2007. NBER Working Paper No. 18221.

Berliner, J.S., 1976. The Innovation Decision in Soviet Industry. MIT Press, Cambridge, Massachusetts.

Blanchard, O., 2004. The economic future of Europe. Journal of Economic Perspectives 18, 3–26.

Bleaney, M., Nishiyama, A., 2002. Explaining growth: a contest between models. Journal of Economic Growth 7, 43–56.

Bloom, N., Sadun, R., van Reenen, J., 2012. Americans do IT better: US multinationals and the productivity miracle. American Economic Review 102, 167–201.

Blundell, R., Griffith, R., van Reenen, J., 1999. Market share, market value and innovation in a panel of British manufacturing firms. Review of Economic Studies 66, 529–554.

Bogart, D., Drelichman, M., Gelderbloom, O., Rosenthal, J.-L., 2010. State and private institutions. In: Broadberry, S., O’Rourke, K.H. (Eds.), The Modern Economic History of Europe: 1700–1870. Vol. 1, Cambridge University Press, Cambridge, pp. 70–95.

Bolt, J., van Zanden, J.L., 2013. The first update of the Maddison project: re-estimating growth before 1820. Maddison Project Working Paper 4. Data available at <http://www.ggdc.net/maddison/maddison-project/data/mpd_2013-01.xlsx>.

Boskin, M.J., Dulberger, E.R., Gordon, R.J., Griliches, Z., Jorgenson, D.W., 1996. Towards a More Accurate Measure of the Cost of Living. US Government Printing Office, Washington, DC.

Bosworth, B.P., Collins, S.M., 2003. The empirics of growth: an update. Brookings Papers on Economic Activity 2, 113–206.

Bosworth, B.P., Collins, S.M., 2008. Accounting for growth: comparing China and India. Journal of Economic Perspectives 22, pp. 45–66.

Bourguignon, F., Morrisson, C., 2002. Inequality among world citizens: 1820–1992. American Economic Review 92, 727–744.

Brandt, L., Rawski, T.G. (Eds.), 2008a. China's Great Economic Transformation. Cambridge University Press, Cambridge.

Brandt, L., Rawski, T.G., 2008b. China's great economic transformation. In: Brandt and Rawski (2008a), pp. 1–26.

Brandt, L., Rawski, T.G., Sutton, J., 2008. China's industrial development. In: Brandt and Rawski (2008a),, pp. 569–632.

Brandt, L., Hsieh, C.-T., Zhu, X., 2008. Growth and structural transformation in China. In: Brandt and Rawski (2008a), pp. 683–728.

Brandt, L., Ma, D., Rawski, T., forthcoming. From divergence to convergence: re-evaluating the history behind China's economic boom. Journal of Economic Literature.

Branstetter, L., Lardy, N., 2008. China's embrace of globalization. In: Brandt and Rawski (2008a), pp. 633–682.

Brezis, E.S., Krugman, P.R., Tsiddon, D., 1993. Leapfrogging in international competition: a theory of cycles in national technological leadership. American Economic Review 83, 1211–1219.

Broadberry, S., 1997. Anglo-German productivity differences 1870–1990: a sectoral analysis. European Review of Economic History 1, 247–267.

Broadberry, S., 1998. How did the United States and Germany overtake Britain? A sectoral analysis of comparative productivity levels, 1870–1990. Journal of Economic History 58, 375–407.

Broadberry, S., 2006. Market Services and the Productivity Race, 1850–2000: British Performance in International Perspective. Cambridge University Press, Cambridge.

Broadberry, S., 2013. Accounting for the Great Divergence. Paper presented to CAGE/CEPR Conference, Long-Run Growth: Unified Growth Theory and Economic History, University of Warwick.

Broadberry, S., Campbell, B., Klein, A., Overton, M. van Leeuwen, B., 2010. British economic growth: 1270–1870. University of Warwick CAGE Working Paper No. 35.

Broadberry, S., Crafts, N., 1996. British economic policy and industrial performance in the early post-war period. Business History 38, 65–91.

Brown, W., Bryson, A., Forth, J., 2008. Competition and the retreat from collective bargaining. National Institute of Economic and Social Research Discussion Paper No. 318.

Brynjolfsson, E., Hitt, L.M., 2003. Computing productivity: firm-level evidence. Review of Economics and Statistics 85, 793–808.

Buccirossi, P., Ciari, L., Duso, T., Spagnolo, G., Vitale, C., forthcoming. Competition policy and productivity growth: an empirical assessment. Review of Economics and Statistics.

Buera, F.J., Monge-Naranjo, A., Primiceri, G.E., 2011. Learning the wealth of nations. Econometrica 79, 1–45.

Cain, L.P., Paterson, D.G., 1986. Biased technical change, scale, and factor substitution in American industry: 1850–1919. Journal of Economic History 46, 153–164.

Cameron, G., Wallace, C., 2002. Macroeconomic performance in the Bretton Woods era and after. Oxford Review of Economic Policy 18, 479–494.

Carreras, A., Josephson, C., 2010. Aggregate growth, 1870–1914: growing at the production frontier. In: Broadberry, S., O'Rourke, K.H. (Eds.), The Cambridge Economic History of Modern Europe: 1870 to the Present. Vol. 2. Cambridge University Press, Cambridge, pp. 30–58.

Chandler, A.D., 1977. The Visible Hand: The Managerial Revolution in American Business. Harvard University Press, Cambridge, Massachusetts.

Clemens, M.A., Williamson, J.G., 2004. Why did the tariff-growth correlation change after 1950? Journal of Economic Growth 9, 5–46.

Collier, P., 2007. Growth strategies for Africa. Commission on Growth and Development Working Paper No. 9.

Collier, P., Goderis, B., 2012. Commodity prices and growth: an empirical investigation. European Economic Review 56, 1241–1260.

Collier, P., Gunning, J.W., 1999. Explaining African economic performance. Journal of Economic Literature 37, 64–111.

Comin, D., Dmitriev, M., Rossi-Hansberg, E., 2013. The spatial diffusion of technology. Mimeo.

Comin, D., Hobijn, B., 2010. An exploration of technology diffusion. American Economic Review 100, 2031–2059.

Comin, D., Hobijn, B., Rovito, E., 2006. Five facts you need to know about technology diffusion. NBER Working Paper No. 11928.

Conway, P., de Rosa, D., Nicoletti, G., Steiner, F., 2006. Regulation, competition and productivity convergence. OECD Economics Department Working Paper No. 509.

Corrado, C., Hulten, C., Sichel, D., 2009. Intangible capital and US economic growth. Review of Income and Wealth 55, 661–685.

Cosh, A.D., Guest, P., Hughes, A., 2008. UK corporate governance and takeover performance. In: Gugler, K., Yurtoglu, B.B. (Eds.), The Economics of Corporate Governance and Mergers. Edward Elgar, Cheltenham, pp. 226–261.

Costa, D., 2001. Estimating real income in the United States from 1888 to 1994: correcting CPI bias using Engel curves. Journal of Political Economy 109, 1288–1310

Crafts, N., 1989. Revealed comparative advantage in manufacturing: 1899–1950. Journal of European Economic History 18, 127–137.

Crafts, N., 1992a. Productivity growth reconsidered. Economic Policy 15, 387–414.

Crafts, N., 1992b. Institutions and economic growth: recent British experience in an international context. West European Politics 15, 16–38.

Crafts, N., 1999. East Asian growth before and after the crisis. IMF Staff Papers 46, 139–166.

Crafts, N., 2002. The Solow productivity paradox in historical perspective. CEPR Discussion Paper No. 3142.

Crafts, N., 2004a. Productivity growth in the industrial revolution: a new growth accounting perspective. Journal of Economic History 64, 521–535.

Crafts, N., 2004b. Steam as a general purpose technology: a growth accounting perspective. Economic Journal 114, 338–351.

Crafts, N., 2005. The first industrial revolution: resolving the slow growth/rapid industrialization paradox. Journal of the European Economic Association 3, 525–534.

Crafts, N. 2009a. Solow and growth accounting: a perspective from quantitative economic history. History of Political Economy 41, 200–220.

Crafts, N., 2009b. The Celtic Tiger in historical and international perspective. In: Mulreany, M. (Ed.), Economic Development 50 Years On, 1958–2008. IPA, Dublin, pp. 64–76.

Crafts, N. 2010. Cliometrics and technological change: a survey. European Journal of the History of Economic Thought, 17, 1127–1147.

Crafts, N., 2011. Explaining the first industrial revolution: two views. European Review of Economic History 15, 153–168.

Crafts, N., 2012. British relative economic decline revisited: the role of competition. Explorations in Economic History 49, 17–29.

Crafts, N. 2013a. Long-term growth in Europe: what difference does the crisis make? National Institute Economic Review, 224, R14–R18.

Crafts, N. 2013b. Returning to growth: policy lessons from history. Fiscal Studies 34, 255–282.

Crafts, N., Mills, T.C., 2005. TFP growth in British and German manufacturing: 1950–1996. Economic Journal 115, 649–670.

Crafts, N., Mills, T.C., 2009. From Malthus to Solow: how did the Malthusian economy really evolve? Journal of Macroeconomics 31, 68–93.

Crafts, N., Mulatu, A., 2006. How did the location of industry respond to falling transport costs in Britain before World War I? Journal of Economic History 66, 575–607.

Crafts, N., Toniolo, G., 1996. Postwar growth: an overview. In: Crafts, N., Toniolo, G. (Eds.), Economic Growth in Europe since 1945. Cambridge University Press, Cambridge, pp. 1–37.

Crafts, N., Toniolo, G., 2008, European economic growth: 1950–2005: an overview. CEPR Discussion Paper No. 6863.

Crafts, N., Venables, A.J., 2003. Globalization in history: a geographical perspective. In: Bordo, M.D., Taylor, A.M., Williamson, J.G. (Eds.), Globalization in Historical Perspective. University of Chicago Press, Chicago, pp. 323–364.

Criscuolo, C., Haskel, J., Martin, R., 2004. Import competition, productivity and restructuring in UK manufacturing. Oxford Review of Economic Policy 20, 393–408.

Crouch, C., 1993. Industrial Relations and European State Traditions. Clarendon Press, Oxford.

Das, M., N'Diaye, P., 2013. Chronicle of a decline foretold: has China reached the Lewis turning point? IMF Working Paper No. WP/13/26.

David, P.A., 1975. Technological Choice Innovation and Economic Growth: Essays on American and British Experience in the 19th Century. Cambridge University Press, Cambridge.

David, P.A., 1985. Clio and the economics of QWERTY. American Economic Review 75, 332–337.

David, P.A., Wright, G., 1997. Increasing returns and the genesis of American resource abundance. Industrial and Corporate Change 6, 203–245.

David, P.A., Wright, G., 1999. Early twentieth century productivity growth dynamics: an inquiry into the economic history of "our ignorance." University of Oxford Discussion Paper in Economic and Social History No. 33.

De Long, J.B., 1988. Productivity growth, convergence, and welfare: comment. American Economic Review 78, 1138–1154.

DeLong, J.B., Eichengreen, B., 1993. The Marshall Plan: history's most successful structural adjustment program. In: Dornbusch, R., Nölling, W., Layard, R. (Eds.), Postwar Economic Reconstruction and Lessons for the East Today. MIT Press, Cambridge, Massachusetts, pp. 189–230.

Denny, K., Nickell, S.J., 1992. Unions and investment in British industry. Economic Journal 102, 874–887.

Domar, E.D., 1970. The causes of slavery or serfdom: a hypothesis. Journal of Economic History 30, 18–32.

Duval, R., de la Maissonneuve, C., 2010. Long-run growth scenarios for the world economy. Journal of Policy Modeling 32, 64–80.

Easterly, W., Kremer, M., Pritchett, L., Summers, L.H., 1993. Good policy or good luck? Country growth performance and temporary shocks. Journal of Monetary Economics 32, 459–483.

Easterly, W., Levine, R., 2003. Tropics, germs, and crops: how endowments influence economic development. Journal of Monetary Economics 50, 3–39.

Eaton, J., Kortum, S., 1999. International technology diffusion: theory and measurement. International Economic Review 40, 537–570.

Edgerton, D.E.H., Horrocks, S.M., 1994. British industrial research and development before 1945. Economic History Review 47, 213–238.

Edquist, H., 2010. Does hedonic price indexing change our interpretation of economic history? Evidence from Swedish electrification. Economic History Review 63, 500–523.

Eichengreen, B., 1992. Golden Fetters: The Gold Standard and the Great Depression: 1919–1939. Oxford University Press, Oxford.

Eichengreen, B., 1996. Institutions and economic growth: Europe after World War II. In: Crafts, N., Toniolo, G. (Eds.), Economic Growth in Europe since 1945. Cambridge University Press, Cambridge, pp. 38–72.

Eichengreen, B., 2007. The European Economy Since 1945: Coordinated Capitalism and Beyond. Princeton University Press, Princeton.

Eichengreen, B., Park, D., Shin, K., 2012. When fast-growing economies slow down: international evidence and implications for China. Asian Economic Papers 11, 42–87.

Engerman, S.L., Sokoloff, K.L., 1997. Factor endowments, institutions, and differential paths of growth among new world economies: a view from economic historians of the United States. In: Haber, S. (Ed.), How Latin America Fell Behind: Essays on the Economic Histories of Brazil and Mexico: 1800–1914. Stanford University Press, Stanford, pp. 260–304.

Ennew, C., Greenaway, D., Reed, G., 1990. Further evidence on effective tariffs and effective protection in the UK. Oxford Bulletin of Economics and Statistics 52, 69–78.

Ergas, H., 1987. Does technology policy matter? In: Guile, B.R., Brooks, H. (Eds.), Technology and Global Industry: Companies and Nations in the World Economy. National Academy Press, Washington, DC, pp. 191–245.

Estevadeordal, A., Taylor, A.M., forthcoming. Is the Washington consensus dead? Growth, openness, and the great liberalization, 1970s–2000s. Review of Economics and Statistics.

Feinstein, C.H., 1981. Capital accumulation and the industrial revolution. In: Floud, R., McCloskey, D.N. (Eds.), The Economic History of Britain since 1700. vol. 1. Cambridge University Press, Cambridge, pp. 128–142.

Feinstein, C.H., Matthews, R.C.O., Odling-Smee, J.C., 1982. The timing of the climacteric and its sectoral incidence in the UK: 1873–1913. In: Kindleberger, C.P., di Tella, G. (Eds.), Economics in the Long View: Essays in Honour of W.W. Rostow, vol. 2. Macmillan, London, pp. 168–185.

Findlay, R., O'Rourke, K.H., 2007. Power and Plenty: Trade, War, and the World Economy in the Second Millennium. Princeton University Press, Princeton.

Fogel, R.W., 1964. Railroads and American Economic Growth: Essays in Econometric History. Johns Hopkins Press, Baltimore.

Foreman-Peck, J., 1991. Railways and late Victorian economic growth. In: Foreman-Peck, J. (Ed.), New Perspectives on the Victorian Economy: Essays in Quantitative Economic History, 1860–1914. Cambridge University Press, Cambridge, pp. 73–95.

Foreman-Peck, J., Hannah, L., 2012. Extreme divorce: the managerial revolution in UK companies before 1914. Economic History Review 65, 1217–1238.

Gerschenkron, A., 1962. Economic Backwardness in Historical Perspective: A Book of Essays. Harvard University Press, Cambridge Massachusetts.

Gilmore, O., 2009. Corporatism and Growth: Testing the Eichengreen Hypothesis. MSc. Dissertation, University of Warwick.

Goldin, C., Katz, L.F., 2008. The Race Between Education and Technology. Harvard University Press, Cambridge, Massachusetts.

Gómez-Galvarriato, A., Williamson, J.G., 2009. Was it prices, productivity or policy? Latin American industrialisation after 1870. Journal of Latin American Studies 41, 663–694.

Gragnolati, U., Moschella, D., Pugliese, E., 2011. The spinning jenny and the industrial revolution: a reappraisal. Journal of Economic History 71, 458–462.

Gregg, P., Machin, S., Metcalf, D., 1993. Signals and cycles? Productivity growth and changes in union status in British companies: 1984–1989. Economic Journal 103, 894–907.

Gregory, P.R., 1991. The role of the state in promoting economic development: the Russian case and its general implications. In: Sylla, R., Toniolo, G. (Eds.), Patterns of European Industrialization: The 19th Century. Routledge, London, pp. 64–79.

Griffith, R., 2001. Product market competition, efficiency and agency costs: an empirical analysis. Institute for Fiscal Studies Working Paper No. 01/12.

Griffith, R., Harrison, R., Simpson, H., 2010. Product market reform and innovation in the EU. Scandinavian Journal of Economics 112, 389–415.

Griliches, Z., 1996. The discovery of the residual: a historical note. Journal of Economic Literature 34, 1324–1330.

Habakkuk, H.J., 1962. American and British Technology in the 19th Century: The Search for Labour-Saving Inventions. Cambridge University Press, Cambridge.

Hall, P.A., Soskice, D., 2001. An introduction to varieties of capitalism. In: Hall, P.A., Soskice, D. (Eds.), Varieties of Capitalism: The Institutional Foundations of Comparative Advantage. Oxford University Press, Oxford, pp. 1–68.

Hall, R.E., Jones, C.I., 1999. Why do some countries produce so much more output per worker than others? Quarterly Journal of Economics 114, 83–116.

Hanushek, E.A., Wössmann, L., 2012. Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. Journal of Economic Growth 17, 267–321.

Harley, C.K., 1991. Substitution for prerequisites: endogenous institutions and comparative economic history. In: Sylla, R., Toniolo, G. (Eds.), Patterns of European Industrialization: The 19th Century. Routledge, London, pp. 29–44.

Harris, R., Siegel, D.S., Wright, M., 2005. Assessing the impact of management buyouts on economic efficiency: plant-level evidence from the United Kingdom. Review of Economics and Statistics 87, 148–153.

Harrison, M., 2002. Coercion, compliance, and the collapse of the Soviet command economy. Economic History Review 55, 397–433.

Haskel, J., 1991. Imperfect competition, work practices and productivity growth. Oxford Bulletin of Economics and Statistics 53, 265–279.

Hausmann, R., Pritchett, L., Rodrik, D., 2005. Growth accelerations. Journal of Economic Growth 10, 303–329.

Henisz, W.J., 2002. The institutional environment for infrastructure investment. Industrial and Corporate Change 11, 355–389.

Heston, A., Sicular, T., 2008. China and development economics. In: Brandt and Rawski (2008a), pp. 27–67.

Hicks, J., 1969. A Theory of Economic History. Oxford University Press, Oxford.

Hitiris, T., 1978. Effective protection and economic performance in UK manufacturing industry: 1963–1968. Economic Journal 88, 107–120.

Honohan, P., Walsh, B., 2002. Catching up with the leaders: the Irish hare. Brookings Papers on Economic Activity 1, 1–57.

Høj, J., Jimenez, M., Maher, M., Nicoletti, G., Wise, M., 2007. Product market competition in the OECD countries: taking stock and moving forward. OECD Economics Department Working Paper No. 575.

Hsieh, C.-T., Klenow, P.J., 2009. Misallocation and manufacturing TFP in China and India. Quarterly Journal of Economics 124, 1403–1448.

Hsieh, C.-T., Klenow, P.J., 2010. Development accounting. American Economic Journal: Macroeconomics 2, 207–223.

Hulten, C.R., 1979. On the "importance" of productivity change. American Economic Review 69, 126–136.

Humphries, J., 2013. The lure of aggregates and the pitfalls of the patriarchal perspective: a critique of the high wage economy interpretation of the British industrial revolution. Economic History Review 66, 693–714.

Isham, J., Woolcock, M., Pritchett, L., Busby, G., 2005. The varieties of resource experience: natural resource export structures and the political economy of economic growth. World Bank Economic Review 19, 141–174.

Ito, T., 2001. Growth, crisis, and the future of economic recovery in East Asia. In: Stiglitz, J.E., Yusuf, S. (Eds.), Rethinking the East Asian Miracle. Oxford University Press, Oxford, pp. 55–94.

James, J.A., Skinner, J.S., 1985. The resolution of the labor-scarcity paradox. Journal of Economic History 45, 513–540.

Jerven, M., 2010. African growth recurring: an economic history perspective on African growth episodes, 1690–2010. Economic History of Developing Regions 25, 127–154.

Jerzmanowski, M., 2007. Total factor productivity differences: appropriate technology vs. efficiency. European Economic Review 51, 2080–2110.

Jones, C.I., 1995. Time series tests of endogenous growth models. Quarterly Journal of Economics 110, 495–525.

Kindleberger, C.P., 1967. Europe's Postwar Growth: The Role of Labor Supply. Harvard University Press, Cambridge, Massachusetts.

Kindleberger, C.P., 1973. The World in Depression 1929–1939. University of California Press, Berkeley.

Kishtainy, N., 2011. Social Orders, Property Rights and Economic Transition: a Quantitative Analysis. Ph. D. thesis, University of Warwick.

Klein, A., Crafts, N., 2012. Making sense of the manufacturing belt: determinants of US industrial location: 1880–1920. Journal of Economic Geography 12, 775–807.

Knack, S., Keefer, P., 1995. Institutions and economic performance: cross-country tests using alternative institutional measures. Economics and Politics 7, 207–227

Kraay, A., 2000. Household saving in China. World Bank Economic Review 14, 545–570.

Krantz, O., Schön, L., 2007. Swedish Historical National Accounts, 1800–2000. Almqvist and Wiksell International, Lund.

Krugman, P., 1994. The myth of Asia's miracle. Foreign Affairs 73, 62–78.

Krugman, P., 2013. Hitting China's wall. New York Times, July 18.

Krugman, P.R., Venables, A.J., 1995. Globalization and the inequality of nations. Quarterly Journal of Economics 110, 857–880.

Kuznets, S., 1966. Modern Economic Growth: Rate, Structure, and Spread. Yale University Press, New Haven.

Lee, I.H., Syed, M., Lui, X., 2012. Is China over-investing and does it matter? IMF Working Paper No. WP/12/277.

Lehmann, S.H., O'Rourke, K.H., 2011. The structure of protection and growth in the late 19th century. Review of Economics and Statistics 93, 606–616.

Leunig, T., 2001. New answers to old questions: explaining the slow adoption of ring spinning in Lancashire: 1880–1913. Journal of Economic History 61, 439–466.

Levy, D.M., Peart, S.J., 2011. Soviet growth and American textbooks: an endogenous past. Journal of Economic Behavior and Organization 78, 110–125.

Lewis, W.A., 1954. Economic development with unlimited supplies of labour. The Manchester School 22, 139–191.

Lewis, W.A., 1969. Aspects of Tropical Trade 1883–1965. Almqvist and Wiksell, Uppsala.

Lewis, W.A., 1970. The export stimulus. In: Lewis, W.A. (Ed.), Tropical Development 1880–1913. Northwestern University Press, Evanston, pp. 13–45.

Lewis, W.A., 1978. Growth and Fluctuations 1870–1913. George Allen & Unwin, London.

Lipsey, R.G., Bekar, C., Carlaw, K., 1998. What requires explanation? In: Helpman, E. (Ed.), General Purpose Technologies and Economic Growth. MIT Press, Cambridge, Massachusetts, pp. 15–54.

Liu, D., Meissner, C. M., 2013. Market potential and the rise of US productivity leadership. NBER Working Paper No. 18819.

Lucas, R.E., 2000. Some macroeconomics for the 21st century. Journal of Economic Perspectives 14, 159–168.

Lucas, R.E., 2009. Trade and the diffusion of the industrial revolution. American Economic Journal: Macroeconomics 1, 1–25.

Machin, S., Wadhwani, S., 1991. The effects of unions on organisational change and employment. Economic Journal 101, 835–854.

Maddison, A., 1987. Growth and slowdown in advanced capitalist economies: techniques of quantitative assessment. Journal of Economic Literature 25, 649–698.

Maddison, A., 1995. Monitoring the World Economy: 1820–1992. OECD, Paris.

Maddison, A., 2005. Measuring and interpreting world economic performance 1500–2001. Review of Income and Wealth 51, 1–35.

Maddison, A., 2010. Statistics on world population, GDP and per capita GDP, 1–2008 AD. Available at <http://www.ggdc.net/maddison/Historical_Statistics/vertical-file_02-2010.xls>.

Maddison, A., Wu, H.X., 2008. Measuring China's economic performance. World Economics 9, 13–44.

Madsen, J.B., 2007. Technology spillover through trade and TFP convergence: 135 years of evidence for the OECD countries. Journal of International Economics 72, 464–480.

Magee, G., 2004. Manufacturing and technological change. In: Floud, R., Johnson, P. (Eds.), The Cambridge Economic History of Modern Britain, vol. 2. Cambridge University Press, Cambridge, pp. 74–98.

Matthews, R.C.O., Feinstein, C.H., Odling-Smee, J.C., 1982. British Economic Growth: 1856–1973. Oxford University Press, Oxford.

McMillan, J., Whalley, J., Zhu, L., 1989. The impact of China's economic reforms on agricultural productivity growth. Journal of Political Economy 97, 781–807.

Mehlum, H., Moene, K., Torvik, R., 2006. Institutions and the resource curse. Economic Journal 116, 1–20.

Melman, S., 1956. Dynamic Factors in Industrial Productivity. Basil Blackwell, Oxford.

Mitch, D., 1999. The role of education and skill in the British industrial revolution. In: Mokyr, J. (Ed.), The British Industrial Revolution: An Economic Perspective, second ed. Westview Press, Oxford, 241–279.

Morrison, C.J., 1993. A Microeconomic Approach to the Measurement of Economic Performance: Productivity Growth, Capacity Utilization, and Related Performance Indicators. Springer-Verlag, New York.

Morrisson, C., Murtin, F., 2009. The century of education. Journal of Human Capital 3, 1–42.

Mowery, D., 2009. Plus ca change: industrial R & D in the "third industrial revolution". Industrial and Corporate Change 18, 1–50.

Mowery, D., Rosenberg, N., 2000. Twentieth-century technological change. In: Engerman, S.L., Gallmann, R.E. (Eds.), The Cambridge Economic History of the United States. The Twentieth Century. Vol. 3, Cambridge University Press, Cambridge, pp. 803–925.

Murphy, K.M., Shleifer, A., Vishny, R.W., 1989. Industrialization and the big push. Journal of Political Economy 97, 1003–1026.

National Science Board, 2012. Science and Engineering Indicators, 2012. Washington, DC.

Naughton, B., 2008. A political economy of China's economic transition. In: Brandt and Rawski (2008a), pp. 91–135.

Nelson, R.R., Wright, G., 1992. The rise and fall of American technological leadership: the postwar era in historical perspective. Journal of Economic Literature 30, 1931–1964.

Nicholas, T., 2010. The role of independent invention in US technological development: 1880–1930. Journal of Economic History 70, 57–82.

Nickell, S., Nicolitsas, D., Dryden, N., 1997. What makes firms perform well? European Economic Review 41, 783–796.

Nicoletti, G., Scarpetta, S., 2003. Regulation, productivity and growth: OECD evidence. Economic Policy 36, 9–72.

North, D.C., 1990. Institutions. Institutional Change and Economic Performance. Cambridge University Press, Cambridge.

North, D.C., Thomas, R.P., 1973. The Rise of the Western World: A New Economic History. Cambridge University Press, Cambridge.

North, D.C., Wallis, J.J., Weingast, B.R., 2009. Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History. Cambridge University Press, Cambridge.

Ofer, G., 1987. Soviet economic growth: 1928–1985. Journal of Economic Literature 25, 1767–1833.

Ogilvie, S., 2007. "Whatever is, is right"? Economic institutions in pre-industrial Europe. Economic History Review 60, 649–684.

Ó Gráda, C., O'Rourke, K.H., 1996. Irish economic growth, 1945–1988. In: Crafts, N., Toniolo, G. (Eds.), Economic Growth in Europe since 1945. Cambridge University Press, Cambridge, pp. 388–426.

Ó Gráda, C., O'Rourke, K.H., 2000. Living standards and growth. In: O'Hagan, J. (Ed.), The Economy of Ireland: Policy and Performance of a European Region. Gill and Macmillan/St Martin's Press, Dublin, pp. 178–204.

Oliner, S.D., Sichel, D.E., Stiroh, K.J., 2007. Explaining a productive decade. Brookings Papers on Economic Activity 1, 81–152.

Olson, M., 1982. The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities. Yale University Press, New Haven.

O'Mahony, M., Timmer, M.P., 2009. Output, input and productivity measures at the industry level: the EU KLEMS database. Economic Journal 119, F374–F403.

O'Rourke, K.H., 2000. Tariffs and growth in the late 19th century. Economic Journal 110, 456–483.

O'Rourke, K.H., Williamson, J.G., 2002. When did globalisation begin? European Review of Economic History 6, 23–50.

Oulton, N., 1976. Effective protection of British industry. In: Corden, W.M., Fels, G. (Eds.), Public Assistance to Industry. Macmillan, London, pp. 46–90.

Oulton, N., 2012. Long-term implications of the ICT revolution: applying the lessons of growth theory and growth accounting. Economic Modeling 29, 1722–1736.

Pavitt, K., Soete, L., 1982. International differences in economic growth and the international location of innovation. In: Giersch, H. (Ed.), Emerging Technologies. Mohr, Tübingen, pp. 105–133.

Prados de la Escosura, L., Rosés, J., 2009. The sources of long-run growth in Spain, 1850–2000. Journal of Economic History 69, 1063–1091.

Prais, S.J., 1982. Productivity and Industrial Structure. Cambridge University Press, Cambridge.

Pratten, C.F., Atkinson, A.G., 1976. The use of manpower in British industry. Department of Employment Gazette 84, 571–576.

Prescott, E.C., 2004. Why do Americans work so much more than Europeans? Federal Reserve Bank of Minneapolis Quarterly Review 28, 2–13.

Pritchett, L., 1997. Divergence, big time. Journal of Economic Perspectives 11, 3–17.

Pritchett, L., 2000. Understanding patterns of economic growth: searching for hills among plateaus, mountains, and plains. World Bank Economic Review 14, 221–250.

Proudman, J., Redding, S., 1998. A summary of the openness and growth project. In: Proudman, J., Redding, S. (Eds.), Openness and Growth. Bank of England, London, pp. 1–29.

Redding, S., Venables, A.J., 2004. Economic geography and international inequality. Journal of International Economics 62, 53–82.

Rhode, P., 2002. Gallman's annual output series for the United States: 1834–1909. NBER Working Paper No. 8860.

Robertson, D.H., 1938. The future of international trade. Economic Journal 48, 1–14.

Rodríguez, F., Rodrik, D., 2001. Trade policy and economic growth: A skeptic's guide to the cross-national evidence. In: Bernanke, B., Rogoff, K. (Eds.), Macroeconomics Annual 2000. MIT Press, Cambridge, Massachusetts, pp. 261–325.

Rodrik, D., 1995. Getting interventions right: how South Korea and Taiwan grew rich. Economic Policy 20, 55–97.

Rodrik, D., 1997. TFPG controversies, institutions and economic performance in East Asia. CEPR Discussion Paper No. 1587.

Rodrik, D., Subramanian, A., 2005. From "Hindu Growth" to productivity surge: the mystery of the Indian growth transition. IMF Staff Papers 52, 193–228.

Rosenstein-Rodan, P., 1943. Problems of industrialisation of eastern and south-eastern Europe. Economic Journal 53, 202–211.

Rubinstein, W.D., 1992. The structure of wealth-holding in Britain, 1809–1839: a preliminary anatomy. Historical Research 65, 74–89.

Sachs, J.D., McArthur, J.W., Schmidt-Traub, G., Kruk, M., Bahadur, C., Faye, M., McCord, G., 2004. Ending Africa's Poverty Trap. Brookings Papers on Economic Activity 1, 117–240.

Sachs, J.D., Warner, A.M., 1995. Economic reform and the process of global integration. Brookings Papers on Economic Activity 1, 1–95.

Sachs, J.D., Warner, A.M., 1997. Fundamental sources of long-run growth. American Economic Review 87, 184–188.

Sandberg, L.G., 1979. The case of the impoverished sophisticate: human capital and Swedish economic growth before World War I. Journal of Economic History 39, 225–241.

Schneider, M.R., Paunescu, M., 2012. Changing varieties of capitalism and revealed comparative advantages from 1990 to 2005: a test of the Hall and Soskice claims. Socio-Economic Review 10, 731–753.

Schulze, M.-S., 2007. Origins of catch-up failure: comparative productivity growth in the Habsburg Empire: 1870–1910. European Review of Economic History 11, 189–218.

Sokoloff, K.L., Engerman, S.L., 2000. History lessons: institutions, factor endowments, and paths of development in the New World. Journal of Economic Perspectives 14, 217–232.

Sokoloff, K.L., Zolt, E.M., 2007. Inequality and the evolution of institutions of taxation: evidence from the history of the Americas. In: Edwards, S., Esquivel, G., Márquez, G. (Eds.), The Decline of Latin American Economies: Growth, Institutions, and Crises. University of Chicago Press, Chicago, 83–136.

Solow, R.M., 1957. Technical change and the aggregate production function. Review of Economics and Statistics 39, 312–320.

Song, Z., Storesletten, K., Zilibotti, F., 2011. Growing like China. American Economic Review 101, 196–233.

Sumner, M., 1999. Long-run effects of investment incentives. In: Driver, C., Temple, J. (Eds.), Investment, Growth and Employment: Perspectives for Policy. Routledge, London, pp. 292–300.

Svejnar, J., 2008. China in light of the performance of the transition economies. In: Brandt and Rawski (2008a), pp. 68–90.

Sylla, R., 1991. The role of banks. In: Sylla, R., Toniolo, G. (Eds.), Patterns of European Industrialization: The 19th Century. Routledge, London, pp. 45–63.

Sylla, R., Toniolo, G., 1991. Introduction: patterns of European industrialization during the 19th century. In: Sylla, R., Toniolo, G. (Eds.), Patterns of European Industrialization: The 19th Century. Routledge, London, pp. 1–26.

Symeonidis, G., 2008. The effect of competition on wages and productivity: evidence from the United Kingdom. Review of Economics and Statistics 90, 134–146.

Tanzi, V., 1969. The Individual Income Tax and Economic Growth. Johns Hopkins University Press, Baltimore.

Temin, P., 2002. The Golden Age of European growth reconsidered. European Review of Economic History 6, 3–22.

Temple, J., Johnson, P.A., 1998. Social capability and economic growth. Quarterly Journal of Economics 113, 965–990.

Timmer, M.P., Inklaar, R., O'Mahony, M., van Ark, B., 2010. Economic Growth in Europe: A Comparative Industry Perspective. Cambridge University Press, Cambridge.

Triplett, J.E., 1999. The Solow productivity paradox: what do computers do to productivity? Canadian Journal of Economics 32, 309–334.

van Ark, B., Melka, J., Mulder, N., Timmer, M., Ypma, G., 2003. ICT investments and growth accounts for the European Union. Groningen Growth and Development Centre Research, Memorandum GD-56.

van der Ploeg, F., 2011. Natural resources: curse or blessing? Journal of Economic Literature 49, 366–420.

Venables, A.J., 2010. Economic geography and African development. Papers in Regional Science 89, 469–483.

Voth, H.-J., 2001. The longest years: new estimates of labor input in England: 1760–1830. Journal of Economic History 61, 1065–1082.

Wacziarg, R., Welch, K.H., 2008. Trade liberalization and growth: new evidence. World Bank Economic Review 22, 187–231.

Wade, R., 1990. Governing the Market: Economic Theory and the Role of Government in East Asian Industrialization. Princeton University Press, Princeton.

Wallis, G., 2009. Capital services growth in the UK: 1950 to 2006. Oxford Bulletin of Economics and Statistics 71, 799–819.

Weitzman, M.L., 1970. Soviet postwar economic growth and capital-labor substitution. American Economic Review 60, 676–692.

Whalley, J., Xin, X., 2010. China's FDI and non-FDI economies and the sustainability of future high Chinese growth. China Economic Review 21, 123–135.

Williamson, J.G., 2011. Trade and Poverty: When the Third World Fell Behind. MIT Press, Cambridge, Massachusetts.

World Bank, 1993. The East Asian Miracle: Economic Growth and Public Policy. Oxford University Press, Oxford.

Wössmann, L., Lüdemann, E., Schütz, G., West, M.R., 2007. School accountability, autonomy, choice, and the level of student achievement: international evidence from PISA 2003. OECD Education Working Paper No. 13.

Wren, C., 1996. Industrial Subsidies: The UK Experience. Macmillan, London.

Wright, G., 1990. The origins of American industrial success, 1879–1940. American Economic Review 80, 651–668.

Wright, G., Czelusta, J., 2007. Resource-based growth past and present. In: Lederman, D., Maloney, W.F. (Eds.), Natural Resources: Neither Curse nor Destiny. Stanford University Press, Palo Alto, 183–211.

Xu, C., 2011. The fundamental institutions of China's reforms and development. Journal of Economic Literature 49, 1076–1151.

Young, A., 1995. The tyranny of numbers: confronting the statistical realities of the East Asian growth experience. Quarterly Journal of Economics 110, 641–680.

Zeira, J., 1998. Workers, machines, and economic growth. Quarterly Journal of Economics 113, 1091–1117.

CHAPTER SEVEN

Historical Development

Nathan Nunn
Department of Economics, Harvard University, NBER, and BREAD, 1805 Cambridge Street, Room M25, Cambridge, MA 02138, USA

Abstract

This chapter surveys a growing body of evidence showing the impacts that historical events can have on current economic development. Over the past two decades historical persistence has been documented in a wide variety of time periods and locations, and over remarkably long time horizons. Although progress continues to be made identifying and understanding underlying mechanisms, the existing evidence suggests that cultural traits and formal institutions are both key in understanding historical persistence.

Keywords

Persistence, Colonialism, Institutions, Norms, Culture, Path dependence

JEL Classification Codes

H11, N00, O10, 050, P51, R58, Z13

7.1. INTRODUCTION

In recent years, a new dynamic, empirical literature has emerged examining whether historical events are important determinants of current economic performance.[1] The origins of this literature can be traced to three lines of research that began approximately a decade and a half ago: Engerman and Sokoloff (1997, 2002), La Porta et al. (1997, 1998), and Acemoglu et al. (2001, 2002). Although each line of research had different motivations, what was common to them was that each provided an analysis, and supporting evidence, for how one important historical event—Europe's colonization of the globe—was important for long-term economic growth.

Since this time, the literature has developed in a number of ways. Most notably, other important events have also been examined. These range from systems of labor coercion, Africa's slave trades, medieval long-distance trade, Atlantic trade, the Protestant Reformation, overseas missionary work, the French Revolution, the Mexican Revolution, the forced opening of China's treaty ports, the adoption of new food crops during the

[1] See Nunn (2009) and Spolaore and Wacziarg (forthcoming) for recent reviews of this literature.

ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53538-2.00007-1

Columbian Exchange, the adoption of the plough, the invention of the printing press, the Neolithic Revolution, and various environmental catastrophes.

The typical study involves the collection and compilation of new and impressive data. Although this in and of itself is an important contribution, the real contribution is the use of the data to convincingly test hypotheses related to historical development. The most enlightening papers are able to trace the full impacts of a historical event through time, while examining specific channels and mechanisms.

This chapter provides a summary of this new literature. As we will see, once one surveys the progress made to date, it is impressive what the recent wave of quantitative historical studies has taught us about historical economic growth and development.

7.1.1 The Origins of the Literature

The origins of the historical development literature can be found in three sets of papers. What the three papers have in common is that they all examine European colonial rule. However, their motivations are very different.

The first study, written by economic historians Engerman and Sokoloff (1997), is a historical narrative, supported with descriptive statistics. In it they examine the importance of factor endowments and colonial rule for the subsequent economic development of colonies within the Americas. They argue that New World societies that were endowed with soil and climate suitable for growing lucrative, globally traded commodities, such as sugar, tobacco, and cotton, developed plantation agriculture, and with it, the use of slave labor. In the Spanish colonies, characterized by sizable indigenous populations and large reserves of gold and silver, forced labor was instituted. The use of slavery and forced labor resulted in economic and political inequality, both of which inhibited long-term economic development.

Interestingly, the other two seminal articles were not inherently interested in better understanding the history of European colonial rule. For example, the interest of Acemoglu et al. (2001) was in testing whether domestic institutions are a fundamental determinant of economic prosperity today. The interest of La Porta et al. (1997, 1998) was in identifying the causal impact of investor protection on financial development. What motivated both studies to examine colonial rule is the fact that the historical episode provides a source of variation in domestic institutions (in the case of Acemoglu et al.) and in investor protection (in the case of La Porta et al.). Both studies exploited European colonial rule as a natural experiment, focusing on a different dimension or characteristic that was argued to provide exogenous variation that they could use to identify their effect of interest.

La Porta et al. (1997, 1998) argue that because the legal tradition of the colonizer was transplanted to the colonies, the identity of the colonizer had an important impact on the legal system that evolved and, in particular, on contemporary investor protection. In particular, they show that former colonies with a legal code based on Roman civil law—these were the colonies of France, Spain, and Portugal—had weaker investor protection and less financial development relative to former British colonies with a legal system based on common law.

Acemoglu et al. (2001) argue that a primary determinant of the form (and long-term impacts) of colonial rule was the disease environment faced by potential European settlers. In temperate areas, like Canada, Australia, and the United States, European mortality rates were moderate enough to facilitate European settlement on a large scale. In these areas, the Europeans brought with them their values and beliefs and developed European-like institutions that emphasized the protection of property rights. In areas such as sub-Saharan Africa, where European mortality was high due to diseases like malaria and yellow fever, Europeans did not settle. Instead, they engaged in an extractive strategy. Rather than settling in a location, they set out to extract natural resources without regard for the consequences. Arguably, this strategy was facilitated by a lack of property rights and other similar institutions. Motivated by this historical narrative, Acemoglu et al. used a measure of early settler mortality as an instrument for contemporary domestic institutions to estimate the causal impact of institutions on long-term economic development.

The analysis of the three lines of research showcased how insights can be gained by examining economic growth and development from a historical perspective. Specifically, they showed how historical episodes can provide econometrically useful sources of exogenous variation. More importantly, they also showed that history matters and that it can have long-term persistent impacts that continue to influence growth and development today.[2]

Following these early studies, a large number of subsequent papers have emerged examining economic growth and development from a historical perspective. In the following section, I begin an overview of this literature by first describing a number of studies that examine other dimensions and aspects of European colonial rule, the historical event that has received the most attention in the literature. In Section 7.3, I then turn to an examination of studies that have investigated the long-term impacts of other historical events. These include the Columbian Exchange; various episodes of increased trade and globalization; episodes of warfare and armed conflict; expulsions and forced population movements; religious reformations; and important technological innovations. Following this, in Section 7.4, I turn to an important insight that has emerged from the literature: geography has important impacts on development today working through its impacts on historical events.

After having surveyed the evidence for the importance of history for contemporary economic development, I then turn to causal mechanisms. In Section 7.5, I summarize

[2] As with any seminal paper, extensions, comments, criticisms, criticisms of comments, criticisms of criticisms, etc. soon emerged. In an effort not to get lost in the weeds, I do not discuss these papers here. See for example, Easterly and Levine (2003), Glaeser (2004), Olsson (2004), Rodrik et al. (2004), Austin (2008), Albouy (2012), and Acemoglu et al. (2012).

the evidence that has been uncovered for the relative importance of various channels of persistence, including multiple equilibria and path dependence; domestic institutions; cultural values and beliefs; and genetic traits.

The final two sections of the chapter, Sections 7.6 and 7.7, discuss unresolved questions in the literature and offer concluding thoughts.

7.2. EUROPEAN COLONIAL RULE

7.2.1 Americas

The studies that examine the impacts of colonial rule in the Americas tend to focus on testing the hypothesis that initial endowments affected the extent of economic and political inequality, both of which were detrimental for long-term economic development (Engerman and Sokoloff, 1997). In a followup study, Engerman and Sokoloff (2005) provide additional evidence for their hypothesis by documenting a positive relationship between economic inequality and political inequality, measured by the inclusiveness of voting rights. Sokoloff and Zolt (2007) document a link between inequality and lower taxes on wealth and income and less spending on public goods such as education.

While the evidence put forth by Engerman and Sokoloff in support of their hypothesis primarily takes the form of historical narrative and descriptive statistics, a number of studies have undertaken more formal tests of their hypothesis. Bruhn and Gallego (2012) examine variation across 345 regions within 17 countries from North and South America. They identify a strong negative correlation between long-run development and initial colonial specialization, in what they call "bad" activities, which Engerman and Sokoloff (1997) argue display economies of scale and therefore relied heavily on exploited labor, e.g. sugar, coffee, rice, cotton, and mining. They provide additional evidence, also consistent with Engerman and Sokoloff (1997), that other activities, like subsistence farming, cattle raising, or manufacturing, are not negatively related to long-term development, unless there were large native populations that could potentially be exploited in the production process.

Naritomi et al. (2012) provide evidence consistent with Bruhn and Gallego (2012), but focus on Brazil and two commodities, gold and sugar. They examine variation across approximately 5000 Brazilian municipalities and quantify each municipality's historical involvement in the gold boom (during the 1700s) and the sugarcane boom (1530–1760). The authors show that the municipalities that experienced the sugar boom have greater land inequality today, while municipalities that experienced the gold boom have worse domestic institutions today.

The key mechanism in Engerman and Sokoloff's hypothesis is inequality, both economic and political. A number of studies provide evidence that calls into question their assertion that greater economic inequality is associated with greater political inequality and less development. Dell (2010) examines the *mita* forced labor system, which was instituted by the Spanish in Peru and Bolivia between 1573 and 1812. The *mita* system required that over 200 communities supply one seventh of their adult male population to work in the silver mines of Potosí and mercury mines of Huancavelica. The study combines contemporary household survey data, geographic data, and data from historical records; and uses a regression discontinuity estimation strategy to estimate the long-term impacts of the *mita* system. Dell's study exploits the fact that the boundary of the *mita* conscription area was clearly defined and that other relevant factors likely vary smoothly close to the *mita* boundary. Because of this, comparing the outcomes of *mita* and non-*mita* districts very close to the border provides an unbiased estimate of the long-term effects of the *mita*. The study finds that the *mita* system had an adverse effect on long-term economic development. All else equal, former *mita* districts now have an average level of household consumption that is approximately 25% lower than households in former non-*mita* districts. The study finds that a significant proportion of the difference can be explained by lower levels of education and less developed road networks.

Dell argues that the underdevelopment of *mita* districts was due to an absence of large haciendas. These haciendas lobbied the crown for public goods, like education and roads, and provided these goods directly. Therefore, in contrast to the Engerman-Sokoloff hypothesis, she finds better long-run development outcomes in locations with large haciendas and greater inequality (not less).

Acemoglu et al. (2008) also question Engerman-Sokoloff's inequality hypothesis. The authors first examine municipalities within Cundinamarca, Colombia and show that late 19th century land inequality is positively associated with late 20th century secondary school enrollment. They further question the presumption that economic and political inequality go hand in hand. After constructing a measure of political inequality using data on the identity of mayors for 4763 appointments held by 2300 different individuals between 1875 and 1895, they show that economic inequality and political inequality are not positively correlated. In fact, they argue that greater land inequality was better for long-term development because the landowners provided greater checks on the actions of the political elite.

Examining variation across US states and counties and across countries within the Americas, Nunn (2008b) also considers the role of inequality. Although he does find that, consistent with Engerman and Sokoloff, there is a negative relationship between slave use and current income, he fails to find evidence that inequality is the intervening channel. Although past slave use is positively correlated with historical and current inequality, controlling for historical land inequality does not reduce the negative impact of slavery on current income. Further, there is no relationship, either in the past or today, between inequality and income.

7.2.2 Asia

Early European contact with India occurred through overseas trade, beginning in 1613. Colonization of the subcontinent occurred through a number of battles. Beginning with the Battle of Plassey in 1757, the British East India Company (EIC) gained control

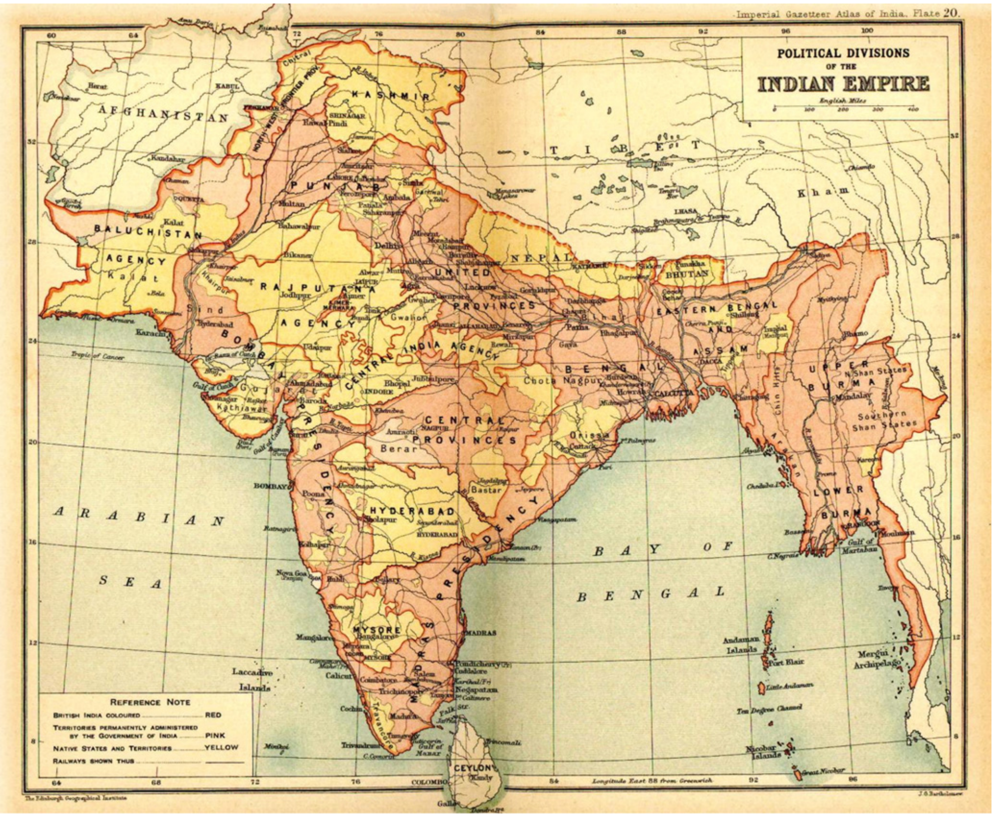


Figure 7.1 Directly ruled British districts and princely states within the Indian Empire. *Source: Imperial Gazeteer Atlas of India, Plate 20.*

of Bengal and Bihar, and by the early 19th century, the British controlled large parts of the Indian subcontinent, the other portions being the "princely states." The British EIC continued to annex the princely states during the 19th century until the mutiny of the British EIC's army (the Sepoy Mutiny) in 1857. After this, the British government ruled the subcontinent, establishing the British Raj.

The long-term impacts of British control on the Indian subcontinent have been examined empirically in a series of recent papers. Iyer (2010) examines the long-term impacts of direct British rule versus indirect rule—i.e. the princely states. The portions of the subcontinent under the two forms of rule (in 1909) are shown in Figure 7.1.

Looking across 415 districts, Iyer (2010) estimates the effect of direct British rule versus indirect British rule on investment in agriculture and agricultural productivity today. To help uncover causal estimates, she exploits the Doctrine of Lapse, a British policy that was in place between 1848 and 1856, that stated that a native ruler's adopted heirs were not to be recognized by the British government. This allowed the British to annex several states where the native ruler died without a natural heir. Iyer instruments for direct British rule using a district-specific indicator variable that equals one if the ruler died without a natural heir between 1848 and 1856, the period when the Doctrine of Lapse was in place. Examining the subset of states that had not yet been annexed by 1848, she shows that the IV procedure estimates no statistically significant difference between directly ruled districts and the princely states. This is in contrast to the OLS estimates which suggest that direct British rule is positively associated with agricultural investment and productivity. The most likely explanation for the difference is that the British annexed the most productive parts of the continent, which also had the greatest growth prospects.

In subsequent analysis, Iyer (2010) examines other contemporary outcomes, including the availability of public goods such as health, education, and roads. She continues to find that the IV estimates of the impact of direct British rule on public goods provision are lower than the OLS estimates, again suggesting that the British selected the "better" states, with greater long-run growth potential. In addition, for many of the public goods outcome variables, the IV estimates suggest that British rule actually exerted a negative long-term effect, a finding that is consistent with earlier research by Banerjee et al. (2005) who examine an even larger set of 27 different public good measures.

Other research, rather than examining differences between directly and indirectly ruled districts, looks at variation within the directly ruled districts of India. Banerjee and Iyer (2005) show that differences in the institutions initially implemented by the British had long-term growth effects. In particular, they examine the different revenue collection systems that were established and compare districts where revenue was collected directly by British officials to those where revenue was collected by native landlords. They find that after independence, districts with non-landlord systems have higher levels of health, education, and agricultural technology investments relative to landlord systems.

To determine the extent to which this correlation is causal, the authors exploit the fact that in the parts of India conquered between 1820 and 1856, non-landlord revenue collection was implemented. They argue that the historical determinants of the form of revenue collection are orthogonal to district characteristics and determined primarily by the date of British conquest, which they use as an instrument for the revenue collection system. Their IV estimates are consistent with the OLS estimates.

Overall, the existing body of evidence for India suggests that British control of the subcontinent, through the extent to which colonial policies took the form of direct rule vs. indirect rule, had lasting impacts on long-term economic growth.

7.2.3 Island Colonies

In a novel study, Feyrer and Sacerdote (2009) examine the experience of European colonial islands of the Pacific, Indian, and Atlantic Oceans. Although one could argue this is a somewhat obscure set of colonies to examine, the question they attempt to answer is of general interest. Their motive for looking at their particular sample stems from their interest in the obvious, but difficult-to-answer question of whether colonial rule, on average, was good or bad. In particular, is a longer period of colonial rule associated with better or worse long-term economic growth?

Feyrer and Sacerdote (2009) argue that, for islands, the date of discovery was determined in part by its location relative to prevailing wind patterns and that these wind patterns most likely do not affect long-term development through channels other than the island's date of discovery. They argue that the wind vectors surrounding an island can be used as instruments to estimate the causal effect of the length of colonial rule on subsequent development. Their baseline set of instruments, which are constructed from satellite imagery data, reported monthly on a one-degree by one-degree global grid, include the annual mean and variance of monthly east-west wind vectors.

Their first-stage estimates show that stronger westerly winds are associated with earlier discovery and more years under colonial rule. According to their 2SLS estimates, the length of colonial rule has a positive effect on per capita income in 2000. In other words, conditional on being a colony (within their sample), a longer period of colonial rule was better for economic development. They are quick to point out, however, that their results cannot address the question of whether or not the island colonies are better off because they were colonized.

7.2.4 Africa

A number of studies that examine the impacts of colonial rule within Africa find evidence of long-term impacts that persist until today. This is perhaps surprising since, for the vast majority of the continent, the period of colonial rule was short relative to the rest of the world. Africa was the last continent to be colonized, with the Berlin Conference of 1884/1885 marking the beginning of large-scale colonial rule within Africa.

An important source of evidence documenting the long-term impacts of colonial rule within Africa is Huillery (2009). In the study, her analysis combines data from historical documents from archives in Paris and Dakar with household surveys from the 1990s. She shows that looking across districts in French West Africa, there is a positive relationship between early colonial investments in education, health, and infrastructure and current levels of schooling, health outcomes, and access to electricity, water, and fuel. Most interestingly, she provides evidence of persistence that is specific to a particular public good and outcome. In other words, she finds that greater education spending during the colonial period is associated with more education in the post-colonial period, but not better health outcomes or more infrastructure. Similarly, she finds more infrastructure investment during the colonial period is associated with greater access to infrastructure today, but not with the other outcomes; and more health investments during the colonial period are associated with better health outcomes today, but not with the other outcomes.

While the exact mechanisms behind this somewhat extreme form of persistence are not yet well understood, she does provide evidence that early investments subsequently lead to more of the same investments. An alternative explanation is that persistent omitted factors, which have impacts that are public-good specific, are driving the results. However, her analysis undertakes a number of strategies to rule out this explanation, including matching districts based on geographic proximity.

A commonly cited adverse consequence of European colonial rule in Africa is that it resulted in the creation of country boundaries that paid little or no attention to pre-existing kingdoms, states, or ethnic groups. During the Berlin Conference of 1884/1885, the European powers divided among themselves lands that they had yet to explore, and lakes, rivers, and mountains that they had yet to discover. Although it has long been hypothesized that one of the legacies of colonial rule in Africa may be the artificial nature of national boundaries that it created, it was not until recently that this assertion was tested formally. Michalopoulos and Pappaioannou (2011) combine information on the pre-colonial locations of 834 ethnic groups from Murdock (1959) with the current boundaries of nation states and test for differences between ethnic groups that were partitioned by a country's border and those that were not. The authors create, for each ethnicity, two indicator variables. The first equals "1" if the ethnicity was partitioned by a border and greater than 10% of the area of the group lies on both sides of the border. The second equals "1" if the ethnicity was partitioned by a border but less than 10% of the area of the group is located on one side of the border. The examination of the two partition measures is motivated by potential measurement error due to imprecision in the location of ethnic group as mapped by Murdock (1959). Even if borders do not split ethnic groups, measurement error will generate splitting. This form of partitioning is more likely to occur in the second (less than 10%) measure.

The authors examine two sets of outcomes at the ethnicity level: economic development, measured by the density of night-time lights, and the number of civil conflicts between 1970 and 2005. They find that partitioned ethnic groups are associated with lower economic development (as measured by lower light density) and more civil war. Both partition measures are statistically significant, although the magnitude of the less than 10% measure is often smaller, which is consistent with this variable being measured with greater error. Therefore, their findings confirm the conventional belief that colonial rule, because of the way it artificially divided up the continent between European powers, had detrimental impacts.

Another potentially adverse consequence of colonial rule—particularly the policy of indirect rule—was that it often resulted in heightened hostilities between ethnic groups. In Rwanda, colonial policies intentionally deepened racial differences between the Hutus and Tutsis. The Census of 1933–1934 institutionalized the Hutu distinction, creating identity cards that reported individual ethnicity. In addition, an educational system with separate streams for Hutus and Tutsis was implemented. Prior to the arrival of the Europeans, the distinction between Hutus and Tutsis was more one of class than of race, and with much movement between the two groups. This is illustrated by the fact that it was actually quite difficult for the Belgians to group individuals neatly into one of the two groups. As a result, they developed the "Ten Cow Rule". When there was doubt, if an individual had more than ten cows, they were designated "Tutsi"; otherwise, they were "Hutu."

Many scholars, and perhaps most notably Mamdani (2001), have argued that the ethnic hostilities between the Hutus and Tutsis, culminating in the 1994 genocide, have their roots in Belgian colonial rule. However, there is far from full consensus on this issue. Vansina (2004), relying primarily on oral evidence and early written accounts, argues against this view. He instead argues that Hutu and Tutsi identities arose in the 17th century and became further entrenched during the 19th century, both of which occurred before the colonial period.

7.3. OTHER IMPORTANT HISTORICAL EVENTS

The literature's initial focus on European colonial rule was perfectly natural given that the event was one of the most important in human history and arguably the single most important for shaping the current distribution of the world's income. However, more recently, researchers have begun to turn to other important historical events to empirically examine their long-term impacts and importance for economic development today. I now turn to a discussion of these studies.

7.3.1 The Columbian Exchange

The Columbian Exchange refers to the transfer of crops, disease, ideas, and people between the Americas and the rest of the world following Christopher Columbus's voyage to the New World in 1492. The exchange brought diseases that decimated the Native American populations. It introduced the Eastern Hemisphere to a variety of new plants that were widely adopted, including tomatoes, the white potato, the sweet potato, cassava, corn, chillis, peppers, cacao, vanilla, and tobacco. In addition, Europeans were introduced to the chincona tree, which produces quinine, a prophylactic against malaria. The New World also provided abundant fertile land that could grow valuable Old World commodities, such as sugar and cotton.[3]

A number of papers have documented the impacts of the transfer of new foods from the Americas to the rest of the world. Nunn and Qian (2011) estimate the impact of the introduction of the potato to Europe, Asia, and Africa. Since the potato was calorically

[3] See Grennes (2007), Nunn and Qian (2010), and Mann (2011) for further descriptions of the Columbian Exchange.

and nutritionally superior to Old World staples like wheat, barley, rye, and rice, for the parts of the Old World that were able to adopt the potato, its diffusion from the New World resulted in a large positive shock to agricultural productivity. Using a difference-in-differences estimation strategy, the authors compare differences in population growth, urbanization, and adult heights before introduction relative to the population growth after introduction between locations that were able to adopt potatoes and those that were not. A location's ability to cultivate potatoes is measured using GIS-based climate and soil data from the FAO.

At the country level, they find that the introduction of the potato had a positive impact on total population and urbanization rates. Consistent with their urbanization finding, they also find that city growth, both globally and within Europe, was positively impacted by the potato. In an attempt to examine mechanisms more closely, the authors examine height data from France. They show that after the diffusion of the potato to France, individuals born in villages that could cultivate potatoes were between one-half and three quarters of an inch taller as adults.[4]

Other studies have also examined the impacts of the post-1492 diffusion of other food crops from the Americas. Chen and Kung (2012) examine the introduction of maize to China and find that although maize had a large positive impact on population, there is no evidence that it spurred urbanization rates. Jia (forthcoming) examines the diffusion of the sweet potato in China. One characteristic of the sweet potato is that it is much more drought resistant than the pre-existing staple crops in China, rice and wheat. Her analysis shows that prior to the sweet potato, there is a close relationship between the occurrence of drought and peasant uprisings. After the diffusion of the drought resistant sweet potato, this relationship weakened significantly.[5]

The Columbian Exchange not only brought New World crops to the Old World, but it also brought Old World crops to the New World. For many crops the soils and climates in the Americas were much more suitable for cultivation than in the Old World. The prime example is sugar. Hersch and Voth (2009) calculate the welfare gains in Europe from the increased supply of sugar that resulted from its large-scale cultivation in the Americas. They also consider the welfare gains from the introduction of tobacco, the highly popular New World crop, to Europe. They calculate that by 1850, the increased availability of sugar and tobacco had increased English welfare by approximately 6% and 4%, respectively. These results have important implications for our understanding of European well-being over time. It is generally presumed, based on real wage figures that do not account for new goods, that English welfare did not begin to improve until after

[4] Cook (2013b) provides evidence of a complementarity between milk and potatoes. He finds that the population effects of Nunn and Qian (2011) are greater among populations that exhibit lactase persistence (i.e. are lactose tolerant).

[5] This finding is particularly interesting because it shows that the impacts of weather shocks differ depending on the historical environment. This is a point that I return to below.

1800 (Clark, 2005). However, according to Hersch and Voth (2009) welfare increased significantly before this time from the availability of "new" goods that were a result of the Columbian Exchange.

7.3.2 International Trade and Globalization

A number of studies have traced the impacts of increased international trade in various periods of time. An important insight that has emerged from this literature is that international trade, by altering the evolution of domestic institutions, can have important effects for long-term growth.[6]

The seminal papers making this point are Greif (1993 and 1994), where it is shown how the development paths of two long-distance trading groups, the Maghribi and Genoese, have their origins in the manner in which trading contracts were enforced. Among the Maghribi, merchants relied on a collective enforcement strategy, where all merchants collectively punished any agent who had cheated; while among the Genoese, enforcement was achieved through an individual punishment strategy. These contracting institutions led to the development of different institutions outside of the trading environment. While the Genoese developed a formal legal system and formal organizations to facilitate exchange; the Maghribis continued to rely on collective informal enforcement mechanisms like group punishment.

A number of papers have empirically examined the long-term impacts of increased international trade and domestic institutions and long-term economic growth. The existing evidence seems to suggest that trade can have very different impacts depending on initial conditions and the specifics of the environment. The heterogeneous impacts of international trade are clearly illustrated in the study by Puga and Trefler (2012), which examines medieval trade in Venice between 800 and 1350. They show that trade initially changed the balance of power, which enabled new merchants to push for greater political openness (e.g. the end of the hereditary Dogeship and the begining of the Venetian parliament) and better contracting institutions (e.g. the colleganza). These institutional improvements led to economic growth. However, over time, wealth was increasingly concentrated in the hands of a relatively small number of merchants and this power was used to block further institutional reforms and limit political access. Therefore, international trade first led to the rise and then decline of growth-promoting inclusive institutions and economic growth.

The heterogeneous impacts of international trade can also be seen in the cross-section. For some parts of the world, evidence suggests that increased trade had beneficial impacts. Jha (2008) considers medieval India and shows that looking across cities within India, participation in overseas trade is associated with less religious conflict during the late 19th and early 20th centuries. Jha addresses the endogeneity of the selection of medieval

[6] On this point, see the recent survey by Nunn and Trefler (forthcoming).

ports by using the existence of natural harbors as an instrument for whether a coastal city was a trading port and by using propensity score matching techniques.

According to his estimates, being a town that was a medieval trading port made it less likely that the town later experienced Hindu-Muslim riots. Using historical evidence, Jha argues that because Muslims provided access to the markets of the Middle East, the returns to Hindu-Muslim cooperation were much higher in the towns connected to this overseas trade. As a result, institutions that supported exchange and a peaceful coexistence between Hindus and Muslims were developed.

The existing evidence suggests that international trade was also beneficial within Europe. Acemoglu et al. (2005) examine the impacts of the Atlantic three-corner trade on European institutions. They first show that the rise of Western Europe after the 16th century was driven by the economic growth of countries (and port cities) heavily involved in Atlantic trade. They argue that the primary benefit of the trade was not a direct benefit of the profits from the trade but an indirect impact that worked through domestic institutions. Profits from the trade shifted political power toward commercial interests and resulted in the development of growth-promoting institutions. They provide support for their hypothesis by showing that the Atlantic trade also resulted in better institutions, measured using an index of a country's "constraints on its executive." They also show empirically that among countries with access to the Atlantic trade, only for those that had non-absolutist institutions initially (i.e. in the 15th and 16th centuries) did trade generate improved institutions. If the monarchy was too strong, initially, it simply monopolized the trade—as in Spain and Portugal—and this limited the benefits to the commercial class and therefore limited institutional change.[7]

Beneficial impacts of international trade have also been estimated in 19th century China. Jia (forthcoming) examines the impacts of increased trade due to the forced opening to trade first imposed by Britain and later by the United States. The Treaty of Nanking (1842), which followed the Qing Dynasty's defeat by Britain during the First Opium War, named five cities as treaty ports: Guangzhou, Shanghai, Fuzhou, Ningbo and Xiamen (Fairbank, 1953). By 1896, 16 more treaty ports were added to the original five, with 28 more added between 1896 and 1911 (Tai, 1918; Wang, 1998).

Jia (forthcoming) examines a sample of 57 prefectures all with geographic access to overseas trade over 11 periods between 1776 and 2000. She estimates a difference-in-

[7] Another interesting source of heterogeneity related to the Atlantic trade is shown by Dittmar (2011). He examines the growth of port cities in Europe and shows that port cities that adopted the printing press by the late 15th century grew significantly faster than those that did not. According to his estimates, once one accounts for the importance of the printing press, Atlantic port cities no longer appear to grow faster than non-Atlantic port cities. These findings are consistent with Dittmar's hypothesis that the printing press, by making print media more widely available, fostered numeracy, literacy, and innovations in bookkeeping and accounting, all of which were particularly valuable in cities with significant commercial opportunities due to overseas trade.

difference specification, controlling for prefecture fixed effects and time period fixed effects. She finds that on average, prefectures with treaty ports experienced faster population growth during this period. The population increase occured prior to the Communist revolution of 1949, as well as after China's more recent period of increased openness after 1980. Between 1958 and 1980, when China's economy was heavily regulated, treaty ports did not experience faster growth.

The clear exception to the beneficial impacts of trade and institutional and economic development lies in the experiences of Africa and (arguably) Latin America, following the Age of Discovery. As we have seen, in the Americas, specialization of production in commodities that Bruhn and Gallego (2012) classify as "bad" or "ugly" (recall Section 7.2) led to long-term underdevelopment. Further, as we discuss below, the impacts of the Atlantic trade within Africa were extremely detrimental. Within Africa, participation in trade meant warfare, theft, and banditry to supply slaves for export to the Americas. As shown by Nunn (2008a) and Nunn and Wantchekon (2011), participation in the slave trade had long-term adverse consequences. The parts of Africa that were heavily involved in the trade today are poorer, have worse domestic institutions, and exhibit lower levels of trust. We discuss these impacts in further detail below.

7.3.3 Warfare and Conflict

History is filled with episodes of warfare and conflict. A number of recent studies show that many instances of violence have had important effects on human history. The most well-known hypothesis regarding warfare and long-term development is Tilly's (1990) hypothesis that an important determinant of the rise of Europe was interstate warfare which promoted the development of strong states. According to Tilly, beginning in the early-modern period, warfare and interstate competition resulted in the development of centralized governments and institutions that were able to raise sufficient capital and maintain large populations that could be used to wage war. In other words, according to Tilly, war made states.

While Tilly's argument has been very influential, very few studies have actually formally tested any version of the Tilly hypothesis. Aghion et al. (2012) provide evidence, which is very much in the spirit of Tilly, that in the modern period an increased threat of warfare is associated with increased education. The authors examine a yearly panel of 137 countries from 1830 to 2001. Controlling for country fixed effects and year fixed effects, they show that education, measured by primary school enrollment and education reforms, is positively associated with conflict in the previous 10 years and with a contemporaneous measure of the existence of a military rivalry with another country. In other words, they find evidence that war made education.

Proponents of the Tilly hypothesis have argued that it may also apply outside of Europe. For example, Bates (forthcoming) and Reid (2013) argue that interstate conflict bred larger, more centralized states. Bates (forthcoming) shows that among the African societies documented in the Standard Cross Cultural Sample (SCCS), there is a positive relationship between rates of warfare and the degree of political centralization.

Other evidence suggests that a history of warfare can also have negative long-term impacts. An innovative paper by Jancec (2012) provides evidence that interstate conflict has a long-term adverse impact on trust in the political system today. Using individual-level data from Slovenia, Croatia, Serbia, Montenegro, Romania, and Ukraine, the author shows that individuals living in regions that experienced more frequent changes in the ruling nation-state between 1450 and 1945 have lower levels of trust in political institutions today. In other words, populations whose ancestors more regularly experienced conquest trust the government less today. In addition, Jancec (2012) also finds that frequent border change is associated with greater identification with an individual's locality rather than with the nation, and with less participation in national politics, as measured by voting. Since Jancec's (2012) analysis always controls for broad region effects, he is able to rule out worse governance as the explanation for the findings. All estimates are derived holding national institutions constant. The most likely explanation is that a history of conquest generated less identification with the nation-state and less trust for national leaders and institutions.

A theme that has developed in the literature highlights important historical links between warfare and religion. For example, Botticini and Eckstein (2005, 2007) argue that the burning of the Second Jewish Temple by the Romans in 70AD had lasting important consequences for the Jewish religion and subsequent economic development. After the burning of the temple, the Jewish religion was transformed from one that centered around religious sacrifices in the Temple to one that required all Jewish males to read the Torah and teach their sons to do the same in the synagog. According to Botticini and Eckstein (2005), the literacy and numeracy generated by this religious requirement resulted in the migration of Jewish farmers to cities, beginning in the fifth and sixth centuries, where they engaged in urban occupations. In Babylon, Jews moved into urban centers and engaged in shopkeeping and artisanal activities such as tanning; linen and silk production; dyeing; and glassware-making. Jewish migration continued in the Muslim empire between the mid-8th and early-9th centuries, with Jews entering a variety of skilled occupations, including handicrafts and jewelry production; ship building; money lending; and long-distance trading. Overall, their analysis provides compelling evidence that one important violent event—the Roman burning of the Jewish Temple—had impacts that affected subsequent human capital accumulation and the trajectory of economic prosperity of the Jewish people.

Another example is the link between warfare with the Ottoman Empire and the rise of Protestantism in Europe. Iyigun (2008) tests the established hypothesis that the Ottoman military incursions into continental Europe from the mid-15th to late 16th centuries allowed Protestantism to develop. Chronologically, a cursory examination of the dates of Ottoman and counter-reformation conflicts is consistent with the hypothesis. For example, the deadliest of all religious wars, the Thirty Years War (1618–1648), followed the decline of the Ottoman Empire, marked by the Battle of Lepanto (1571) where the Holy League (a coalition of Catholic maritime states organized by Pope St. Pius V) defeated the Ottoman Empire's primary naval fleet. Iyigun shows that the hypothesis receives support when examined more rigorously. He analyzes annual data from 1450 to 1700 and shows that in years with more European-Ottoman conflict there is less conflict between countries within continental Europe, and there are less conflicts due to Catholic-Protestant religious differences.

Acemoglu et al. (2011a) examine the beneficial impacts of one of the most important episodes of European history: the French Revolution. Within a few short years (1789–1799), the revolution displaced traditional values regarding order and hierarchy with new enlightenment values of equality, citizenship, and inalienable rights. In 1792, the new republic declared war on Austria and its allies, including Prussia. Acemoglu et al. show that the institutional reforms that were imposed on conquered territories had lasting impacts in Germany and Prussia. Examining 19 regions at six different points in time between 1700 and 1900, the authors show that regions that experienced longer periods of French occupation between 1793 and 1815, subsequently experienced faster economic development, measured by urbanization rates. Constructing an index of reforms that quantifies the abolition of feudalism and guilds, and the implementation of the French civil code, they provide evidence consistent with these institutional reforms being the source of the increased urbanization rates. Locations that experienced longer French occupation also had more intensive reforms.

7.3.4 Expulsions and Forced Population Movements

Closely related to warfare and conflict are expulsions and forced population movements. The most dramatic example of forced population movement is the export of African slaves during the trans-Atlantic, trans-Saharan, Indian Ocean, and Red Sea slave trades. Slaves were captured through kidnappings, raids, and warfare. Historical accounts suggest that the pervasive insecurity, violence, and warfare had detrimental impacts on state formation, inter- and intra-group co-operation, and institutional, social, and economic development generally (e.g. Inikori, 2000, 2003).

The most illustrative example of this is the experience of the Kongo Kingdom, which was discovered in 1493 by Diogo Cão. Initially, a diverse array of products were traded between Kongo and the Portuguese, including copper, textiles, ivory, and slaves. The first slaves that were traded were prisoners of war and criminals. However, the increasing external demand for slaves, the presence of Portuguese slave traders, and competition for the throne within the Kingdom all resulted in a dramatic and uncontrollable increase in slave capture and raiding throughout the Kingdom. As early as 1514, King Afonso was already writing to the Portuguese to complain of Portuguese merchants colluding with noblemen to enslave Kongo citizens. In 1526, Afonso asked for the removal of all

Portuguese merchants and the end of trade. This attempt was unsuccessful and through the 16th century the process continued, culminating in the Jaga invasion of 1568–1570. The large-scale civil war from 1665 to 1709 resulted in the complete collapse of the Kingdom (Heywood, 2009).

Nunn (2008a) empirically examines the impacts the slave trades had on the long-term development of the African continent. Combining information from historical shipping records with information from a variety of historical sources—plantation inventories, marriage records, death records, slave runaway notices, etc.—that report the ethnic identity of the slaves shipped from Africa, Nunn constructs estimates of the total number of slaves shipped from modern-day African countries during each of the four slave trades.

The study finds that the parts of Africa from which the largest number of slaves were taken are the poorest today. The core issue in interpreting this correlation is selection into the slave trades. If, for example, the societies with the most poorly functioning institutions and the poorest future growth prospects selected into the slave trades, then this would explain the negative relationship even if the external trade in slave trades had no direct impact on societies within Africa. Nunn tests whether selection is driving the results by looking at the evidence on the nature of selection during the slave trades. He finds that the descriptive and quantitative evidence suggest that it was not the least developed societies that selected into the slave trade, but it was actually the more developed and more densely populated societies that supplied the largest numbers of slaves. Nunn also constructs instruments based on the distance of each country from the external locations of demand for the slaves. He argues that although the location of the demand for slaves influenced the location of supply, the reverse was not true. The IV estimates provide estimates consistent with the OLS estimates. Nunn concludes that the empirical evidence suggests that Africa's external trade in slaves did have a significant negative impact on the economic development of regions within Africa.

Subsequent studies have documented other important impacts of Africa's slave trades. Nunn and Wantchekon (2011) provide evidence that the slave trades adversely affected subsequent levels of trust within Africa. They show that the lower levels of trust arise through two channels: a deterioration of domestic institutions that enforce trustworthy behavior and an increase in the prevalence of cultural norms of distrust. They estimate that quantitatively the second determinant is about twice as important as the former.[8]

Dalton and Leung (2011) and Fenske (2012) provide evidence that the trans-Atlantic slave trade resulted in a long-term increase in the prevalence of polygamy. This is due to

[8] Deconinck and Verpoorten (forthcoming) update the results of Nunn and Wantchekon (2011) using the more recent (2008) round of the Afrobarometer survey. This increases the sample by two additional countries (from 17 to 19) and expands the number of ethnic groups from 185 to 228. With the more recent and expanded sample, they find estimates that are very similar to Nunn and Wantchekon (2011). Also see Pierce and Snyder (2012) who show that in countries that were more impacted by the slave trades, firms have less access to external financing today, whether it be through formal or informal means.

the fact that primarily males were captured and shipped to the Americas, which resulted in a shortage of men within Africa. Interestingly, there is no evidence of such an impact for the Indian Ocean slave trade, where there was not a strong preference for male slaves. Dalton and Leung (2011) conclude that Africa's history of the slave trades is the primary explanation for why polygamy is much more prevalent in West Africa than in East Africa, today.

Within Europe, studies have also found evidence of large persistent impacts of forced migration. Acemoglu et al. (2011b) examine the long-term impacts of the mass movement and murder of Jewish populations in Russia during World War II. Examining variation across 278 cities, the authors show that Jewish depopulation during the holocaust is associated with significantly slower population growth, which was still detectable 50 years later in 1989, the last year of their sample. The authors confirm these results by looking across 48 oblasts, identifying a relationship between Jewish depopulation and lower per capita income in 2002.

A number of studies have also examined the persistent impacts of the 1609 expulsion of approximately 300,000 Moriscos (Spanish Muslims) from the Iberian Peninsula. Chaney (2008) and Chaney and Hornbeck (2013) examine the effects in the Kingdom of Valencia, where 130,000 Muslims—equal to one-third of its total population—were expelled. Chaney and Hornbeck (2013) show that after the expulsion, total output responded quickly although total population did not, resulting in higher per capita incomes in districts where a greater share of the population had been expelled. The persistently higher output per capita is potentially explained by the presence of more extractive institutions with a higher tax rate that inhibited population growth. Chaney (2008) also examines the impacts of the 1609 expulsion in Valencia, but considers spillover impacts on neighboring districts as low-skilled migrants moved to newly available land.

Forced movements of indigenous populations were also common in the Americas. Dippel (2011) examines the long-term development impacts of forced integration of different tribal bands onto the same reservation in the 19th century. He measures forced integration by combining information on the indigenous integration of the bands within a tribe (specifically, whether bands within a tribe were politically integrated prior to the 19th century) with information on which bands were subsequently forced to live on the same reservation. Forced integration occurs when bands that were previously independent were forced to live on the same reservation. He finds that reservations that experienced forced integration have 30 percent lower per capita GDP in 2000. He provides convincing evidence that this effect is causal and that it is due to dysfunctional political institutions.

Feir (2013) considers the impacts of the policy of forcibly removing indigenous children from their homes and sending them to residential schools. She finds that in Canada the schools were successful in their intended goal of eroding indigenous culture.

Individuals attending residential schools, as adults are 16 percent less likely to participate in traditional activities and 8 percent less likely to speak their indigenous language. Residential schools were notorious for the presence of mental, physical, and sexual abuse. Collecting data on the number of proven abuse claims by school, Feir shows that attendance in the least abusive schools is associated with increased educational attainment and more employment. On the other hand, attending the most abusive residential schools is not associated with increased education, but is associated with lower employment, lower rates of marriage, and increased alcohol consumption.

7.3.5 Religion

A number of studies provide evidence of the persistent long-run impacts of important religious historical events. The episode that has received the most attention in the literature is the Protestant Reformation, whose origin dates back to October 31, 1517 when Martin Luther posted the *Ninety-Five Theses on the Power and Efficacy of Indulgences* on the doors of All Saints' Church in Wittenberg. He objected to corruption in the Catholic Church, and in particular, to the selling of indulgences. His teachings quickly spread, partly facilitated by the recent invention of the printing press (Rubin, 2011).

According to Weber (1930), the new religion that emerged, Protestantism, was significant because, in contrast to Catholicism, it approved the virtues of hard work and the accumulation of wealth and that these values provided the moral foundation that spurred the transition to a modern market-based industrial economy. Another significant feature of the Protestant religion is its emphasis on the ability of individuals to read the Bible. With this came a belief in the importance of education.

A large number of studies have examined the historical and persistent impacts of the Protestant religion. Becker and Woessmann (2009) examine the two potential impacts of Protestantism, namely increased education and a change in values related to accumulation, thrift, and hard work. Their analysis examines variation in the intensity of Protestant and Catholic denominations across 452 counties in late 19th century Prussia. They find that the Protestant religion is associated with higher literacy. To better understand whether the correlations reflect a causal impact of the Protestant religion, they use a county's distance from Wittenberg, the origins of the Reformation, as an instrument for the share of the population that is Protestant in 1871.[9] Using the same empirical structure, the authors also identify a positive impact of Protestantism on various measures of economic development. This finding is consistent both with Protestantism increasing education which increases income, and with Protestantism affecting beliefs and values which increase income. The authors attempt to disentangle the two by estimating the impact of Protestantism on income after netting out the level of income explained by education (which they estimate

[9] The determinants and dynamics of the adoption of Protestantism are an interesting subject of analysis in its own right. For more on this, see Rubin (2011) and Cantoni (2012).

directly and take from previous studies). The findings from this procedure indicate that Protestantism's positive impact on income can be almost fully explained by its impact on education.

The link between Protestantism and education also receives support from studies examining the long-term impacts of missionary activities outside of Europe. A large literature has emerged documenting this relationship in various locations and time periods. An early contribution is provided by Woodberry (2004), who documents a positive relationship between measures of the historical presence of missionaries and current per capita income and democracy across former non-settler colonies. According to Woodberry these benefits arise not only from increased education, but because missionaries, particularly Protestant missionaries, fought against injustices against native populations during colonial rule, which helped to foster better institutions, improved civil liberties, and increased democracy in the long-run (Woodberry, 2004, 2012).

Others have also examined the impact of missionary activities, but use a more micro-approach that focuses on a specific region or country. For example, Bai and Kung (2011b) look within China and examine county-level data from 1840 and 1920. They identify a positive relationship between Protestant missionary activity and economic development, measured using urbanization rates.

A recent insight within this literature is the identification of differences between religious denomination or orders within the Protestant and Catholic religions. Waldinger (2012) examines variation within colonial Mexico and shows differences in the long-term impacts of four different Catholic orders: the Franciscans, Dominicans, Augustinians, and Jesuits. She finds that the three Mendicant orders (Franciscan, Dominican, Augustinian), which shared a strong commitment to alleviating poverty and educating the poor, had a long-term impact on educational attainment. By contrast, the Jesuits, who focused their educational efforts on the colonial elites only, appear to have had long-term effects on conversion to Catholicism, but not on increased educational attainment.

Andersen et al. (2011) analyze the Catholic Order of Cisterians in England during the early modern period. One defining characteristic of the Catholic order, which after being established in France in 1098, quickly spread across England in the following century, was the belief and emphasis on a strong work ethic and promotion of thrift. Examining county-level data for England from 1377 to 1801, the authors show that counties with a greater presence of Cisterian monasteries exhibited greater population growth during this period.

Akçomak et al. (2012) empirically trace the impacts of the founding of the Brethren of the Common Life, a Roman Catholic community established by Geert Groote in the late 14th century. The movement arose because of dissatisfaction with the Catholic Church and set to reform the Church by educating citizens and enabling them to read the Bible in the vernacular. In addition to their strong emphasis on literacy and education, the Brethren of the Common Life also promoted hard work and productive labor.[10] The authors empirically trace the historical impact of the Brethren of the Common Life within the Netherlands. Examining a sample of Dutch cities, the authors show that cities with Brethren of the Common Life communities had higher rates of literacy in 1600, more book production from 1470 to 1500, and faster population growth between 1400 and 1560. Of course, these correlations may be driven by reverse causality or omitted variables bias. The authors attempt to better understand whether the correlations are causal by using a city's distance from Deventer, the birthplace of Geert Groote and the origins of the movement.

Gender differences between the Protestant and Catholic religions are another aspect that has been examined by the literature. Because Protestants believed that reading the Bible directly was important for salvation, even for women, they placed greater importance on female education than Catholics. Using data from the first Prussian census of 1816, Becker and Woessmann (2008) show that Protestantism is associated with a smaller gender gap in education. Evidence for a greater emphasis on female education among Protestants is also found in Nunn's (forthcoming) analysis of the impacts of Catholic and Protestant colonial African missions on long-term education. He finds that although both had positive impacts on long-term education, the impact of Protestant missions was concentrated among females, while the impact of Catholic missions was concentrated among males.

7.3.6 Technological Innovation

Findings from a number of recent studies suggest a link between innovative activities in the past and subsequent economic outcomes. For example, Comin et al. (2010) document a positive correlation between the measure of a society's level of technology in the past (either 1000 BC, 0AD, or 1500AD) and either its level of technology or per capita income today. The authors hypothesize that this is driven by increasing returns to technology adoption: a higher level of technology lowers the cost of discovering new technologies. That is, a higher level of technology in the past affects the ease of accumulating subsequent technology which impacts technology in the future. Of course, their findings are also consistent with omitted persistent factors impacting both technology and development in the past and today. An example is the persistence of governance and institutional quality as has been documented by Bockstette et al. (2002).

Other studies, by zooming in on specific innovations, have been more successful at establishing persistent long-term impacts. Dittmar (2011) examines the long-term effects of the printing press, which was first established in Mainz, Germany between 1446 and 1450. He constructs a panel of European city-level data at 100-year intervals

[10] The similarity between Protestant beliefs and the Brethren of the Common Life is not a coincidence, as Martin Luther studied under the Brethren of the Common Life at Magdeburg before attending university.

between 1300 and 1800, combining data on city populations with information on the early adoption of the printing press. His analysis shows that cities that adopted the printing press between 1450 and 1500 experienced faster population growth during the 16th to 19th centuries. The impacts he estimates are extremely large. They imply that the printing press accounts for 18% of city growth between 1500 and 1600.

Dittmar uses the panel dimension of his data to validate his cross-sectional finding, showing that cities that adopted the printing press in the late 15th century were not growing more quickly prior to adoption. This suggests that the results are not driven by unobserved time-invariant differences between cities. He also provides additional evidence for a causal interpretation of his estimates using distance from the invention of the printing press—Mainz, Germany—as an instrument for adoption in the late 15th century. The IV estimates are consistent with the OLS estimates.

In a follow-up study, Dittmar (2012) calculates the impact of the printing press on aggregate welfare. Using data from England on the price and consumption of printed books in England between the 1490s and 1700 (and assumptions about consumers' utility functions), he estimates that the printed book increased welfare by an equivalent of 4% of income; by the mid-17th century this figure was 3–7%.

Baten and van Zanden (2008) provide complementary evidence of the importance of the printed book for long-term growth. The authors construct an impressive dataset of the production of printed books in eight Western European countries every 50 years between 1450 and 1800. The authors show that book production correlates strongly with literacy, and in panel regressions with time-period fixed effects, initial per capita book production is positively associated with faster growth in real wages during the next 50 years.

Alesina et al. (2013) examine the long-term impacts of the plough, an important technological innovation used in agriculture. The tool, which was able to prepare large amounts of soil for cultivation in a shorter period of time than previous tools, was first invented between 6000 and 4000 BC in Mesopotomia (Lal et al. 2007). Although the impacts of the plough were likely vast, the authors focus on one consequence that was highlighted by Boserup (1970). Because the use of the plough required significant upper body strength, it tended to generate a gender division of labor where men worked outside the home in the fields while women specialized in home production and other domestic activities. Boserup argues that this gender division of labor resulted in deeply held beliefs about the role of women in society. In societies that traditionally use plough agriculture, less equal beliefs about the roles of men and women evolved. Alesina et al. (2013) test this hypothesis by linking ethnographic data with contemporary individual- and country-level measures of gender role attitudes. They find that traditional plough agriculture is associated, even today, with less equal beliefs about the roles of men and women in society.[11]

[11] Also see Hansen et al. (2012), who find that a longer history of agriculture is associated with more unequal gender roles.

The findings of Alesina et al. are consistent with evidence that in the early phase of agriculture, prior to the adoption of the plough, societies tended to be matriarchal and characterized by gender equality (Gimbutas, 2007). Recent excavations from Çatalhöyük, a Neolithic town of 8000 people on the plains of central Turkey inhabited approximately 9000 years ago, provide additional evidence of the equality of the sexes during this time (Hodder, 2005). Analysis of male and female skeletal remains shows carbon deposits inside the ribs, due to indoor wood fires and a lack of ventilation in the homes. The smoke was ingested causing soot to build up in the lungs resulting in a lining of carbon inside the ribs. Hodder (2005) finds that the average amount of carbon in the ribs of men and women was equal, suggesting that men and women tended to spend roughly equal amounts of time within and outside the home. In addition, the archeological evidence from Çatalhöyük suggests that men and women had similar diets and were buried in similar positions and locations, both of which also suggest roughly equal social status.

The growth-promoting impacts of the plough have been studied by Andersen et al. (2013). Examining 316 European regions between 500 and 1300AD, the authors show that the adoption of the heavy plough is associated with greater population growth and increased urbanization. According to the authors' diff-in-diff estimates, the heavy plough accounts for 10 percent of the increase in population and urbanization during this time.

7.4. GEOGRAPHY AND HISTORY

7.4.1 The Historical Impacts of Geography

One of the important insights that has arisen from the historical development literature concerns the relationship between geography and contemporary development. Specifically, a common finding in the literature is that geography can have important impacts on current development through its persistent historical effects. Further, evidence also suggests that this historical impact of geography may be larger than its contemporaneous impact. For example, the findings from Acemoglu et al. (2001) show that the disease environment at the time of European colonization crucially affected subsequent institutional development. The authors argue that the primary impact of a country's disease environment works through this historical channel rather than through contemporary channels. The line of research by Engerman and Sokoloff (1997, 2002) also shows that small geographic differences become magnified through historical events and as a result end up having large impacts on long-term economic development. As they argue, differences in soil and climate made plantation agriculture and its reliance on slavery more or less profitable in different parts of the Americas, which in turn affected long-term economic development.

Even more dramatic examples of the long-term historical effect of geography are documented by Jared Diamond in his book *Guns, Germs and Steel*. The book is devoted to exploring the answer to the question of why Europeans colonized the rest of the world and not the other way around. Part of Diamond's answer lies in the fact that in Eurasia, crops and animals were domesticated earlier and in more varieties than in other parts of the world.

In addition, the domestication of plants and animals quickly spread east and west throughout Eurasia, but diffused much less quickly south to the African continent. When moving east or west, the length of the day does not change, and the climate is generally not drastically different. However, this is not true when moving north or south, where the length of the day changes and the climate typically is very different. More generally, for continents with a north-south orientation, such as the Americas or Africa, domestication or technological advance tended not to spread as quickly as in Eurasia with its east-west orientation.

Because of the early domestication of animals in Eurasia (and its more rapid diffusion), humans lived in close proximity to animals. As a result of this, new animal-based diseases, such as measles, tuberculosis, influenza, and smallpox developed, and over time humans developed genetic resistances to the diseases. In contrast, the parts of the world without domesticated animals did not develop the diseases or genetic resistance. According to Diamond, this explains why European diseases decimated native populations and not the other way around. As Diamond points out, the spread of disease was as important a factor as the military for European conquest of the Americas.

Diamond's explanation for Europe's global dominance illustrates clearly the large effect that geography can have through history. The historical origins of European colonization of the globe lie in two deep determinants: (i) being endowed with wild plants and animals suitable for domestication; and (ii) being located on a continent with an east-west orientation.

Although Diamond's hypothesis is intuitive in many ways, there are reasons to be sceptical. First, having domesticated plants and animals is potentially endogenous. For example, Diamond asserts that although the horse was domesticable, its close relative the zebra was not (Diamond, 2002). However, this assertion is difficult to assess since we do not observe the wild ancestors of the horse and so cannot compare it to the zebra. All we observe is the domesticated version, which has undergone centuries of selective breeding. Perhaps there are other historical determinants—be they economic, cultural, institutional, etc.—that caused horses to become domesticated in Eurasia but not zebras in Africa. Interestingly, there are examples of Europeans attempting to tame zebras. Rosendo Ribeiro, a doctor in Kenya, made house calls on a zebra. In England, Lord Walter Rothschild, pictured in Figure 7.2, would frequently drive a carriage pulled by zebras through the streets of London. However, despite these examples, the zebra never become widely domesticated.

Olsson and Hibbs (2005) take Diamond's hypothesis to the data. Using modern countries as the unit of analysis, the authors show that consistent with Diamond's descriptive accounts, countries with richer biological and geographic environments experienced the

Figure 7.2 Lord Lionel Walter Rothschild with his zebra-drawn carriage, 1895. *source: The Picture Magazine.*

transition to agriculture at an earlier date and have higher per capita GDP in 1997. The geographic environment is measured using an index that includes the axis orientation of the continent, suitability of the climate for agriculture, latitude, and the size of the landmass within which the country is located. The measure of biological conditions is based on an index that comprises the number of annual or perennial wild grasses known to exist in prehistory and with a mean kernel weight exceeding 10 mg, as well as the number of domesticable mammals known to exist in prehistory and weighing more than 45 kg. Overall, the authors find that their estimates confirm Diamond's hypotheses.

Evidence from a wide range of empirical studies provides additional evidence of the importance of historical impacts of geography. Ashraf and Michalopoulos (2011) provide evidence that geography was an important determinant of the timing of the Neolithic Revolution, arguably the most important event in human history. Looking across countries globally, and across archaeological sites within Europe and the Middle East, they document an inverted U-shaped relationship between year-to-year variability in temperature and early adoption of agriculture.[12]

Michalopoulos (2012) shows that geography was an important determinant of the evolution of ethnic identity and hence ethnic diversity (which is known to be highly correlated with economic development today). Michalopoulos (2012) provides evidence

[12] Because fine-grained temperature data are not available prior to 1500, the authors are forced to use post-1500 variability as a proxy. The assumption is that the rank ordering of variability after 1500 is similar to the ordering prior to 1500. They show that this is true comparing data from 1500 to 1900 and 1900 to 2000.

that the pattern of agricultural suitability and terrain slope were important determinants of the interaction between ethnic groups and their proclivity to merge into and identify as larger ethnicities. His analysis combines fine-grained geographic data with information on the locations of ethnic groups globally. The world is then divided into grid-cells that are 2.5° by 2.5°. Michalopoulos (2012) shows that grid-cells that exhibit greater variation in soil quality and in elevation are also more linguistically diverse. The most likely explanation for the finding is that greater geographic variation prevented trade and migration between societies, and conquest of one society over others, all of which have homogenizing impacts.

Interestingly, Michalopoulos (2012) shows that the link between geography and ethnic diversity is due to geography's impact prior to 1500. Among the parts of the world that witness significant population changes after 1500 (due to death and voluntary and involuntary migrations), there is no relationship between geographic diversity and linguistic diversity.

Durante (2010) provides evidence that within Europe, historical variability in weather conditions created greater benefits for cooperation, which increased the level of cooperation in societies. He hypothesizes that greater spatial variability in temperature and precipitation generates output shocks that are less correlated, providing an increased incentive for trade, thus increasing trust and cooperation. As well, greater temporal variability of weather increases the benefits to large storage facilities and irrigation, which require large-scale cooperation. Durante therefore argues that locations characterized by greater spatial and temporal variability may have higher levels of trust and cooperation today. He tests these predictions using monthly historical climatic data from 1500 to 2000, measured across grid-cells within Europe. He finds that greater year-to-year variability in both temperature and precipitation is associated with higher levels of trust today, and that less correlated weather shocks over space is also associated with more trust today.

Of course, there are a number of potential alternative explanations for these correlations. Therefore, as a further test of his channel, Durante measures variability in growing season months and months outside of the growing season. He finds that only historical variability during the growing season is correlated with current trust. He also examines weather variability from 1500 to 1750, which was prior to the industrial revolution when Europe was primarily agricultural, and from 1900 to 2000, which is after industrialization. He finds that only weather variability during the agricultural period is correlated with trust today.[13]

Recent findings from Alsan (2012) suggests that geography also had important historical impacts within Africa. A large literature attributes many of Africa's unique

[13] A subsequent study by Ager and Ciccone (2012) raises the questions of whether the increased trust found in Durante (2010) is due, in part, to increased religiosity. Although in a different context—the 19th century United States—Ager and Ciccone (2012) show that increased variability in annual rainfall (looking across counties) is associated with increased church membership.

characteristics to the fact that it is land abundant and labor scarce. Alsan (2012) considers a potential explanation for this: the tsetse fly. The fly, which is unique to Africa, transmits the parasite *trypanosomiasis*, which causes sleeping sickness in humans and nagana in domesticated animals. The tsetse fly, both directly through its impact on humans, and indirectly through its impact on domesticated animals, may be responsible for Africa's low population densities historically.

The author uses 19th century climate data measured at the grid-cell level to construct a measure of the historical suitability of each cell for the tsetse fly. The index is a highly nonlinear function of temperature and humidity. Examining variation across ethnic groups within Africa, she shows that ethnicities with climates more suitable for the tsetse fly, at the end of the 19th century, were less likely to use draft animals for trade and agriculture, were less likely to use the plough, and were more likely to use shifting cultivation rather than more intensive agricultural techniques. Because tsetse-suitable areas did not develop plough agriculture, women were more likely to participate in agriculture, and because the use of animals was not feasible, slaves were more likely to be used. Additionally, the less intensive agricultural techniques resulted in lower population densities, fewer urban centers, and less developed states. Her findings provide strong evidence that geographic suitability for the tsetse fly had a formative impact on the nature and prosperity of societies within Africa.

Because the tsetse fly did not exist outside of Africa, Alsan is able to undertake a falsification test by examining the correlation between tsetse suitability and the outcome of interest in the other parts of the world. If her estimates are really capturing the causal impact of the tsetse fly on long-term development, then in the parts of the world where there was no fly, we should not observe the same correlations. This is indeed what she finds. The tsetse suitability index has no predictive power outside of Africa. Overall, her findings provide strong evidence that the tsetse fly, by inhibiting the development of intensive agriculture using draft animals, resulted in lower populations, less urbanization, and less state development.

Fenske (2011) also considers the question of how geographic conditions affected the history of state development in Africa. The author tests the hypothesis that ecological diversity, by increasing the benefits of peaceful exchange between locations, increased the need of a state to provide the institutional setting to facilitate trade. This in turn resulted in the development of larger more developed states. Combining data on the boundaries of African ethnic groups in the 19th century with information on 18 ecological zones within Africa, Fenske constructs a measure of each ethnic group's ecological diversity. He finds that ethnic groups that were more ecologically diverse also had larger and more developed states.

A large number of studies also examine historical weather shocks and show that they had important historical impacts, many of which continue to be felt today. For example, Fenske and Kala (2013) show that during the slave trade, cooler temperatures near the slave ports were associated with increased slave exports. Therefore, due to the persistent impacts of the slave trade, temperature fluctuations during this time period had long-term impacts. Bai and Kung (2011a) examine the impacts of rainfall on Sino-Nomadic attacks in Han China between 220BCE and 1839CE. They identify a negative relationship between conflict and precipitation showing that climate was also an important determinant of conflict in the region.

Chaney (forthcoming) shows that in Ancient Egypt deviant Nile floods had important politcial impacts. Because deviant floods increased social unrest, this increased the bargaining power of the religious leaders relative to military leaders. Consistent with this, Chaney (forthcoming) shows that from 641 to 1437CE, deviant floods are associated with higher food prices, more conflict, less turnover of the highest ranking religious leader, and more construction of religious structures (relative to secular ones).

Haber and Menaldo (2010) also argue that climate can have important political effects. They show that there exists an inverted U-shaped relationship between average rainfall and democracy. They argue that this relationship is explained by the non-linear relationship between rainfall and suitability of a location for sedentary agriculture, which they argue provides a foundation more suitable for democracy than nomadic modes of subsistence. Bentzen et al. (2012) also argue for a link between geography/climate and modern political institutions, but, motivated by Wittfogel (1957), focus on the extent to which a location's agricultural output is increased by investments in irrigation. They argue that the large-scale investment and coordination needed for irrigation promoted strong authoritarian leadership and autocratic institutions, and this has persistent impacts even today. Using data from the FAO on yields with and without irrigation, the authors construct a measure of irrigation potential. Looking across 160 countries, they find greater irrigation potential is associated with less democracy today. Bentzen et al. (2012) show that in their specification the non-linear effect found in Haber and Menaldo (2010) no longer exists once one controls for their measure of irrigation potential.

Overall, there is a large body of evidence—only some of which is reviewed here—that suggests that a significant effect of geography—if not the largest effect of geography—on current economic development arises due to its influence on past events rather than through its direct effect on economic outcomes today.

7.4.2 Geography's Changing Impact Over Time and Space

Once one recognizes the fact that geography had important impacts historically as well as today, it is natural to ask whether the impacts of geography have been roughly constant throughout time or whether its impact varies in a systematic manner across time and/or space. This is a point also addressed in Acemoglu et al. (2001). Their empirical and historical narrative is that the disease environment generally, and in particular today, does not have large impacts on economic development. However, during the period of

European colonization of much of the globe it had a crucial impact. In locations with a disease environment that threatened European survival, Europeans did not migrate and did not establish growth-promoting institutions. Acemoglu et al.'s (2001) assumption that this disease environment only mattered during this specific historical episode is what allows them to use initial settler mortality as an instrument for a country's domestic institutions in explaining current per capita income. According to them, this particular geographic characteristic—the severity of the disease environment for Europeans—only had impacts during the colonial period.

Along similar lines, a number of papers find evidence that weather shocks can have significant long-term impacts during specific windows of time and no long-term effects during others. For example, Dell (2012) shows that within Mexico, drought experienced by municipalities between 1906 and 1910 had a large positive impact on violence and insurgency during the Mexican Revolution (1910–1918), resulting in a greater prevalence of *ejidos* (communal farms), which are less economically developed today. This implies that drought experienced between 1906 and 1910 had a long-term persistent impact on underdevelopment in Mexico. She shows that drought in other periods (between 1960 and 1995) are uncorrelated with long-term development.

Osafo-Kwaako (2012) also finds evidence of weather shocks mattering during a specific window of time. He shows that within Tanzania and during the early process of the government's establishment of development villages in the early 1970s, drought provided a motivation for peasants to agree to the villagization process. One therefore observes a positive relationship between droughts in 1973–1975 and the subsequent extent of villagization. The author then documents the persistent impacts of villagization. Although it increased education levels, political awareness, and community participation, it has also led to increased poverty and lower consumption today. Like Dell, Osafo-Kwaako also shows that the long-term impacts of drought are specific to this one narrow window.

Fenske and Kala (2013), in their study of the link between climate and slave exports in 18th and 19th century Africa, also provide some suggestive evidence of climate being particularly important during the height of the slave trade. They estimate the cross-sectional relationship between contemporary light density (a commonly used measure of economic development at the sub-national level) and historical weather shocks. Their findings provide evidence of the greater importance of temperature shocks during the height of the trans-Atlantic slave trade, which is consistent with the shocks having a large impact on contemporary development through their historical impacts on the supply of slaves.

Nunn and Puga (2012) focus on geography and provide an example of its impact varying over both time and space. They show that for most of the world, terrain ruggedness has a negative contemporaneous impact on economic development. All else equal, rugged terrain makes it more difficult to build buildings, roads, bridges, and other infrastructure;

agriculture and irrigation is also more difficult; and trade is more costly. They further show that within Africa, ruggedness had very different impacts than outside. Within Africa, greater ruggedness is associated with higher incomes, not lower.

The authors argue and provide evidence that this can be explained by an indirect historical impact of geography that was specific to Africa because of its history of the slave trades. During the slave trades, societies were able to use rugged terrain to protect and hide from slave raiders and kidnappers. This allowed individuals, villages, and societies to partially defend against the negative effects of the slave trades documented in Nunn (2008a). Therefore, for the African continent, which was exposed to the slave trade, ruggedness also had a historical indirect positive effect on income. Ruggedness allowed certain areas to evade the slave trade, thereby increasing long-term economic growth.

Nunn and Qian's (2011) study of the introduction of the potato to the Old World during the Columbian Exchange directly exploits the fact that the importance of geography changes over time. Specifically, their analysis relies on the fact that having climate and soil suitable for cultivating potatoes was important only after the potato was introduced from the Americas. Despite not having spatially or temporally extensive data on potato production or consumption, they infer the impacts of the potato by comparing the evolution of populations, city sizes, urbanization rates, and adult heights, before and after the adoption of the potato, in the places suitable for potato cultivation relative to unsuitable locations. Their estimates show that after the introduction of the potato, the places suitable for cultivation witnessed significant population growth, city growth, increased urbanization rates, and increased heights.

Overall, evidence continues to accumulate suggesting that geography can have very different impacts at different points in time and in different locations. The impacts of geography depend crucially on the particular historical context.

7.5. MECHANISMS UNDERLYING HISTORICAL PERSISTENCE

I next turn to the important question of why historical events have persistent impacts. In particular, I discuss the existing evidence for path dependence, culture, institutions, and genetic traits as important channels underlying historical persistence.

7.5.1 Multiple Equilibria and Path Dependence

Although it is far from obvious why historical events have persistent impacts, particularly in the long-run, once one acknowledges the possibility of multiple equilibria, then historical events can have long-term impacts if they move the society from one equilibrium to another. A large number of models show how easily multiple equilibria arise, even in very simple environments. See, for example, Murphy et al. (1993), Acemoglu (1995), Mehlum et al. (2003), and Nunn (2007).

Less formally, many examples of multiple equilibria in daily life have been identified, the most well known being the adoption of the less-efficient QWERTY keyboard over other more efficient configurations like the DVORAK keyboard (David, 1985). The QWERTY keyboard design was developed by Christopher Sholes and patented in 1873. That same year, it was sold to Remington, which used the configuration for their typewriters. The configuration was chosen because it separated the most commonly used keys, which kept the arms of the typewriter from jamming. In other words, the format was chosen because it effectively reduced typing speeds.[14]

A number of studies have undertaken the task of formally testing for the existence of multiple equilibria. A common strategy that has been employed is to examine cases where there has been an extremely large temporary shock to an equilibrium. The studies then test whether the temporary shock causes a permanent movement to a new equilibrium. If so, this is evidence for the existence of multiple equilibria.

Davis and Weinstein (2002, 2008) examine the effect of bombings on 114 Japanese cities during World War II and show that after the bombings, the cities returned to their pre-bombing populations, regained their shares in total manufacturing output, and most surprisingly, also regained their pre-existing industrial composition. Overall, the results point toward the existence of a unique stable equilibrium of production, rather than the existence of multiple equilibria.

Although these results appear to suggest the existence of one unique equilibrium, a second possibility is that the shock was not sufficient to move the society away from the current equilibrium. The US bombings during WWII were dramatic and severe, but they did not alter property rights or the ownership of assets. It is likely that these are the fundamental determinants of where people live and where production occurs.

The findings in Miguel and Roland's (2011) analysis of the long-term effects of the US bombings in Vietnam are consistent with the finding from Davis and Weinstein (2002, 2008). The authors find that the bombings had no long-term effects on populations, poverty, or consumption 25 years later. However, in this case, the authors show that the return can be explained by reconstruction efforts intentionally aimed at rebuilding the hardest hit parts of the country. In other words, policy intentionally helped the country return to its original equilibrium.

An innovative study by Redding et al. (2011) tests for the existence of multiple equilibria in a very different setting. The study examines the location of airport hubs in Germany before and after the division of Germany following World War II. It is shown that after division, the location of West Germany's primary airport hub switched

[14] Liebowitz and Margolis (1990) argue that the efficiency difference between the QWERTY and DVORAK keyboards is lower than argued in David (1985). The authors provide some evidence for this. However, even if the efficiency gap is lower than previously thought as they contend, the QWERTY keyboard still provides a clear example of multiple equilibria and path dependence, which is the central point of David's (1985) original argument.

from Berlin to Frankfurt. After reunification in 1990, the location of the hub did not switch back to Berlin. Redding et al. show that this shift cannot be explained by changes in fundamentals over the time period. Thus, the evidence suggests that the temporary division of Germany resulted in a permanent movement of the location of Germany's largest airport hub.

Bleakely and Lin (2012) examine a very specific and seemingly innocuous geographic characteristic and show that even though it only mattered for a narrow window of time, it had lasting and important impacts on urban development within the United States. The characteristic they examine is the existence of rapids or falls, which occur when a river crosses a fault line. In these locations, river transport required hauling goods and boats over land. This is known as portage. These locations were a focal point for commercial activity and entrepot trade.

The shipment of goods by boat was a dominant form of transportation until the early to mid-19th century, when canals and railways were developed. Combining geographic data with population at the census tract level, the authors show that today, looking either along rivers or along fault lines, populations are concentrated where rivers cross fault lines—i.e. at historical portage sites. The authors then turn to historical populations, examining the relationship between portage and population density from 1790 to 2000. The authors show that after 1850 (and the decline in the use of water transport and portage), the population actually became more (not less) concentrated at portage sites. Their findings are consistent with portage sites serving as a focal point that helped determine the location of early cities (i.e. the equilibrium population distribution) among a large set of possible multiple equilibria.

7.5.2 Domestic Institutions

Even without the existence of multiple equilibria, historical events can still affect economic development in the long-run if they alter deep determinants of long-term economic growth. The determinant that has received the greatest attention in the literature is domestic institutions. This focus is illustrated by the fact that in each of the seminal papers by Acemoglu et al. (2001, 2002), Engerman and Sokoloff (1997, 2002), and La Porta et al. (1997, 1998), the mechanism through which colonial rule affects current development is institutions.

The focus on institutions as a causal mechanism has also continued in subsequent research. An example of this is Acemoglu et al.'s (2005) study of the effect that early Atlantic trade had in Europe. The authors argue that in countries with access to the lucrative Atlantic three-corner trade, economic and political power shifted toward commercial interests. As the merchant class became more powerful, they were able to alter domestic institutions to protect their interests against the interests of the royalty, and these institutional changes in turn had a positive effect on long-term prosperity. Using data on historical urbanization rates and per capita incomes, the study first shows that the rise of

Europe was actually a rise of nations with access to the lucrative Atlantic trade, namely Britain, France, the Netherlands, Portugal, and Spain.

The authors argue that profits alone are not able to explain the divergent growth of Atlantic traders and that the evolution of domestic institutions played an important role in the process.[15] To test this hypothesis, the authors extend the Polity IV data back to 1350 and show that Atlantic trade increased the quality of domestic institutions as measured by an index of the constraints on the executive. They further hypothesize that the process of institutional change could only occur in countries that initially had non-absolutist political institutions. They show that the data are also consistent with this. The increase in economic growth generated by Atlantic trade was higher for countries with better initial domestic institutions, again measured by the constraint on the executive.

Other examples of studies documenting the persistent importance of historical institutions include Dell's (2010) analysis of the impact of the early forced labor institutions in colonial Peru and Bolivia, as well as Banerjee and Iyer's (2005, 2008) studies of the effects of early land tenure institutions in colonial India.

The recent study by Gennaioli and Rainer (2007) also provides evidence of the persistence of early institutions, but within the African context. The authors use ethnographic data to construct a measure of the level of state development in pre-colonial African societies. Their OLS estimates show that there is a positive correlation between pre-colonial political development and the provision of public goods today. More recently, Michalopoulos and Pappaioannou (2013) combine the same ethnographic data used in Gennaioli and Rainer (2007) with satellite data on night-light density. Examining within-country variation, the authors find that the only robust correlate of night-light density is an ethnicity's pre-colonial level of political development. This finding echoes Gennaioli and Rainer's finding of the importance of this variable.

These results can be combined with evidence from Nunn (2008a) showing that the parts of Africa from which more slaves were taken had less developed political systems after the slave trade (and before official colonial rule).[16] Although the evidence for both relationships is based on correlations and therefore one must be cautious when drawing conclusions, the combined evidence from Gennaioli and Rainer (2007), Michalopoulos and Pappaioannou (2013), and Nunn (2008a) is consistent with a chain of causality where the slave trade resulted in a deterioration of domestic political institutions, which in turn had a long-term adverse impact on the provision of public goods. Therefore, the body of evidence provides support for the notion that history can matter through the evolution and persistence of early institutions.

Overall, the literature since Acemoglu et al. (2001) has succeeded at providing additional evidence showing that institutions are an important channel through which history

[15] See Inikori (2002) for the alternative view that the profits that accrued by Western Europe during the three-corner Atlantic trade explain much of its growth during that time.

[16] Also see Whatley (forthcoming) for micro-level evidence for this relationship.

matters. However, much work remains to be done before we have a clear understanding of the effect that historical events have on the formation of early institutions and their persistence and importance for long-term development. For example, in past studies (typically at the macro-level) institutions have been conceptualized and measured as a broad cluster of institutions. The result of this is that, by-and-large, institutions have remained a black box that we do not clearly understand the details of.[17] As empirical research continues to examine specific examples of institutional change and persistence at the micro-level, our understanding of the causes and consequences of specific institutions will naturally improve.

7.5.3 Cultural Norms of Behavior

Another way in which historical events can have long-term impacts is if these past events permanently affect culture or norms of behavior. While in economics the notion of culture often remains vague, other disciplines place much more emphasis on precisely defining culture. For example, evolutionary anthropologists have long recognized that there are clear micro-foundations that explain the existence of a phenomenon like culture (e.g. Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1985). If information acquisition is either imperfect or costly, then selection favors short-cuts to learning. Individuals, rather than using scarce resources to acquire all of the information needed for every decision to be made, will instead develop "rules-of-thumb". These short-cuts then become internalized as individuals come to believe that certain behaviors are the "right" behaviors in certain situations.[18] For a fuller exposition of this definition of culture see Nunn (2012).

The idea that norms of behavior may be a channel through which history can affect long-term economic development is not new. One of the most famous links between history, culture, and development is Max Weber's (1930) hypothesis that the Protestant Reformation was instrumental in facilitating the rise of industrial capitalism in Western Europe. He argues that Protestantism, in contrast to Catholicism, approved the virtues of hard work and the accumulation of wealth, and that these values, referred to as the

[17] An exception is the study by Acemoglu and Johnson (2004), where the authors distinguish "property rights institutions" from "contracting institutions." According to their definitions, property rights institutions protect individuals from theft or expropriation by the government or elites, and contracting institutions enforce private contracts written between individuals. They find that property rights institutions have a positive and significant effect on income, investment, and financial development. On the other hand, contracting institutions appear to have a much more limited impact, only affecting the form of financial intermediation.

[18] Within economics, examples of models of cultural evolution include Verdier (2000, 2001) and Tabellini (2008).

"Protestant work ethic," provided the moral foundation that spurred the transition to a modern market-based industrial economy.[19]

One of the earliest studies empirically examining the possibility that cultural norms may be historically determined was undertaken by a group of social psychologists (Cohen et al. 1996). The authors test whether there is a culture of honor in the US south, where a special importance is placed in defending one's reputation and honor, even if this requires aggression and violence. Their explanation for why this culture exists in the US south and not the north lies in the different histories of settlement in the two areas. The north was settled by groups with a farming background, while the south was settled primarily by the Celts who had been herders since prehistorical times and had never engaged in large-scale agriculture. They argue that in herding cultures, with their low population densities and weak states, protection of one's property was left to the individual and therefore norms of aggressive behavior developed as a means to protect one's herd.

To test the culture of honor hypothesis, Cohen et al. (1996) conducted a series of experiments involving white males from the US north and US south. In the experiments, each individual was bumped by an accomplice and called an "asshole." (The participants did not know this was part of the experiment.) Cohen et al. use a number of methods including direct observation, psychological tests, and saliva tests to compare the effects of this incident on southerners relative to northerners. They find that southerners became more upset, were more likely to feel that their masculinity was threatened, became more physiologically and cognitively primed for aggression as measured by a rise in testosterone and cortisol levels, and were more likely to engage in aggressive behavior, subsequently.

A number of studies provide additional evidence for the historical origins of current cultural differences. For example, Guiso et al. (2008) empirically examine the well-known hypothesis put forth by Putnam et al. (1993) that within Italy, city states that became independent during the 1000–1300 period developed higher levels of social capital, and these higher levels of social capital continue to persist today. The authors bring Putnam et al.'s hypothesis to the data by collecting various city level measures of social capital. They show that looking across 400 Italian cities, there is a positive relationship between their measures of social capital and whether the city was free in 1176.

Nunn and Wantchekon (2011) consider the historical determinants of trust within the African context. The authors examine whether the trans-Atlantic and Indian Ocean slave trades influenced the amount of distrust within society. This is done by combining household survey data with estimates of the number of slaves taken from each ethnic group in Africa. The study finds a negative relationship between an individual's reported trust in others (either neighbors, relatives, local governments, co-ethnics, and those from

[19] A more recent example is Mokyr's (2008) argument that an important determinant of the Industrial Revolution was the development of a social norm he calls "gentlemanly culture" that emphasized honesty, commitment, and cooperation.

other ethnicities) and the number of slaves taken from the individual's ethnic group during the slave trades.

The study attempts to distinguish between the two most plausible channels through which the slave trades could have adversely affected trust. One channel is that they altered the cultural norms of the ethnic groups exposed to the trade, making them inherently less trusting. A second channel is that the slave trades resulted in a long-term deterioration of legal and political institutions, which causes individuals to be less trusting of others today.

The authors undertake a number of tests to distinguish between these two channels. One test examines individuals' trust in the local government and attempts to control for the quality of domestic institutions using the individuals' perceived quality of the local government, extent of corruption, and whether local councillors listen to their concerns, as well as measures of the quality of public goods provision.

Another test controls for a second measure of slave exports: the average number of slaves that were taken from the geographic location that each individual is currently living in. This is different from the first measure, which is the average number of slaves taken from an individual's ethnic group. The second slave export variable is motivated by the fact that when an individual relocates the individual's internal norms move with them, but the external institutional environment is left behind. In other words, institutions, which are external to the individual, are much more geographically fixed, relative to cultural beliefs which are internal to the individual. Therefore, the two variables can be used to distinguish the extent to which the slave trade affects trust through the culture channel versus through the institutions channel. If the slave trade affects trust primarily through internalized norms and cultural beliefs, which are ethnically based and internal to the individuals, then when looking across individuals, what should matter is whether their ancestors were heavily enslaved. If the slave trade affects trust primarily through its deterioration of domestic institutions, which are external to the individual and geographically immobile, then what should matter is whether the external environment the individual is living in was heavily affected by the slave trades.

The results of each of the tests indicate that the slave trades adversely affect trust through both cultural norms and institutions, but that the magnitude of the culture channel is always greater than the institutions channel.

Another cultural consequence of the slave trade that has received attention is the practice of polygamy. Because significantly more men than women were enslaved during the trans-Atlantic slave trade, the ratio of men to women in Africa was significantly affected. It has been hypothesized that this gave rise to the practice of polygamy. Combining Nunn and Wantchekon's (2011) estimates of ethnicity-level slave exports and information from household survey data, Dalton and Leung (2011) and Fenske (2012) find a positive relationship between slave exports and the prevalence of polygamy.

Other examples of evidence for the historical origins of current cultural traits include Alesina et al.'s (2013) study of the relationship between traditional plough use and current gender roles, as well as Durante's (2010) analysis of the link between historical weather variability and current trust. Both have been described earlier in the chapter.

7.5.4 The Interplay Between Culture and Institutions

Generally, studies of the historical importance of culture and studies of the historical importance of institutions are done in isolation of each other. However, there is evidence that there are important complementarities and interdependencies between culture and institutions. I now turn to a discussion of these.

7.5.4.1 Culture Affecting Formal Institutions

Historically, there are many examples of culture impacting the evolution of domestic institutions. Arguably, the most obvious are the European migrant communities established around the globe after the Age of Exploration. At a macro-level, this has been illustrated by Acemoglu et al.'s (2001) colonial origins hypothesis. A more micro-level analysis of the process (at least for the United States) is provided in David Hackett Fischer's (1989) book *Albion's Seed*, where he demonstrates that the institutions and social structures initially established by European migrants arose from the values and beliefs brought with them from the Old World. In other words, the institutions first established were endogenous to the cultural beliefs of the early migrants.

Fisher documents four waves of early migration to North America—the Puritans (1629–1641), the Anglican Cavaliers (1642–1675), the Quakers (1675–1725), and the Scotch-Irish (1717–1775)—and shows how differences in the values of each immigrant wave generated differences in the institutions that were established. The Puritans, believing in the importance of universal education and in a well-functioning society, established universal education, significant taxes, sizable governments, and heavy-handed justice. The Virginia Cavaliers, who believed that inequality was natural and were primarily concerned with maintaining existing forms of hierarchy, implemented limited education, lower taxes, less government spending, and an informal system of justice based on hierarchical violence. The institutions established by the Quakers in the Delaware Valley reflected their belief in the central importance of personal freedoms. All citizens were granted equal legal rights, there was limited government involvement in personal and religious affairs, and taxes were limited. The institutions implemented by the Scotch-Irish were an outgrowth of their belief in freedom from the constraints imposed by government. This resulted in a limited formal justice system (and a reliance on ad hoc vigilante justice), limited political institutions, light taxes, and strong rights to armed resistance from authority.

European mass migration provides one episode that clearly illustrates the endogeneity of institutions to cultures. Other studies also provide similar evidence from other contexts. For example, Zerbe and Anderson (2001) document that the initial property rights institutions established during the 1848 California Gold Rush reflected the values and beliefs that miners brought with them westward. The beliefs—which included individualism, respect for property, and the view that rewards should be commensurate with effort—first developed into collectively practiced norms of behavior (i.e. informal institutions) before being formalized as written laws.

As well, the work by Greif (1994) on the cultural differences between the Maghribi and Genoese medieval traders also illustrates the role of culture in shaping the formation of formal institutions. The Genoese developed institutions that arose from their individualist cultural beliefs, including a formal legal system as well as other formal organizations that helped to facilitate exchange. By contrast, the institutional structures of the Maghribis grew out of their collectivist cultural beliefs. Because the Maghribis continued to rely on informal enforcement mechanisms, organizations remained limited in size and scope.

7.5.4.2 Institutions Affecting Culture

There is also the possibility of feedback effects, with formal institutions affecting the evolution of cultural traits. A number of recent studies have found evidence for this. For example, Guido Tabellini (2010) explains variation across regions of Europe in levels of trust, respect, and confidence in the returns to individual effort. He identifies a strong positive relationship between the prevalence of these cultural traits and measures of the average quality of domestic institutions between 1600 and 1850. The estimates show that European regions that had less well-developed institutions in the past have less trust in others, less respect for others, and believe less in the value of individual effort today.

Evidence for the impact of institutions on culture also comes from a number of studies that use a regression discontinuity strategy, focusing on particularly important historical borders that today lie within the same country. Becker et al. (2011) examine Eastern European villages lying within the same country today, but on either side of the historical Habsburg border. They show that villages that were formerly part of the Habsburg Empire, with its greatly respected and well-functioning bureaucracy, today have greater trust in their local government. Grosjean (2011b) examines location pairs within Eastern Europe and shows that the longer a pair was under the same Empire historically, the more similar the reported social trust of the locations' citizens today. Peisakhin (2010) surveys 1,675 individuals living in 227 villages located within 25 km of the Habsburg-Russian border that divided Ukraine between 1772 and 1918. Relying on information on cultural traits based on answers to survey questions, Peisakhin (2010) documents a wide range of statistically significant cultural differences between the two groups.

7.5.4.3 Coevolution of Culture and Institutions

Tabellini (2008) provides a formal model of the interplay between culture and institutions in an environment in which both are endogenous and co-evolve. In the model, there are two potential cultural traits with one valuing cooperation (or believing cooperation is the right thing to do) more than the other. Vertical transmission of these values is modeled explicitly with parents exerting costly effort to instill values of cooperation. One of the primary innovations of the paper is to also model the endogenous formation of institutions (which enforce cooperation) through majority voting. Tabellini shows that the co-evolution of culture and institutions generates strategic complementarity and multiple equilibria. A culture that values cooperation prefers institutions that strongly enforce cooperation, which in turn increases the returns to cooperation, reinforcing this cultural trait. Conversely, a culture that does not value cooperation prefers institutions that weakly enforce cooperation, which in turn decreases the returns to cooperation, reinforcing a culture that does not value cooperation.

Recent papers that have empirically studied contemporary institutions and culture provide evidence of interactions between culture and institutions. Aghion et al. (2011) examine contemporary labor markets and identify a negative cross-country relationship between the existence of cooperative labor relations and the severity of minimum wage regulation by the state. Similarly, Aghion et al. (2010) identify a negative cross-country relationship between general trust and government regulation.

Both studies then develop models of the interplay between institutions/policies and culture/beliefs. In both, greater government regulation crowds out beneficial behavior of citizens. In Aghion et al. (2011), higher minimum wage regulation reduces the benefits to workers of trying to cooperate with firms. Therefore, more stringent minimum wage regulations crowd out cooperation between firms and workers. In turn, less cooperative firm-worker relationships increase the demand for minimum wage regulation. Thus, this interdependence explains the observed negative relationship between minimum wage and cooperative labor relations.

In Aghion et al. (2010), a low level of civic mindedness in the economy results in a greater need for regulation to protect citizens from the negative externalities imposed by those that are not civic-minded. The high level of regulation in the economy also reinforces the low level of civic mindedness, as it is these individuals that are comfortable paying and demanding bribes. The result is that greater trust is observed in economies with lower levels of government regulation.

What the three studies described here have in common is their analysis of the two-way relationship between culture and institutions. Given this interdependence, both institutions and culture co-evolve, and this can generate multiple stable equilibria with different sets of institutions and cultural norms that are self-enforcing.

7.5.5 Genetics

It is possible that historical events that affect the distribution of individuals in different locations—i.e. through genocide, forced expulsions, or voluntary migrations—could have long-term impacts through a genetic channel. Given that genetic traits tend to be fairly persistent over time, if they have an impact on economic outcomes, then it is theoretically possible that events that impact the genetic distribution of the population may have long-term economic impacts.

A number of recent studies provide some evidence that genetics can impact human behavior. For example, Cesarini et al. (2008) exploit variation in genetic differences between monozygotic and dizygotic twins. The authors compare differences in the actions taken in a standard trust game between monozygotic and dizygotic twins from Sweden and the United States. By assuming that similarity of twin behavior can be decomposed into a common environment, common genes, and other individual-specific variables, and that monozygotic twins share the same environment and same genes, and dizygotic twins share the same environment but have half the alleles of genes, they are able to estimate the extent to which behavior is genetically determined. They find that monozygotic twins consistently exhibit more similar behavior than dizygotic twins, and therefore, based on their assumptions, they conclude that an important part of behavior is genetically determined. The same basic procedure is repeated in Cesarini et al. (2009), but examining behavior in a dictator game and measures of individual risk aversion.

At the macro-level, a number of studies have documented relationships between genetic measures and economic outcomes. Spolaore and Wacziarg (2010) show that greater genetic relatedness has a positive impact on the probability that two populations go to war with one another. Spolaore and Wacziarg (2009) show that, across country-pairs, bilateral genetic distance is positively associated with current income differences. In other words, genetically similar countries are economically more similar.

Ashraf and Galor (2012) provide evidence that genetic diversity within a country is non-monotonically related to per capita income. There is an inverted-U relationship between the two, with income reaching a maximum at an intermediate level of diversity. Too much diversity and too little diversity are both associated with low per capita income.

Cook (2013a) also examines genetic diversity, but unlike Ashraf and Galor (2012), he considers a specific group of genes associated with resistance and susceptibility to disease, namely the major histocompatibility complex. Within humans this is a cluster of 239 genes on the 6th chromosome. Cook (2013a) measures the variation in allele frequency within this system and shows that, across countries, his measure of genetic variation is positively correlated with Olsson and Hibbs's (2005) measure of the number of domesticable animals and Putterman's (2008) measure of the time since the adoption of agriculture. It is also positively correlated with health, measured in the 1960s. Interestingly, by 1990 the health relationship no longer exists.

In a subsequent study, Cook (2013c) considers another channel through which genetics could have long-term impacts on economic development. This is through lactase persistence (i.e. the ability to digest milk after childhood). Cook hypothesizes that, historically, societies with the gene variant that resulted in lactase persistence had access to an additional source of calories, vitamins, and nutrients, which resulted in increased population densities. The author shows that looking across countries, a greater proportion of the population with lactase persistence is associated with greater population density, measured in 1500.

7.6. UNRESOLVED QUESTIONS AND DIRECTIONS FOR FUTURE RESEARCH

7.6.1 Persistence or Reversals?

A number of studies provide evidence of the persistence in economic development over long periods of time. Societies that were more economically, technologically, or institutionally developed in the past are also the most developed today. For example, Comin et al. (2010) document a positive relationship between historical technology levels (as far back as 1000 BC) and current income per capita across different parts of the world. Along similar lines, Bockstette et al. (2002) empirically document a positive relationship between state antiquity and current economic performance today. Societies that were more politically developed in the past, are more economically developed today. At a more micro-level and over a shorter timespan, Huillery's (2011) analysis of French West Africa shows persistence of prosperity between the pre- and post-colonial period.

These findings of persistence stand in contrast to the "reversal of fortunes" documented in Acemoglu et al. (2002): among a sample of former colonies, those locations that were the most prosperous in 1500 are the most underdeveloped today. This reversal has also been confirmed in alternative studies. For example, Nunn (2008a) shows that among African countries, those that were the most developed prior to the slave trades (measured by population density in 1400) had the largest number of slaves taken and have the lowest incomes today.[20]

These two sets of findings appear to stand in contrast with one another, one showing persistence over long periods of time and the other showing a complete reversal. Which is correct? It turns out that both are, and an important part of the difference arises due to differences in the samples being examined. The persistence studies tend to examine all countries globally, while the reversal studies have samples that only include former colonies.

To illustrate this, consider the bivariate relationship between the natural log of per capita income in 1500 and the natural log of real per capita GDP in 2000. This relationship is reported in columns (1)–(3) of Table 7.1. The sample comprises 85 former colonies and 65 non-colonies examined in Comin et al. (2010). Column (1) reports the relationship among former colonies. This is analogous to the regressions estimated by Acemoglu et al. (2002). As shown, consistent with their findings, there is a negative relationship between

[20] Also related is the question of whether Africa has always been behind the rest of the world. While the conventional wisdom is that Africa has generally been the most underdeveloped continent of the world, there is evidence that this view is misplaced (Ehret, forthcoming).

population density in 1500 and per capita income today. There has been a reversal. Column (2) examines this same relationship among the rest of the sample, which are the countries that were never colonized. Here one observes a very different relationship. The two variables are positively correlated. Among this group there is persistence. Column (3) reports the relationship between the full sample, and shows that on average there is persistence. The coefficient is positive and significant at the 10 percent level. This estimate is analogous to the findings of persistence by Comin et al. (2010), Bockstette et al. (2002), and others.

Acemoglu et al. (2002) argue that among former colonies, the reversal occurred because initial prosperity impacted the institutions that were developed by Europeans. Where initial incomes were low, population was sparse, and Europeans settled, establishing protection of property rights and other growth-promoting institutions. Where initial incomes were high, Europeans undertook an extractive strategy. In some cases, they co-opted existing forced labor traditions, and in others, they promoted enslavement and the sale of indigenous populations. As a result, locations that were initially poor in 1500 today are more developed than those that were initially richer.

A similar but alternative explanation for the reversal, and one that has been stressed in the recent paper by Easterly and Levine (2012), is that the less populated places witness an in-migration of people from more prosperous countries, with higher levels of human capital, culture more conducive to economic growth, and/or other vertically transmitted traits. Therefore these locations are richer today. This alternative explanation suggests that the reversal simply reflects migration and the persistence of prosperity at the society level.

This alternative explanation can be examined using an ancestry-based measure of population density in 1500 and per capita GDP in 2000. The ancestry-based initial population measure is constructed using Putterman and Weil's (2010) *World Migration Matrix*. While the geography-based measure used in columns (1)–(3) is the average income (proxied by population density) of people living on the country's land in 1500, the ancestry-based measure is the average income in 1500 (proxied by population density) of the ancestors of those living in the country today.

The estimates from column (4) show that, all else equal, among former colonies, being descended from ancestors with a high prosperity is positively associated with per capita income today. Therefore, it is plausible that colonial migration of individuals from prosperous societies explains the reversal. Interestingly, the persistence of income along lineages (and not locations) is similarly strong among non-colonies (for which there is less migration) as among former colonies (with much greater migration).[21]

A simple way to examine whether the reversal documented in Acemoglu et al. (2002) is explained by migration combined with the persistence of prosperity across generations

[21] Better understanding the specific transmission mechanisms underlying this persistence is the subject of ongoing research and debate. For an excellent summary of this literature see Spolaore and Wacziarg (forthcoming).

Table 7.1 Persistence and reversals

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Dependent variable: ln per capita GDP in 2000						
	Colonies	Non-colonies	Both	Colonies	Non-colonies	Colonies	Non-colonies
ln population density in 1500:							
Geography-based	−0.228***	0.276***	0.115*			−0.316***	0.003
	(0.070)	(0.090)	(0.061)			(0.058)	(0.319)
Ancestry-based				0.475***	0.319***	0.581***	0.316
				(0.098)	(0.100)	(0.086)	(0.355)
Observations	85	65	150	85	65	85	65
R-squared	0.114	0.129	0.023	0.222	0.140	0.430	0.140

Notes: The table reports coefficients from OLS estimates, with standard errors in parentheses. The dependent variable is the natural log of real per capita GDP in 2000. The independent variables are the natural log of a country's population density in 1500, measured as the historical average of the land of the country (geography-based) or the historical average of the ancestors of the population of the country (ancestry-based). The correlation between the two measures is 0.23 for colonies and 0.96 for non-colonies.

* Indicate significance at the 10% level.

** Indicate significance at the 5% level.

*** Indicate significance at the 1% level.

is to examine the coefficient of the geography-based measure of population density in 1500, while controlling for the ancestry-based measure. This is done in column (6). The magnitude of the coefficient for the geography-based measure of 1500 prosperity does not diminish, and actually increases. The ancestry-based 1500 prosperity measure enters with a large positive and significant coefficient. This suggests the coexistence of two channels. One is the migration of populations from more prosperous societies and the other being the reversal of fortune discussed in Acemoglu et al. (2002). This finding of stronger persistence by ancestry than by location is not new and is an important point made in Putterman and Weil (2010), Comin et al. (2010), and Chanda et al. (2013).

Column (7) examines the same correlations as column (6), but among the sample of non-colonies. Because there is little migration among this group, the two population density measures are highly correlated (the correlation coefficient is 0.96). Due to this multicollinearity, both variables are insignificant. However, the estimated coefficient for the ancestry-based variable provides evidence of persistence across generations that is similar in magnitude but smaller than the estimate for the colonies sample. As expected, the coefficient for the geography-based variable shows no evidence of a reversal-of-fortunes mechanism among non-colonies.

Overall, the correlations reported in Table 7.1 are suggestive of the following facts. First, within former colonies, there has been a reversal of fortunes (looking at geographic locations as the unit of observation). Second, no such reversal exists among non-colonies. Third, there is no reversal once one uses societies (and their descendants) as the unit of observation. Instead one observes extreme persistence, both among former colonies and non-colonies, a fact that has been empirically noted by Putterman and Weil (2010) and discussed in Spolaore and Wacziarg (forthcoming). Fourth, the Acemoglu et al. (2002) reversal exists even after accounting for the migration of populations from more prosperous to less prosperous regions during the colonial period. This does not appear to fully explain the reversal.

Therefore, the existence of reversals and persistence in the data seem to be reconcilable. However, the most recent research along these lines shows a reversal that is not explained by the logic above. Olsson and Paik (2012) document a reversal within Europe from the Neolithic until now. They show that the parts of Europe that adopted agriculture earlier (and were arguably more economically developed during the Neolithic period) are less developed today. Although the authors provide an explanation, the exact reason for this reversal is far from clear. They also find evidence of a reversal within sub-Saharan Africa and East Asia. The reason behind the reversals in these regions is also unclear. Most interestingly, they show that if one looks at a global sample, there is persistence: the parts of the world that adopted agriculture earlier are more developed today. In other words, looking within-regions there are reversals, but looking across regions (and across countries generally) there is persistence.

7.6.2 When Doesn't History Persist?

To date, the primary focus of the literature has been in empirically documenting the persistence of historical shocks, typically arising due to lasting impacts through either domestic institutions or cultural traits. Little or no attention has been placed on examining when historical events *do not* have lasting impacts. This emphasis is logical given the need to first establish that history can matter, which has led to a natural focus on events that have had persistent impacts.

However, there are a few studies that provide some preliminary evidence for when history persists and when it does not. For example, Voigtlaender and Voth (2012) document the persistence of anti-Semitic values and beliefs in Germany between the 14th and 20th centuries. Their analysis examines variation across German villages and documents a remarkable relationship between the prevalence of pogroms during the Black Death (1348–1350) and a number of measures of anti-Semitic sentiment in the early 20th century. The authors then turn to an analysis of the environments in which persistence was more or less strong. One of their most interesting findings is that the persistence of this cultural trait is much weaker among Hanseatic cities, which were self-governed German cities heavily involved in lucrative long-distance trade. This finding may be due to higher rates of migration or to more dynamism arising from greater economic opportunity and growth. Voigtlaender and Voth (2012) also find that (consistent with both mechanisms) there is less persistence among cities with faster population growth, and (consistent with the second mechanism) there is less persistence among cities that were more industrialized in 1933.

Grosjean's (2011a) study of Nisbett and Cohen's (1996) "culture of honor" hypothesis shows a persistent impact of the Scotch-Irish culture of honor, but only within the Southern states of the US. The obvious explanation is that a cultural heuristic of aggression was relatively beneficial in the south, which was more lawless and with less well-developed property rights institutions. However, in the north, with a more established rule of law and better developed property rights protection, norms of aggression and violence were less beneficial, and therefore did not persist. In other words, external characteristics—in this case domestic institutions—by affecting the relative costs and benefits of different cultural norms, influence their persistence.

Another environment in which this can be seen is in Africa in the context of the slave trade. A natural hypothesis is that the detrimental impacts of the slave trades on trust will be more persistent in countries with a poorly functioning legal system. It is in these environments, where individuals are not legally constrained to act in a trustworthy manner, that norms of mistrust, initially developed by the slave trade, may continue to be relatively beneficial and to persist.

This can be tested directly by re-estimating Equation (7.1) from Nunn and Wantchekon (2011), but allowing for the impact of past slave exports on trust today to depend on the quality of country-level domestic institutions, measured at the time of the survey (2005) using the Governance Matters "rule of law" variable. The original index ranges from −2.5 to +2.5, but I normalize the variable to lie between zero and one.[22] The augmented equation is:

$$\begin{aligned}\text{trust}_{i,e,d,c} = \alpha_c + \beta_1 \text{ slave exports}_e + \beta_2 \text{ slave exports}_e \times \text{rule of law}_c \\ + \mathbf{X}'_{i,e,d,c}\boldsymbol{\Gamma} + \mathbf{X}'_{d,c}\boldsymbol{\Omega} + \mathbf{X}'_e\boldsymbol{\Phi} + \varepsilon_{i,e,d,c},\end{aligned} \tag{7.1}$$

where i indexes individuals, e ethnic groups, d districts, and c countries; $\text{trust}_{i,e,d,c}$ denotes one of five individual-level measures of trust that range from 0 to 3; slave exports_e is a measure of the number of slaves taken from ethnic group e during the Indian Ocean and trans-Atlantic slave trades[23]; rule of law_c is the 0-to-1 measure of a country's rule of law in 2005; α_c denotes country fixed effects; and $\mathbf{X}'_{i,e,d,c}$, $\mathbf{X}'_d$, and $\mathbf{X}'_e$ denote vectors of individual-, district-, and ethnicity-level control variables. See Nunn and Wantchekon (2011) for a fuller description of the variables in Equation (7.1).

Estimates of Equation (7.1) are reported in Table 7.2. The table reports estimates of β_1 and β_2. The bottom panel reports estimates of the impact of the slave trade on trust for the country with the lowest measure of rule of law (0.17) and for the country with the highest rule of law measure (0.63).[24] As shown, the estimated coefficient for the interaction term β_2 is positive in all specifications (although the precision of the estimate varies). This indicates a weaker negative impact of the slave trades on trust in countries with better domestic institutions. Further, for all trust measures, the estimated impact of the slave trades on trust is positive and significant for the lowest rule of law country but not statistically different from zero for the highest rule of law country. This is consistent with the adverse impacts of the slave trade being less persistent in countries with a better rule of law. In these countries, well-functioning institutions enforce trustworthy behavior of its citizens and therefore there is less persistence of the mistrust engendered by the slave trades.

An important shortcoming of this exercise arises due to the endogeneity of the country-level rule of law measure. In particular, it is likely endogenous to the slave trade. Ideally, estimates of this nature would rely on exogenous variation in the variable used to test for heterogeneity. However, an important point to bear in mind is that the estimates reported in Table 7.2 and Nunn and Wantchekon (2011) are estimated using within-country variation only. Any impacts that the slave trade had on country-level characteristics are controlled for directly in the regression because of the presence of country fixed effects. In other words, although the rule of law measure is an endogenous

[22] This is done by adding 2.5 to the measure and dividing by 5.

[23] The measure is the natural log of one plus total slave exports normalized by land area.

[24] Zimbabwe is the country with the lowest rule of law measure in the Afrobarometer sample, and Botswana is the country with the highest.

Table 7.2 Testing for heterogenous impacts of the slave trade on trust in Nunn and Wantchekon (2011)

	Trust of relatives	Trust of neighbors	Trust of local council	Intra-group trust	Inter-group trust
	(1)	(2)	(3)	(4)	(5)
ln (1 + exports/area)	−0.172	−0.341***	−0.170***	−0.461***	−0.344***
	(0.141)	(0.115)	(0.064)	(0.102)	(0.082)
ln (1 + exports/area) × rule of law	0.111	0.512*	0.169	0.891***	0.695***
Index 2005	(0.360)	(0.302)	(0.173)	(0.263)	(0.208)
Individual controls	Yes	Yes	Yes	Yes	Yes
District controls	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes
Number of observations	20,062	20,027	19,733	19,952	19,765
Number of ethnicity clusters	185	185	185	185	185
Number of district clusters	1257	1257	1283	1257	1255
Estimated impact for Afrobarometer	−0.153*	−0.252***	−0.141***	−0.305***	−0.223***
country with lowest rule of law	(0.082)	(0.066)	(0.036)	(0.059)	(0.049)
Estimated impact for Afrobarometer	−0.102	−0.018	−0.064	0.102	0.095
country with highest rule of law	(0.098)	(0.087)	(0.054)	(0.073)	(0.058)

Notes: The table reports OLS estimates. The unit of observation is an individual. ln (1 + exports/area) is the number of slaves exported normalized by land area, measured at the ethnicity level. Rule of Law Index 2005 is the Governance Matters VI rule of law measure for 2005, normalized to lie between zero and one. Standard errors are adjusted for two-way clustering at the ethnicity and district levels. The individual controls are for age, age squared, a gender indicator variable, 5 living conditions fixed effects, 10 education fixed effects, 18 religion fixed effects, 25 occupation fixed effects, and an indicator for whether the respondent lives in an urban location. The district controls include ethnic fractionalization in the district and the share of the district's population that is the same ethnicity as the respondent. See Nunn and Wantchekon (2011) for further details.

* Indicate significance at the 10% level.

** Indicate significance at the 5% level.

*** Indicate significance at the 1% level.

variable, its direct (linear) impacts on trust are captured by the country fixed effects in the regression.

Looking at the differences in the estimates of β_1 and β_2 across the five trust measures, it is clear that the heterogeneous impacts of the slave trades are weaker for trust of relatives and trust of the local government. This is true whether one considers the magnitude and significance of β_2 or of the high and low estimates reported in the bottom panel of the table. Interestingly, disputes between relatives and disputes between citizens and the local government are less likely to be resolved through the legal system than disputes between neighbors, co-ethnics, or citizens from different ethnic groups. Given this, we would expect that rule of law would be less successful in enforcing good behavior in these situations, and as a result would be a less important determinant of the persistence of distrust. The estimates reported in Table 7.2 are consistent with this.

7.7. CONCLUSIONS: LOOKING BACK WHILE MOVING FORWARD

This chapter has provided a broad overview of research examining comparative historical economic development. Studies have examined a wide array of historic events, including the Neolithic Revolution, colonial rule, Africa's slave trades, the Industrial Revolution, the Protestant Reformation, the French Revolution, and the Columbian Exchange.

Although the studies reviewed in this chapter have done much to identify important pieces of the larger historical puzzle, many of the pieces are yet to be uncovered. In addition, the more difficult task is understanding exactly how all of the pieces fit together. This is a step that has not been taken by the vast majority of the studies in the literature summarized here. Nearly all examine a particular event in isolation from other events, except possibly to account for other events as covariates in the empirical analysis. However, once one begins thinking of the realities of history, it is soon apparent that historical events impact other historical events in important and sometimes subtle or complicated ways. Further, there are often complex interactions between events, suggesting that the linear specifications typically assumed in studies may be inaccurate.

There are many examples of these interdependencies. For example, Europe's ability to colonize and rule the African continent depended critically on the discovery of the chincona tree in the Andes and its mass production in Asia by the British. This is because quinine, the first effective protection against malaria, is derived from the bark of the tree. Similarly, European knowledge of how to effectively process wild rubber obtained from Native Americans had important consequences for the millions of Africans that were tortured and killed in King Leopold's Congo.

Another example is the interdependence between the printing press and both the Protestant Reformation (Dittmar, 2011; Rubin, 2011) and the Atlantic trade (Dittmar, 2011). We have also seen that Catholic conflict with the Ottoman Empire helped enable the spread of the Protestant religion across Europe (Iyigun, 2008).

We have seen that the presence of the tsetse fly in Africa resulted in less intensive agriculture that did not use animals or the plough (Alsan, 2012). Because the plough was not adopted, women participated actively in agriculture, which generated norms of equality, which continue to persist today (Alesina et al. 2013). This an important explanation for the high levels of female labor force participation that is observed in Africa today.

We have seen that Africa's slave trades resulted in underdeveloped pre-colonial states (Nunn, 2008a), which in turn are associated with less post-colonial public goods provision and lower incomes (Gennaioli and Rainer, 2007; Michalopoulos and Pappaioannou, 2013).

Moving forward, the second major task for the literature to tackle is to better understand channels of causality. In the past decade, we have made significant progress empirically testing whether historical events have lasting impacts. The bulk of this survey is devoted to reviewing this evidence, which overwhelmingly shows that history does matter. What is less clear is exactly why it matters. I have reviewed here the leading candidates: multiple equilibria, cultural norms of behavior, and domestic institutions. The extent to which these mechanisms matter, and in which circumstances, is yet to be fully understood. Further, as discussed, there are also potentially important complementarities between the channels. For example, beliefs and values tend to become codified in formal institutions, which in turn feedback, affecting the evolution of these values. Complementarities between cultural traits and formal institutions are likely an important part of many instances of long-term persistence.

Overall, while much progress has been made to this point, the primary accomplishment has been in establishing the importance of studying the past for understanding current growth and development. The economic literature is increasingly coming to understand that where we are (and therefore how we best move forward) has a lot to do with how we got here.

ACKNOWLEDGMENT

The chapter has benefitted from helpful discussions with Emmanuel Akyeampong, Robert Bates, Samuel Bowles, Claudia Goldin, Joseph Henrich, and James Robinson. I thank Eva Ng for research assistance.

REFERENCES

Acemoglu, Daron, 1995. Reward structure and the allocation of talent. European Economic Review 39, 17–33.

Acemoglu, Daron, Johnson, Simon, 2004. Unbundling Institutions. Journal of Political Economy 113, 949–995.

Acemoglu, Daron, Johnson, Simon, Robinson, James A., 2001. The colonial origins of comparative development: an empirical investigation. American Economic Review 91, 1369–1401.

Acemoglu, Daron, Johnson, Simon, Robinson, James A., 2002. Reversal of fortune: geography and institutions in the making of the modern world income distribution. Quarterly Journal of Economics 117, 1231–1294.

Acemoglu, Daron, Johnson, Simon, Robinson, James A., 2005. The rise of Europe: atlantic trade, institutional change and economic growth. American Economic Review 95, 546–579.

Acemoglu, Daron, Bautista, María Angélica, Querubin, Pablo, Robinson, James A., 2008. Economic and political inequality in development: the case of Cundinamarca, Colombia. In: Helpman, Elhanan (Ed.), Institutions and Economic Performance, Harvard University Press, Cambridge, MA, pp. 181–245.

Acemoglu, Daron, Cantoni, Davide, Johnson, Simon, Robinson, James A., 2011a. The consequences of radical reform: the french revolution. American Economic Review 101 (7), 3286–3307.

Acemoglu, Daron, Hassan, Tarek A., Robinson, James A., 2011b. Social structure and development: a legacy of the Holocaust in Russia. Quarterly Journal of Economics 126 (2), 895–946.

Acemoglu, Daron, Johnson, Simon, Robinson, James A., 2012. The colonial origins of comparative development: an empirical investigation: reply. American Economic Review 102 (6), 3077–3110.

Ager, Philipp, Ciccone, Antonio, 2012. Rainfall Risk and Religious Membership in the Late Nineteenth-Century US. Universitat Pompeu Fabra, Mimeo.

Aghion, Philippe, Algan, Yann, Cahuc, Pierre, Shleifer, Andrei, 2010. Regulation and Distrust. Quarterly Journal of Economics 125 (3), 1015–1049.

Aghion, Philippe, Algan, Yann, Cahuc, Pierre, February 2011. Civil society and the state: the interplay between cooperation and minimum wage regulation. Journal of the European Economic Association 9 (1), 3–42.

Aghion, Philippe, Persson, Torsten, Rouzet, Dorothee, 2012. Education and Military Rivalry. Harvard University, Mimeo.

Akçomak, I. Semih, Webbink, Dinand, Weel, Bas ter, 2012. Why did the Netherlands develop so early? The Legacy of the Brethren of the Common Life, Mimeo.

Albouy, David Y., 2012. The colonial origins of comparative development: an empirical investigation: comment. American Economic Review 102 (6), 3059–3076.

Alesina, Alberto, Giuliano, Paola, Nunn, Nathan, 2013. On the origins of gender roles: women and the plough. Quarterly Journal of Economics 128 (2), 155–194.

Alsan, Marcella, 2012. The Effect of the TseTse Fly on African Development. Harvard University, Mimeo.

Andersen, Thomas Barnebeck, Bentzen, Jeanet, Dalgaard, Carl-Johan, 2011. Religious Orders and Growth Through Cultural Change in Pre-Industrial England. University of Copenhagen, Mimeo.

Andersen, Thomas Barnebeck, Jensen, Peter Sandholt, Skovsgaard, Christian Stejner, 2013. The Heavy Plough and the European Agricultural Revolution in the Middle Ages: Evidence from a Historical Experiment. University of Southern Denmark, Mimeo.

Ashraf, Quamrul, Galor, Oded, 2012. The out of Africa hypothesis. Human genetic diversity, and comparative economic development. American Economic Review 103 (1), 1–46.

Ashraf, Quamrul, Michalopoulos, Stelios, 2011. The Climatic Origins of the Neolithic Revolution: Theory and Evidence. Brown University, Mimeo.

Austin, Gareth, November 2008. The "Reversal of Fortune" thesis and the compression of history: perspectives from African and comparative economic history. Journal of International Development 20 (8), 996–1027.

Bai, Ying, Kung, James Kai-sing, 2011a. Climate shocks and sino-nomadic conflict. Review of Economics and Statistics 93 (3), 970–981.

Bai, Ying, Kung, James Kai-sing, 2011b. Diffusing Knowledge while Spreading God's Message: Protestantism and Economic Prosperity in China, 1840–1920. Hong Kong University of Science and Technology, Mimeo.

Banerjee, Abhijit, Iyer, Lakshmi, 2005. History, institutions and economic performance: the legacy of colonial land tenure systems in India. American Economic Review 95 (4), 1190–1213.

Banerjee, Abhijit, Iyer, Lakshmi, 2008. Colonial Land Tenure, Electoral Competition and Public Goods in India. Harvard Business School Working Paper 08–062.

Banerjee, Abhijit, Iyer, Lakshmi, Somanathan, Rohini, 2005. History, social divisions and public goods in rural India. Journal of the European Economic Association Papers and Proceedings 3 (2–3), 639–647.

Baten, Joerg, van Zanden, Jan Luiten, 2008. Book production and the onset of modern economic growth. Journal of Economic Growth 13, 217–235.

Bates, Robert H., forthcoming. The imperial peace in colonial Africa and Africa's underdevelopment. In: Akyeampong, Emmanuel, Bates, Robert H., Nunn, Nathan, Robinson, James A. (Eds.), Africa's Development in Historical Perspective, Cambridge University Press, Cambridge.

Becker, Sascha O., Woessmann, Ludger, 2008. Luther and the girls: religious denomination and the female education gap in nineteenth-century Prussia. Scandinavian Journal of Economics 110 (4), 777–805.

Becker, Sascha O., Woessmann, Ludger, 2009. Was weber wrong? a human capital theory of protestant economic history. Quarterly Journal of Economics 124 (2), 531–596.

Becker, Sascha O., Boeckh, Katrin, Hainz, Christa, Woessmann, Ludger, 2011. The Empire is Dead, Long Live the Empire! Long-Run Persistence of Trust and Corruption in the Bureaucracy. Warwick University, Mimeo.

Bentzen, Jeanet Sinding, Kaarsen, Nicolai, Wingender, Asger Moll, 2012. Irrigation and Autocracy. University of Copenhagen, Mimeo.

Bleakely, Hoyt, Lin, Jeffrey, 2012. Portage and path dependence. Quarterly Journal of Economics 127, 587–644.

Bockstette, Valeri, Chanda, Areendam, Putterman, Louis, 2002. States and markets: the advantage of an early start. Journal of Economic Growth 7, 347–369.

Boserup, Ester, 1970. Woman's Role in Economic Development. Allen and Unwin, London.

Botticini, Maristella, Eckstein, Zvi, 2005. Jewish occupational selection: education, restrictions, or minorities? Journal of Economic History 65 (4), 922–948.

Botticini, Maristella, Eckstein, Zvi, 2007. From farmers to merchants, voluntary conversions and diaspora: a human capital interpretation of Jewish history. Journal of the European Economic Association 5 (5), 885–926.

Boyd, Robert, Richerson, Peter J., 1985. Culture and the Evolutionary Process. University of Chicago Press, London.

Bruhn, Miriam, Gallego, Francisco A., 2012. Good, bad, and ugly colonial activities: do they matter for economic development? Review of Economics and Statistics 94 (2), 433–461.

Cantoni, Davide, May 2012. Adopting a new religion: the case of Protestantism in 16th century Germany. Economic Journal 122 (560), 502–531.

Cavalli-Sforza, L.L., Feldman, M.W., 1981. Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press, Princeton.

Cesarini, David, Dawes, Christopher T., Fowler, James H., Johannesson, Magnus, Lichtenstein, Paul, Wallace, Bjorn, March 2008. Heritability of cooperative behavior in the trust game. Proceedings of the National Academy of Sciences 105 (10), 3721–3726.

Cesarini, David, Dawes, Christopher T., Fowler, James H., Johannesson, Magnus, Lichtenstein, Paul, Wallace, Bjorn, May 2009. Genetic variation in preferences for giving and risk taking. Quarterly Journal of Economics 124 (2), 809–842.

Chanda, Areendam, Cook, C. Justin, Putterman, Louis, 2013. Persistence of Fortune: Accounting for Population Movements, There was no Post-Columbian Reversal. Brown University, Mimeo.

Chaney, Eric, 2008. Ethnic Cleansing and the Long-Term Persistence of Extractive Institutions: Evidence from the Expulsion of the Moriscos. Harvard University, Mimeo.

Chaney, Eric, Hornbeck, Richard, 2013. Economic Growth in the Malthusian Era: Evidence from the 1609 Spanish Expulsion of the Moriscos. Harvard University, Mimeo.

Chaney, Eric, 2013. Revolt on the Nile: economic shocks, religion and political power. Econometrica.

Chen, Shuo, Kung, James Kai-sing, 2012. The Malthusian Quagmire: Maize and Population Growth in China, 1550–1910. Hong Kong University of Science and Technology, Mimeo.

Clark, Gregory, 2005. The condition of the working class in England, 1209–2004. Journal of Political Economy 113 (6), 707–736.

Cohen, Dov, Nisbett, Richard E., Bowdle, Brian F., Schwarz, Norbert, 1996. Insult, agression, and the southern culture of honor: an "Experimental Ethnography". Journal of Personality and Social Psychology 70 (5), 945–960.

Comin, Diego, Easterly, William, Gong, Erick, 2010. Was the wealth of nations determined in 1000 B.C.? American Economic Journal: Macroeconomics 2 (3), 65–97.

Cook, C. Justin, 2013a. The Long Run Health Effects of the Neolithic Revolution: The Natural Selection of Infectious Disease Resistance. Yale University, Mimeo.

Cook, C. Justin, 2013b. Potatoes, Milk, and the Old World Population Boom. Yale University, Mimeo.

Cook, C. Justin, 2013c. The Role of Lactase Persistence in Precolonial Development. Yale University, Mimeo.

Dalton, J., Leung, T., 2011. Why is Polygamy more Prevalent in Western Africa? An African Slave Trade Perspective. Wake Forest University, Mimeo.

David, Paul A., 1985. Clio and the economics of QWERTY. American Economic Review Papers and Proceedings 75 (2), 332–337.

Davis, Donald R., Weinstein, David E., 2002. Bones, bombs, and breakpoints: the geography of economic activity. American Economic Review 92 (5), 1269–1289.

Davis, Donald R., Weinstein, David E., 2008. A search for multiple equilibria in urban industrial structure. Journal of Regional Science 48 (1), 29–65.

Deconinck, Koen, Verpoorten, Marijke, 2013. Narrow and scientific replication of "The Slave Trade and the Origins of Mistrust in Africa". Journal of Applied Econometrics.

Dell, Melissa, 2010. The persistent effects of Peru's mining mita. Econometrica 78 (6), 1863–1903.

Dell, Melissa, 2012. Insurgency and Long-Run Development: Lessons from the Mexican Revolution. Harvard University, Mimeo.

Diamond, Jared, 1997. Guns, Germs, and Steel. W.W. Norton & Company, New York.

Diamond, Jared, 2002. Evolution, consequences and future of plant and animal domestication. Nature 418, 700–707.

Dippel, Christian, 2011. Coexistence, Forced Coexistence and Economic Development: Evidence from Native American Reservations. University of California Los Angeles, Mimeo.

Dittmar, Jeremiah E., 2011. Information technology and economic change: the impact of the printing press. Quarterly Journal of Economics 126 (3), 1133–1172.

Dittmar, Jeremiah E., 2012. The Welfare Impact of a New Good: The Printed Book. American University, Mimeo.

Durante, Ruben, 2010. Risk, Cooperation and the Economic Origins of Social Trust: An Empirical Investigation. Science Po, Mimeo.

Easterly, William, Levine, Ross, 2003. Tropics, germs and crops: how endowments influence economic development. Journal of Monetary Economics 50, 3–39.

Easterly, William, Levine, Ross, 2012. The European Origins of Economic Development. NBER Working Paper 18162.

Ehret, Christopher, forthcoming. Africa in world history before c. 1440. In: Akyeampong, Emmanuel, Bates, Robert H., Nunn, Nathan, Robinson, James A. (Eds.), Africa's Development in Historical Perspective, Cambridge University Press, Cambridge, (Chapter 2).

Engerman, Stanley L., Sokoloff, Kenneth L., 1997. Factor endowments, institutions, and differential paths of growth among new world economies: a view from economic historians of the United States. In: Harber, Stephen (Ed.), How Latin America Fell Behind, Stanford University Press, Stanford, pp. 260–304.

Engerman, Stanley L., Sokoloff, Kenneth L., 2002. Factor Endowments, Inequality, and Paths of Development Among New World Economies, Working Paper 9259, National Bureau of Economic Research.

Engerman, Stanley L., Sokoloff, Kenneth L., 2005. The evolution of suffrage institutions in the Americas. Journal of Economic History 65, 891–921.

Fairbank, John King, 1953. Trade and Diplomacy on the China Coast, 1842–1854. Harvard University Press, Cambridge.

Feir, Donna, 2013. The Long Term Effects of Indian Residential Schools on Human and Cultural Capital. University of British Columbia, Mimeo.

Fenske, James, 2011. Ecology, Trade and States in Pre-Colonial Africa. Oxford University, Mimeo.

Fenske, James, 2012. African Polygamy: Past and Present. Oxford University, Mimeo.

Fenske, James, Kala, Namrata, 2013. Climate, Ecosystem Resilience and the Slave Trade. CEPR Discussion Paper 9449.

Feyrer, James D., Sacerdote, Bruce, 2009. Colonialism and modern income: islands as natural experiments. Review of Economics and Statistics 91 (2), 245–262.

Fischer, David Hackett, 1989. Albion's Seed: Four British Folkways in America. Oxford University Press, New York.

Gennaioli, Nicola, Rainer, Ilia, 2007. The modern impact of precolonial centralization in Africa. Journal of Economic Growth 12 (3), 185–234.

Gimbutas, Marija, 2007. The Goddesses and Gods of Old Europe: Myths and Cult Images. University of California Press, Berkeley.

Glaeser, Edward L., Porta, Rafael La, Lopez-De-Silanes, Florencio, Shleifer, Andrei, 2004. Do institutions cause growth? Journal of Economic Growth 9, 271–303.

Greif, Avner, 1993. Contract enforceability and economic institutions in early trade: the Maghribi Traders' coalition. American Economic Review 83 (3), 525–548.

Greif, Avner, 1994. Cultural beliefs and the organization of society: a historical and theoretical reflection on collectivist and individualist societies. Journal of Political Economy 102 (5), 912–950.

Grennes, Thomas, 2007. The columbian exchange and the reversal of fortune. Cato Journal 27 (1), 91–107.

Grosjean, Pauline, 2011a. A History of Violence: The Culture of Honor as a Determinant of Homicide in the US South. University of New South Wales, Mimeo.

Grosjean, Pauline, 2011b. The weight of history on European cultural integration: a gravity approach. American Economic Review Papers and Proceedings 101 (3), 504–508.

Guiso, Luigi, Sapienza, Paola, Zingales, Luigi, 2008. Long-Term Persistence, Mimeo.

Haber, Stephen, Menaldo, Victor, 2010. Rainfall and Democracy. Stanford University, Mimeo.

Hansen, Casper Worm, Jensen, Peter Sandholt, Skovsgaard, Christian, 2012. Gender Roles and Agricultural History: The Neolithic Inheritance. Aarhus University, Mimeo.

Hersch, Jonathan, Voth, Hans-Joachim, 2009. Sweet Diversity: Colonial Goods and the Rise of European Living Standards after 1492. Universitat Pompeu Fabra, Mimeo.

Heywood, Linda, 2009. Slavery and its transformation in the kingdom of Kongo: 1491–1800. Journal of African History 21, 1–22.

Hodder, Ian, January 2005. Women and men at Çatalhöyük. Scientific American 15, 34–41.

Huillery, Elise, 2009. History matters: the long-term impact of colonial public investments in French West Africa. American Economic Journal: Applied Economics 1 (2), 176–215.

Huillery, Elise, 2011. The impact of european settlement within French West Africa: did pre-colonial prosperous areas fall behind? Journal of African Economies 20 (2), 263–311.

Inikori, Joseph E., 2000. Africa and the trans-atlantic slave trade. In: Falola, Toyin (Ed.), Africa Volume I: African History Before 1885, Carolina Academic Press, North Carolina, pp. 389–412.

Inikori, Joseph E., 2002. Africans and the Industrial Revolution in England: A Study in International Trade and Economic Development. Cambridge University Press, Cambridge.

Inikori, Joseph E., 2003. The struggle against the trans-atlantic slave trade. In: Diouf, A. (Ed.), Fighting the Slave Trade: West African Strategies, Ohio University Press, Athens, Ohio, pp. 170–198.

Iyer, Lakshmi, 2010. Direct versus indirect colonial rule in India: long-term consequences. Review of Economics and Statistics 92 (4), 693–713.

Iyigun, Murat, 2008. Luther and Suleyman. Quarterly Journal of Economics 123 (4), 1465–1494.

Jancec, Matija, 2012. Do Less Stable Borders Lead to Lower Levels of Political Trust? Empirical Evidence from Eastern Europe. University of Maryland at College Park, Mimeo.

Jha, Saumitra, 2008. Trade, Institutions and Religious Tolerance: Evidence from India. Stanford University, Mimeo.

Jia, Ruixue, forthcoming. The legacies of forced freedom: China's treaty ports. Review of Economics and Statistics.

Jia, Ruixue, forthcoming. Weather shocks, sweet potatoes and peasant revolts in historical China. Economic Journal.

Lal, R., Reicosky, D.C., Hanson, J.D., 2007. Evolution of the plow over 10,000 Years and the rationale for no-till farming. Soil and Tillage Research 93 (1), 1–12.

La Porta, Rafael, Lopez-de-Silanes, Florencio, Shleifer, Andrei, Vishny, Robert, 1997. Legal determinants of external finance. Journal of Finance 52, 1131–1150.

La Porta, Rafael, Lopez-de-Silanes, Florencio, Shleifer, Andrei, Vishny, Robert, 1998. Law and finance. Journal of Political Economy 106, 1113–1155.

Liebowitz, S.J., Margolis, Stephen E., 1990. The fable of the keys. Journal of Law and Economics 33 (1), 1–25.

Mamdani, Mahmood, 2001. When victims become killers: colonialism, nativism, and genocide in Rwanda. Princeton University Press, Princeton, N.J.

Mann, Charles C., 2011. 1493: Uncovering the New World Columbus Created. Alfred A. Knopf, New York.

Mehlum, Halvor, Moene, Karl, Torvik, Ragnar, 2003. Predator or prey? parasitic enterprises in economic development. European Economic Review 47, 275–294.

Michalopoulos, Stelios, 2012. The origins of ethnolinguistic diversity. American Economic Review 102 (4), 1508–1539.

Michalopoulos, Stelios, Pappaioannou, Elias, 2011. The Long-Run Effects of the Scramble for Africa, NBER Working Paper 17620.

Michalopoulos, Stelios, Pappaioannou, Elias, 2013. Pre-colonial ethnic institutions and contemporary African development. Econometrica 81 (1), 113–152.

Miguel, Edward, Roland, Gérard, 2011. The long run impact of bombing Vietnam. Journal of Development Economics 96 (1), 1–15.

Mokyr, Joel, 2008. The institutional origins of the industrial revolution. In: Helpman, Elhanan (Ed.), Institutions and Economic Performance, Harvard University Press, Cambridge, MA, 64–119.

Murdock, George Peter, 1959. Africa: Its Peoples and Their Cultural History. McGraw-Hill Book Company, New York.

Murphy, Kevin M., Shleifer, Andrei, Vishny, Robert W., 1993. Why is rent-seeking so costly to growth. American Economic Review Papers and Proceedings 83 (2), 409–414.

Naritomi, Joana, Soares, Rodrigo R., Assuncao, Juliano J., 2012. Institutional development and colonial heritage within Brazil. Journal of Economic History 72 (2), 393–422.

Nisbett, Richard E., Cohen, Dov, 1996. Culture of Honor: The Psychology of Violence in the South. Westview Press, Boulder.

Nunn, Nathan, 2007. Historical legacies: a model linking Africa's past to its current underdevelopment. Journal of Development Economics 83 (1), 157–175.

Nunn, Nathan, 2008a. The long-term effects of Africa's slave trades. Quarterly Journal of Economics 123 (1), 139–176.

Nunn, Nathan, 2008b. Slavery, inequality, and economic development in the Americas: an examination of the Engerman-Sokoloff hypothesis. In: Helpman, Elhanan (Ed.), Institutions and Economic Performance. Harvard University Press, Cambridge, MA, 148–180.

Nunn, Nathan, 2009. The importance of history for economic development. Annual Review of Economics 1 (1), 65–92.

Nunn, Nathan, 2012. Culture and the historical process. Economic History of Developing Regions 27, 108–126.

Nunn, Nathan, Puga, Diego, February 2012. Ruggedness: the blessing of bad geography in Africa. Review of Economics and Statistics 94 (1), 20–36.

Nunn, Nathan, Qian, Nancy, May 2010. The columbian exchange: a history of disease, food, and ideas. Journal of Economic Perspectives 24 (2), 163–188.

Nunn, Nathan, Qian, Nancy, 2011. The Potato's contribution to population and urbanization: evidence from a historical experiment. Quarterly Journal of Economics 126 (2), 593–650.

Nunn, Nathan, Trefler, Daniel, forthcoming. Domestic institutions as a source of comparative advantage. In: Gopinath, Gita, Helpman, Elhanan, Rogoff, Kenneth (Eds.), Handbook of International Economics, vol. 4, North-Holland, New York.

Nunn, Nathan, Wantchekon, Leonard, 2011. The slave trade and the origins of mistrust in Africa. American Economic Review 101 (7), 3221–3252.

Nunn, Nathan, forthcoming. Gender and missionary influence in colonial Africa. In: Akyeampong, Emmanuel, Bates, Robert, Nunn, Nathan, Robinson, James A. (Eds.), Africa's Development in Historical Perspective.

Olsson, Ola, 2004. Unbundling Ex-Colonies: A Comment on Acemoglu, Johnson, and Robinson, 2001. Goteborg University, Mimeo.

Olsson, Ola, Hibbs Jr., Douglas A., 2005. Biogeography and Long-Run Economic Development. European Economic Review 49, 909–938.

Olsson, Ola, Paik, Christopher, 2012. A Western Reversal Since the Neolithic? The Long-Run Impact of Early Agriculture. University of Gothenburg, Mimeo.

Osafo-Kwaako, Philip, 2012. Legacy of State Planning: Evidence from Villagization in Tanzania. Harvard University, Mimeo.

Peisakhin, Leonid, 2010. Living Historical Legacies: The "Why" and "How" of Institutional Persistence. Yale University, Mimeo.

Pierce, Lamar, Snyder, Jason A., 2012. Trust and Finance: Evidence from the African Slave Trade. University of California Los Angeles, Mimeo.

Puga, Diego, Trefler, Daniel, 2012. International Trade and Institutional Change: Medieval Venice's Response to Globalization. University of Toronto, Mimeo.

Putnam, Robert, Leonardi, Robert, Nanetti, Raffaella, 1993. Making Democracy Work. Simon & Schuster, New York.

Putterman, Louis, 2008. Agriculture, Diffusion an Development: Ripple Effects of the Neolithic Revolution, Economica 75, 729–748.

Putterman, Louis, Weil, David N., 2010. Post-1500 population flows and the long-run determinants of economic growth and inequality. Quarterly Journal of Economics 125 (4), 1627–1682.

Redding, Stephen J., Sturm, Daniel, Wolf, Nikolaus, 2011. History and industrial location: evidence from German airports. Review of Economics and Statistics 93 (3), 814–831.

Reid, Richard, 2013. The fragile revolution: rethinking ware and development in Africa's violent nineteenth century. In: Akyeampong, Emmanuel, Bates, Robert H., Nunn, Nathan, Robinson, James A. (Eds.), Africa's Development in Historical Perspective, Cambridge University Press, Cambridge, p. forthcoming.

Rodrik, Dani, Subramanian, Arvind, Trebbi, Francesco, 2004. Institutions rule: the primacy of institutions over geography and integration in economic development. Journal of Economic Growth 9 (2), 131–165.

Rubin, Jared, 2011. Printing and Protestants: Reforming the Economics of the Reformation. California State University, Fullerton, Mimeo.

Sokoloff, Kenneth L., Zolt, Eric M., 2007. Inequality and the evolution of institutions of taxation: evidence from the economic history of the Americas. In: Edwards, Sebastian, Esquivel, Gerardo, Márquez, Graciela (Eds.), The Decline of Latin American Economies: Growth, Institutions, and Crises, University of Chicago Press, Chicago, 83–136.

Spolaore, Enrico, Wacziarg, Romain, 2009. The diffusion of development. Quarterly Journal of Economics 124 (2), 469–529.

Spolaore, Enrico, Wacziarg, Romain, 2010. War and Relatedness. Tufts University, Mimeo.

Spolaore, Enrico, Wacziarg, Romain, forthcoming. How deep are the roots of economic development? Journal of Economic Literature.

Tabellini, Guido, 2008. The scope of cooperation: values and incentives. Quarterly Journal of Economics 123 (3), 905–950.

Tabellini, Guido, 2010. Culture and institutions: economic development in the regions of Europe. Journal of the European Economic Association 8 (4), 677–716.

Tai, En-Sai, 1918. Treaty Ports in China. Columbia University Press, New York.

Tilly, Charles, 1990. Coercion, Capital and European States, A.D. 990–1990. Blackwell Publishers, Cambridge.

Vansina, Jan, 2004. The Antecedents of Modern Rwanda: The Nyiginya Kingdom. The University of Wisconsin Press, Wisconsin.

Verdier, Thierry, Bisin, Alberto, 2000. Beyond the melting pot: cultural transmission. marriage and the evolution of ethnic and religious traits. Quarterly Journal of Economics 115, 955–988.

Verdier, Thierry, Bisin, Alberto, 2001. The economics of cultural transmission and the dynamics of preferences. Journal of Economic Theory 97, 298–319.

Voigtlaender, Nico, Voth, Hans-Joachim, 2012. Perpetuated persecution: the medieval origins of anti-semitic violence in Nazi Germany. Quarterly Journal of Economics 127 (3), 1339–1392.

Waldinger, Maria, 2012. Missionaries and Development in Mexico. London School of Economics, Mimeo.

Wang, Ke-Wen, 1998. Modern China: An Encyclopedia of History, Culture, and Nationalism. Garland Publisher, New York.

Weber, Max, 1930. The Protestant Ethic and the Spirit of Capitalism. Routledge, London.

Whatley, Warren, forthcoming. The trans-atlantic slave trade and the evolution of political authority in West Africa. In: Akyeampong, Emmanuel, Bates, Robert H., Nunn, Nathan, Robinson, James A. (Eds.), Africa's Development in Historical Perspective, Cambridge University Press, New York.

Wittfogel, Karl A., 1957. Oriental Despotism: A Comparative Study of Total Power. Yale University Press, New Haven.

Woodberry, Robert D., 2004. The Shadow of Empire: Christian Missions, Colonial Policy, and Democracy in PostColonial Societies. PhD Dissertation in Sociology, University of North Carolonia at Chapel Hill.

Woodberry, Robert D., 2012. The missionary roots of liberal democracy. American Political Science Review 106 (2), 244–174.

Zerbe, Richard O., Anderson, C. Leigh, 2001. Culture and fairness in the development of institutions in the California gold fields. Journal of Economic History 61 (1), 114–143.

CHAPTER EIGHT

Institutions and Economic Growth in Historical Perspective

Sheilagh Ogilvie and A.W. Carus
Faculty of Economics, University of Cambridge, United Kingdom

Abstract

This chapter surveys the historical evidence on the role of institutions in economic growth and points out weaknesses in a number of stylized facts widely accepted in the growth literature. It shows that private-order institutions have not historically substituted for public-order ones in enabling markets to function; that parliaments representing wealth holders have not invariably been favorable for growth; and that the Glorious Revolution of 1688 in England did not mark the sudden emergence of either secure property rights or economic growth. Economic history has been used to support both the centrality and the irrelevance of secure property rights to growth, but the reason for this is conceptual vagueness. Secure property rights require much more careful analysis, distinguishing between rights of ownership, use, and transfer, and between generalized and particularized variants. Similar careful analysis would, we argue, clarify the growth effects of other institutions, including contract-enforcement mechanisms, guilds, communities, serfdom, and the family. Greater precision concerning institutional effects on growth can be achieved by developing sharper criteria of application for conventional institutional labels, endowing institutions with a scale of intensity or degree, and recognizing that the effects of each institution depend on its relationship with other components of the wider institutional system.

Keywords

Institutions, Economic growth, Economic history, Private-order institutions, Public-order institutions, Parliaments, Property rights, Contract enforcement, Guilds, Serfdom, The family, Maghribi traders, Champagne fairs

JEL Classification Codes

N01, N03, N04, N07, O17, P00, N05

8.1. INTRODUCTION

The literature on economic growth, old and new, rests on wide-ranging and often unexamined historical assumptions, which therefore raise many fundamental questions. Where and when did economies develop the threshold levels of property rights and market functioning which neoclassical growth models implicitly assume to be met (Aron, 2000)? What are the institutional origins of the asymmetries between sectors which underlie

Handbook of Economic Growth, Volume 2A
ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53538-2.00008-3

dualistic growth models (Lewis, 1954, 1958; Ranis and Fei, 1961)? What institutional arrangements have fostered growth-favoring incentives for human capital investment and innovation in some societies and growth-inhibiting ones in others, as emphasized by endogenous growth models (Romer, 1987, 1990; Aghion and Howitt, 1992; Grossman and Helpman, 1991)? Why have institutional rules favored collective action to resist technological innovations in some societies, but not in others (Parente and Prescott, 2000, 2005)? What are the institutional arrangements that influence demographic behavior and the trade-off between quality and quantity of children in unified growth theory (Galor, 2005a,b)? How has socio-political conflict in past centuries engendered the institutions that foster or stifle economic growth (Acemoglu et al. 2005)?

Recognizing the importance of such questions, the growth literature has increasingly filled in these blanks and made explicit claims about economic history and institutions. Yet some of these claims are not, on closer examination, supported by historical evidence. Others are controversial, and must be revised in the light of what is known. Still other claims are probably right, but not for the reasons given by those who make them. In many ways, then, research in economic history has still hardly been brought to bear on the institutional sources of long-run economic growth.

No single essay could discuss all the implications of economic history regarding the effects of institutions on growth, and this one does not seek to do so either. Instead, we single out eight of the most important lessons historical research can offer economists trying to understand the relationship between institutions and growth.

One common view in the growth literature is that history shows that private-order institutions can substitute for public-order ones in enabling markets to function (North and Thomas, 1970, 1971, 1973; North, 1981; Milgrom et al. 1990; Greif, 1989, 2006c; Greif et al. 1994). Past societies are supposed to have lacked public authorities able and willing to enforce the institutional rules for economic activity, and some of the literature has come to accept the view that private-order substitutes—coalitions, networks, guilds, communities, collective reprisal systems, private judges, serfdom—successfully replaced them. Economic history does not support this view, as emerges repeatedly from the empirical research surveyed in this chapter: on the Maghribi traders and the Champagne fairs in Lesson 1, on merchant guilds in Lesson 3, on peasant communities in Lesson 4, and on serfdom in Lesson 8. Historical evidence suggests strongly that although markets are required for economies to grow, public-order institutions are necessary for markets to function.

This central role of public-order institutions in economic growth has been recognized in parts of the literature (Acemoglu et al. 2005). Parliaments manned by wealth holders are widely viewed as a major component of beneficial public-order institutions, and particular attention has been devoted to the idea that parliamentary powers increased significantly in Britain after 1688, creating the institutional preconditions for the Industrial Revolution three quarters of a century later (North and Weingast, 1989; Acemoglu et al. 2005; Acemoglu and Robinson, 2012). Lesson 2 surveys the historical evidence on

18th century European parliaments in general, and the Glorious Revolution in particular, and finds that parliaments manned by wealth holders historically have a very mixed record of supporting economic growth. Whether a strong parliament manned by wealth holders supported growth in practice depended on underlying institutional mechanisms at lower levels of politics and the economy, which influenced how wealth holders obtained wealth, how they got parliamentary representation, and how parliament could be used to further policies and institutions that fostered rather than stifling growth.

A different way in which the literature has pursued the role of public-order institutions in economic growth is by seeking to classify political as well as economic institutions according to whether they have, historically, proved favorable to growth. One part of the literature has distinguished between open-access social orders which have facilitated economic growth, and closed-access orders which have hampered it (e.g. North et al. 2006, 2009). Another approach has been to distinguish between political and economic institutions that have favored growth by being inclusive, and those that have impeded it by being extractive (e.g. Acemoglu and Robinson, 2012). Lesson 3 surveys these classification systems and suggests that greater precision can be achieved by drawing a more constrained distinction, between generalized institutions (whose rules apply uniformly to all economic agents, regardless of their identity or membership in particular groups), and particularized institutions (which apply only to a subset of agents in the economy). The explanatory potential of this distinction is explored in Lesson 3 in the context of the institutional bases for the growth in long-distance trade during the medieval and early modern Commercial Revolution, and in Lesson 5 in the context of property rights in Britain before and during the Industrial Revolution.

Property rights play an overwhelmingly important role in the entire literature on institutions and economic growth, and history has been employed in this literature in numerous ways. Historical evidence is widely used to support the view that property rights have been the single most important institutional influence on economic growth at least since medieval times (North and Thomas, 1970, 1971, 1973; North, 1981, 1989, 1991; North and Weingast, 1989; Greif et al. 1994; Acemoglu and Johnson, 2005; Acemoglu et al. 2005; Acemoglu, 2009). Other parts of the literature, by contrast, have questioned the very idea that property rights played any role at all in economic growth (Clark, 2007; McCloskey, 2010). Despite the fact that economic history has been mobilized to support both sides of this debate, historical research findings have still not been fully brought to bear on the emergence of property rights, the multiple ways in which they can affect economic growth, and their importance relative to other institutions. Lessons 4–6 address the various challenges this has created. Lesson 4 considers the view that property rights institutions are both separable from, and more important than, contracting institutions (Acemoglu and Johnson, 2005). Historical evidence casts doubt on this idea: both types of institutions involved relationships between ordinary people and rulers, and both had to improve jointly before growth could occur. Lesson 5 asks why property rights are supposed to be good for growth and what precise characteristics they must have in order to provide these benefits. Surveying the evidence for Britain before and during the Industrial Revolution, it finds that in order for property rights to support growth, they not only had to be well defined, private, and secure, but also *generalized* in the sense of applying to all agents in the economy, not just to a privileged subset. Security, however, is the feature of property rights most strongly emphasized as a key to economic growth, both historical and modern. Lesson 6 subjects security of property rights to closer analysis. Surveying the evidence for Europe since the medieval period, it finds that the security of property rights cannot be analyzed without breaking down the concept into three types—security of ownership, security of use, and security of transfer. Security on all three dimensions, the historical evidence reveals, was a matter of gradation rather than outright presence or absence. This explains why it has been possible for the economic history of medieval and early modern Europe to be used to argue both that property rights were irrelevant to economic growth and that they played a central role in causing it to take place.

Although the literature on economic growth has tended to focus on one type of institution at a time, its attempts to classify institutional regimes as favorable or harmful to growth tacitly recognize that institutions are embedded in wider institutional systems. The historical evidence surveyed in this chapter highlights the importance of analyzing not just each institution in isolation but also how it interacts with other components of the surrounding institutional system. This emerges clearly in Lesson 4 where we see how contracting and property institutions were jointly necessary to encourage economic growth during the agricultural revolution. The same importance of the institutional system as a whole emerges from the survey in Lesson 7 of historical demography, which has come to play an increasingly important role in recent literature on economic growth (Galor, 2005a,b; Acemoglu, 2009; Guinnane, 2011). Historically, it turns out that both contributory factors such as demographic responsiveness to economic signals, women's position, and human capital investment; and the over-arching relationship between demographic behavior and economic growth, resulted not from any specific type of family institution in isolation, but rather from the interaction of multiple components of the wider institutional system.

The literature on economic growth has been riven for decades by the debate about whether institutions are merely epiphenomena of more fundamental natural and geographical factors (e.g. Sachs, 2003), are efficient solutions to economic problems (e.g. North and Thomas, 1970, 1973; Greif, 2006c), or result from socio-political conflicts over distribution (e.g. Acemoglu et al. 2005; Ogilvie, 2007b). The historical institutions examined in Lesson 8 provide plentiful evidence that distributional conflicts are central, both to the development of institutions and to their impact on growth. The explanatory power of the conflict approach to institutions is illustrated particularly clearly by the institution of serfdom, which has attracted repeated attention from economists because of its impact on agricultural performance and thus on overall economic growth in the centuries before and during industrialization (Domar, 1970; North and Thomas, 1970, 1973; Acemoglu and Wolitzky, 2011; Acemoglu et al. 2011). The historical evidence on serfdom confirms the centrality of distributional conflicts to the rise, survival, and disappearance of key institutions, and provides a particularly vivid example of how the problem of the lack of a political Coase theorem must be solved in order for institutions to change. But it also shows the importance of analyzing any given institution as one component of a wider institutional system—an analytical point that reappears many times throughout the lessons that follow.

8.2. LESSON 1: PUBLIC-ORDER INSTITUTIONS ARE NECESSARY FOR MARKETS TO FUNCTION

Markets are necessary for economic growth, and this raises the question of what institutions are necessary for markets to function. Economic history is widely supposed to support the claim that the functioning of the market does not necessarily require public-order institutions: private-order institutions can substitute for them. This is taken to imply that modern poor economies can achieve sustained economic growth without good governments or well-functioning legal systems, since private-order substitutes have a successful historical record of sustaining growth (Helpman, 2004; Dixit, 2004, 2009; Dasgupta, 2000; World Bank, 2002). This claim is factually mistaken, as a closer look at the evidence shows.

Private-order institutions are those formed through voluntary collective action by private agents without any involvement of public authorities. Public-order institutions, by contrast, are those associated with the formal public authorities of a society—states, local governments, bureaucracies, legal systems, rulers, courts, and parliaments (Katz, 1996, 2000). A few examples apparently supporting the view that private-order institutions have a successful track record in underpinning markets have attained the status of stylized facts within the economics profession more widely, and are repeatedly cited (Aoki, 2001; Bardhan, 1996; Ba, 2001; Bernstein, 2001; Clay, 1997; Dasgupta, 2000; Dixit, 2004, 2009; Faille, 2007; McMillan and Woodruff, 2000; Miguel et al. 2005; Helpman, 2004; O'Driscoll and Hoskins, 2006; World Bank, 2002). But these examples turn out to be false or misleading. When the evidence is examined more closely, the well-known stylized facts disappear, and there is no indication that private-order institutions could by themselves provide, or ever have provided, an institutional framework for markets.

The only way to show this is to look at the evidence in detail. Since we cannot do this for every such stylized example, we delve more deeply into the two cases that are most widely cited in the literature on economic growth. The first is the case of the Jewish Maghribi traders, who are supposed to have sustained successful commercial growth over long distances between the late 10th and the early 12th century using a private-order institution called a coalition (Greif, 1989, 1993, 2012). The second is the example of the Champagne fairs in what is now northern France, which grew to be the most important European trading center from the late 12th to the late 13th century, and are supposed to have achieved this growth by ensuring contract enforcement through private judges (Milgrom et al. 1990) and community-based reprisals (Greif, 2002, 2006b,c). This section looks at these cases in some detail to demonstrate why these claims are false and cannot be used to support either theory or policy. Later lessons discuss various other institutional arrangements—serfdom, village communities, merchant guilds—which are also widely portrayed as examples of efficient private-order institutions with a track record of supporting growth, and indicate where subsequent research has cast doubt upon their empirical basis.

8.2.1 The Maghribi Traders

A first widely cited historical example of the supposed irrelevance of public-order institutions and the efficacy of private-order substitutes is the Maghribi traders' coalition. The Maghribi traders were a group of Jewish merchants who traded across the Muslim Mediterranean between the 10th and the 12th centuries. Everything we know about them comes from the Geniza (synagogue storeroom) in Old Cairo, the city where most of these merchants lived, so they are often called "the Geniza merchants." There is a debate between those who claim that most of the Geniza merchants came from the Maghreb (essentially the region now occupied by Tunisia and Libya) and rarely established commercial relationships with non-Maghribi Jewish traders (Greif, 1989, 1993, 2012), and others who point out that these merchants neither exclusively came from, nor solely traded in, the Maghreb (see Goldberg, 2005, 2012a,b,c;Toch, 2010; Edwards and Ogilvie, 2012a). Without prejudging this debate, here we use "Maghribi traders" since the term is established in the economics literature, although the term "Geniza merchants" is more widespread among historians.

Two influential articles have argued that these merchants formed a well-defined and cohesive coalition based on Jewish religion and family origins in the Maghreb (Greif, 1989, 1993). According to this account, these medieval Jewish merchants lacked access to effective legal institutions for monitoring and enforcing contracts. Instead, they relied on informal sanctions based on collective relationships inside an exclusive coalition. Members of the Maghribi coalition, according to this view, only used other members as commercial agents. Within this closed ethnic and religious coalition, members conveyed information about each other's misbehavior efficiently to other members, and collectively ostracized members who cheated other members. The Maghribis are supposed to have chosen this type of contracting institution both because there was no effective legal system and because they held collectivist Judaeo-Muslim cultural beliefs which contrasted with the individualistic Christian values held by the medieval Genoese merchants, who consequently chose to enforce their contracts using legal mechanisms (Greif, 1994).

The Maghribis' multilateral reputation mechanism, it is claimed, provided an effective institutional basis for the growth of long-distance trade across the Muslim Mediterranean from the late 10th to the early 12th century, and substituted for the absence of an effective legal system.

This portrayal of the medieval Maghribi traders has been widely deployed to draw lessons for modern economic growth. Some use this characterization of Judaeo-Muslim collectivism versus European individualism to argue that it is cultural differences that are central to both institutions and growth (Aoki, 2001; Mokyr, 2009). Others claim that the Maghribi traders show that economic growth does not require public legal mechanisms but can be based on private-order institutions (Clay, 1997; Faille, 2007; Greif, 1989, 2006b,c; McMillan and Woodruff, 2000; O'Driscoll and Hoskins, 2006), or that the social capital of closely knit networks can effectively support market-based economic growth (World Bank, 2002; Miguel et al. 2005). Still others incorporate this model of the Maghribi traders into their accounts of how informal, reputation-based institutions contributed to long-run productivity growth (Helpman, 2004; Dixit, 2004, 2009; Dasgupta, 2000). According to Helpman, for instance, "If we had data that allowed us to calculate TFP growth during the medieval period, we probably would have found that the institutional innovations of the Maghribi traders … led to TFP growth" (2004, pp. 118–9).

However, the empirical portrayal of the Maghribi traders' coalition (Greif, 1989,1993, 2006c) was based on a limited number of documents, which other scholars, both earlier and later, have interpreted very differently (Goitein, 1966, 1967/1993; Stillman, 1970, 1973; Udovitch, 1977a,b; Gil, 2003, 2004a,b; Friedman, 2006; Ackerman-Lieberman, 2007; Margariti, 2007; Goldberg, 2005, 2012a,b,c; Trivellato, 2009; Toch, 2010; Edwards and Ogilvie, 2012a). The coalition model requires the Maghribi traders to have formed agency relations only with other members of their closed ethnic-religious coalition, yet a number of scholars have pointed out that the Maghribi traders transacted in open and pluralistic constellations rather than a closed or monolithic coalition (Udovitch, 1977a,b; Goldberg, 2005, 2012a,b,c; Toch, 2010). Others have noted that the surviving documents show the Maghribi traders establishing agency relations with non-Maghribi Jews and even with Muslims (Goitein, 1967/1993; Stillman, 1970, 1973; Goldberg, 2005, 2012a,b,c). The existence of business relationships with non-Maghribis shows that the Maghribi traders must have had other mechanisms for contract enforcement that did not rely on collective ostracism inside a closed coalition.

Five cases from the Geniza letters were adduced as providing evidence of the existence of a Maghribi coalition (Greif, 1989, 1993, 2012). Edwards and Ogilvie (2012a) re-analyzed these cases and found that none of them substantiated the existence of a coalition, with no case in which multilateral sanctions were imposed on any opportunistic contracting party by the collectivity of the Maghribi traders. Goldberg (2012b,c) carried out a quantitative and qualitative analysis of hundreds of commercial documents in the Geniza and did not find "any case of an individual being ostracised even after an accusation of serious misconduct spread through the business circle" (Goldberg, 2012b, p. 151). Although there was some evidence that Maghribi traders made use of reputational sanctions, these involved limited transmission of information, primarily to locations and persons directly associated with the conflicting parties. Research studies of businessmen in many economies, including modern ones, find similar reputational sanctions to those observed among the Maghribi traders being used as a complement to legal sanctions (Byrne, 1930; De Roover, 1948; Macaulay, 1963; Goldthwaite, 1987; McLean and Padgett, 1997; Dahl, 1998; Gelderblom, 2003; Court, 2004; Selzer and Ewert, 2005, 2010). The use of reputation mechanisms does not imply that an economy lacks an effective legal framework for contract enforcement or is capable of growing successfully without one.

Other scholars have pointed out that the Geniza documents provide evidence of a wide array of public-order contract-enforcement mechanisms that supported contracts both among Maghribi traders and between them and other Jews and Muslims (Goitein, 1967/1993; Udovitch, 1977a,b; Gil, 2003; Goldberg, 2005, 2012a,b,c; Goitein, 1967/1993; Harbord, 2006; Goitein and Friedman, 2007; Margariti, 2007; Ackerman-Lieberman, 2007; Trivellato, 2009; Toch, 2010; Cohen, 2013). Counter to the claim that the Maghribi traders only used informal reciprocity as a basis for their business associations, with no legal forms of enterprise, the documents reveal these merchants using formal legal partnerships alongside informal business cooperation; even the latter, moreover, involved responsibilities that were recognized in courts of law (Udovitch, 1977a,b; Gil, 2003; Goldberg, 2005, 2012a,b,c; Harbord, 2006; Ackerman-Lieberman, 2007; Trivellato, 2009; Toch, 2010; Cohen, 2013). In a number of cases, Maghribi merchants enforced agency agreements using legal mechanisms; they avoided using the legal system to resolve disputes if possible, but they saw the advantages of a court judgment as a last resort (Goitein, 1967/1993; Gil, 2003; Goldberg, 2005, 2012a,b,c; Goitein and Friedman, 2007; Margariti, 2007; Ackerman-Lieberman, 2007; Trivellato, 2009; Cohen, 2013). This finding resembles those for many groups of merchants and businessmen in commercial societies between the Middle Ages and the modern day, who typically preferred to avoid litigation if at all possible, but used it as a last resort (Gelderblom, 2003; Edwards and Ogilvie, 2008, 2012a).

Commercial divergence between Maghribi and Italian traders can be explained by the broader institutional framework the two groups faced, in which public-order institutions played an important role (Goitein, 1967/1993; Stillman, 1970; Epstein, 1996; Gil, 2004a,b; Goldberg, 2005, 2012a,b,c; Van Doosselaere, 2009; Edwards and Ogilvie, 2012a). The Maghribi traders were a Jewish minority in a Muslim-ruled polity, while Genoese merchants enjoyed full political rights as citizens in their own autonomous city-state. The two groups' contrasting socio-political status had inevitable repercussions for their respective economic privileges, legal entitlements, political influence, and relations with the majority population (Goitein, 1967/1993; Epstein, 1996; Goldberg, 2005, 2012a,b,c). Political and military instability increased commercial insecurity in the central Mediterranean from the mid-11th century on, which caused the Maghribi traders to reduce the geographical scope of their trade and intensify their involvement in intraregional commerce and local industry (Stillman, 1970; Gil, 2004a,b; Goldberg, 2005, 2012a,b,c). Genoese merchants, by contrast, were protected from commercial insecurity by the Genoese navy, precisely because merchants were important in the Genoese polity (Epstein, 1996;Van Doosselaere, 2009). Finally, at the beginning of the 13th century, a powerful association of Muslim merchants, the Karimis, secured privileges from the political authorities granting it an extensive legal monopoly and excluding outsiders, including Jewish traders, from many aspects of long-distance trade (Goitein, 1967/1993).

The current state of research therefore does not empirically confirm the idea that the Maghribi traders enforced contracts through a private-order coalition. The Maghribis used reputation mechanisms indistinguishable from those used by businessmen in most economies, both historical and modern, buttressed by public-order institutions including legal partnership contracts, powers of attorney, litigation in state courts, and appeals to the local and central political authorities. The broader framework of public-order institutions also played a role in the Maghribis' ability to sustain commercial growth. The Maghribi traders therefore do not support the idea that private-order institutions substituted for missing public-order ones.

8.2.2 The Champagne Fairs

A second historical example which is widely used in support of the idea that public-order institutions are irrelevant for growth because of the effectiveness of private-order substitutes is that of the Champagne fairs. These were a cycle of trade fairs held annually in the county of Champagne, a polity governed almost autonomously by the counts of Champagne until it was annexed by the Kingdom of France in 1285. The Champagne fairs operated as the undisputed fulcrum of international exchange in Europe from c. 1180 to c. 1300, and were central to the substantial acceleration of European trade known as the medieval Commercial Revolution (Bautier, 1953, 1970; Verlinden, 1965; Edwards and Ogilvie, 2012b).

Two well-known papers by economists have argued that the Champagne fairs achieved their success with a private-order institution substituting for public-order ones. Milgrom et al. (1990) claimed that commercial growth at this most important medieval European trading location was fostered by private law-courts intermediated by "law merchants" who enforced contracts and guaranteed property rights in trade goods and capital. An alternative account was provided by Greif (2002, 2006b,c), who claimed that the Champagne fairs were sustained by a "community responsibility system," consisting of collective reprisals between corporative groups of businessmen. Both theories are based on the assumption that there were no public-order institutions able or willing to guarantee property rights or enforce contracts in 13th-century Europe, and that this compelled businessmen to devise their own private-order institutional arrangements. These ideas are widely referred to in the economics literature, but closer scrutiny casts doubt upon their empirical basis.

8.2.2.1 Private Judges

Milgrom et al. (1990) argued that the medieval expansion of international trade in centers such as the Champagne fairs was made possible by private-order courts in which private judges kept records of traders' behavior. Before agreeing on any deal, a merchant would ask a private judge about the reputation of his potential trading partner. By communicating reputational status of traders on demand, the private judges enabled merchants to boycott those who had previously defaulted on contracts. The private judges are also supposed to have levied fines for misconduct, which merchants voluntarily paid because non-payment meant losing all future opportunities to trade at the Champagne fairs. Institutional arrangements combining private judges and individual merchants' reputations created incentives for all merchants to fulfill contractual obligations, even though state enforcement was absent and repeated interactions between trading partners were rare. From this portrayal of the Champagne fairs, it was concluded that international trade expanded in medieval Europe through merchants' developing "their own private code of laws," employing private judges to apply these laws, and deploying private-order sanctions against offenders—all "without the benefit of state enforcement of contracts" (Milgrom et al. 1990, p. 2).

This view of the Champagne fairs is widely accepted by economists and policy-makers, and is used to underpin far-reaching conclusions about the institutional basis for exchange in modern economies. Dixit (2004, pp. 12–13, 47–8, 98–9) mentions private judges providing enforcement to merchant customers at the Champagne fairs as an example of a well-functioning private government. Davidson and Weersink (1998) use the Champagne fairs to specify the conditions necessary for markets to function in developing economies without adequate state enforcement. Swedberg (2003, pp. 12–13), places this portrayal of private courts in medieval Champagne at the center of his view of medieval merchant law as "laying the legal foundations for modern capitalism." Richman (2004, p. 2334–5) argues that private judges at the Champagne fairs show how "coordination among a merchant community can support multilateral exchange without relying on state-sponsored courts."

Economic historians, by contrast, have pointed out for some decades that there were no private judges at the Champagne fairs. On the contrary, the Champagne fairs were supported by a rich array of public-order legal institutions, which were voluntarily utilized by international merchants (Bautier, 1953, 1970; Terrasse, 2005; Edwards and Ogilvie, 2012b). One component of these public-order legal institutions consisted of a dedicated fair court which operated throughout the duration of each fair. The fair wardens who decided the cases in this court were princely officials, not private judges. But there were also several other levels of the princely justice-system which foreign merchants used to enforce their commercial contracts—the high tribunal of the count of Champagne as the prince, the courts of the count's bailiffs, and the courts of the district provosts (Arbois de Jubainville and Pigeotte, 1859–66; Arbois de Jubainville, 1859; Bourquelot, 1839–40; Benton, 1969; Edwards and Ogilvie, 2012b). In addition, the towns in which the fairs were held operated their own municipal courts, also attracting commercial business from international merchants. Local abbeys also had the right to operate courts at the fairs, and foreign merchants made intensive use of these abbey courts (Bautier, 1953; Terrasse, 2005). The jurisdiction of the various legal tribunals which guaranteed property rights and contract enforcement at the Champagne fairs emanated not from the merchants using the fairs, but from the public authorities, since even the municipal and abbey courts operated under devolved jurisdiction granted by the rulers of Champagne. Furthermore, there is no evidence in any surviving documents relating to the Champagne fairs that any of these tribunals applied a private, merchant-generated law-code (Edwards and Ogilvie, 2012b). The Champagne fairs therefore provide no support for theories of economic growth arguing that private-order institutions can substitute for missing public-order institutions in enabling markets to function. Markets are necessary for growth, and the Champagne fairs support the view that public-order institutions are necessary for markets.

8.2.2.2 Community-Based Reprisals

A second set of claims concerning private-order institutions at the Champagne fairs postulated that commercial growth both at these fairs and elsewhere in medieval Europe was underpinned by collective reprisals between corporative communities of businessmen (Greif, 2002, 2006b,c). In this portrayal, public law-courts did exist in medieval Europe, but could not support economic growth because they were controlled by local interests which refused to protect foreign merchants' property rights or enforce their contracts impartially. Instead, it is claimed, a private-order institution called the community responsibility system stepped into the breach by providing incentives for local courts to supply impartial justice. According to this account, all long-distance traders were organized into communities or guilds. If a member of one community defaulted on a contract with a member of another, and the defaulter's local court did not provide compensation, the injured party's local court would impose collective reprisals on all members of the defaulter's community, incarcerating them and seizing their property to secure compensation. The defaulter's community could only avoid such sanctions by ceasing to trade with the injured party's community. If this prospect was too costly, the defaulter's community had an incentive to provide impartial justice. It is claimed that this combination of corporative justice and collective reprisals provided the institutional basis for economic growth in the early centuries of the Commercial Revolution, and that the Champagne fairs were a prime example of this private-order institution in operation. This interpretation of medieval history is used to draw wider implications for economic growth, including the claim that state involvement in contract enforcement is not a precondition for impersonal exchange (Greif, 2002, pp. 201–2; Greif, 2006b, pp. 232–4).

Two main arguments were advanced in support of the view that private-order institutions effectively substituted for missing public-order institutions in supporting economic growth at the Champagne fairs (Greif, 2002, 2006b). First, it was claimed that the Champagne fairs did not have a legal system with jurisdiction over visiting merchants. The fair authorities "relinquished legal rights over the merchants once they were there. An individual was subject to the laws of his community—represented by a consul—not the laws of the locality in which a fair was held" (Greif, 2006b, p. 227). The second claim was that enforcement of merchant contracts relied on the exclusion of defaulting debtors and all their compatriots from the fairs. This threat of collective reprisals, it was argued, made merchants' communal courts compel defaulters to fulfill their contracts (Greif, 2002, p. 185).

However, there are also serious difficulties with this second private-order theory. The rulers of Champagne did not relinquish legal rights over visiting merchants and did not ever permit them to be subject solely to the laws of their own communities. For the first 65 years during which the fairs were international trading centers (c. 1180–1245), all visiting merchants were subject to the public legal system prevailing at the fairs, which consisted of courts operated by the ruler's officials or by municipal governments and abbeys under devolved jurisdiction from the ruler (Bourquelot, 1839–40, 1865; Tardif, 1855; Arbois de Jubainville, 1859; Arbois de Jubainville and Pigeotte, 1859–66; Goldschmidt, 1891; Davidsohn, 1896–1901; Bassermann, 1911; Alengry, 1915; Chapin, 1937; Bautier, 1953, 1970;Terrasse, 2005; Edwards and Ogilvie, 2012b). In 1245, the count of Champagne issued a charter exempting a subset of visiting foreign merchants (Roman,Tuscan, Lombard, and Provençal traders visiting one of the six annual fairs) from judgment by his officials, but only by bringing them under his direct jurisdiction as ruler (Bourquelot, 1865, p. 174). The ruler of Champagne neither relinquished legal rights over visiting merchants nor left them to the jurisdiction of their own communities.

The evidence indicates that the role of merchant communities at the Champagne fairs was minimal (Bautier, 1953, 1970; Edwards and Ogilvie, 2012b). No merchants had community consuls (judges) at the fairs for the first 60 years of the fairs' international importance, from c. 1180 to c. 1240. Many important groups of merchants at the fairs never had consuls or communities at all. And even the few groups of merchants that did have community consuls in later phases of the fairs' existence (after c. 1240) could only use them for internal contract enforcement and relied on the public legal system to enforce contracts between their members and merchants of different communities (Edwards and Ogilvie, 2012b). The Champagne fairs flourished as the most important centre of international trade in Europe for 80 years with no recorded collective reprisals, which were only used, in a limited way, in the final phase of the Champagne fairs' ascendancy, after c. 1260 (Bautier, 1953, 1970).

The evidence casts doubt on the claim that collective reprisals were a private-order substitute for missing public-order institutions to enforce contracts. The reprisal system was fully integrated into the public legal system; the right of reprisal required a series of formal legal steps in public law-courts; and the enforcement of reprisals relied on state coercion (Tai, 1996, 2003a,b; Boerner and Ritschl, 2002; Ogilvie, 2011). The few merchant communities at the Champagne fairs played no observable role in implementing reprisals. Rather, reprisals were imposed and enforced by the public authorities, via the public legal system (Edwards and Ogilvie, 2012b). The economic history of the Champagne fairs does not support the idea that private-order collective reprisals underpinned economic growth in the absence of public-order institutions.

8.2.2.3 Public-Order Institutions and the Champagne Fairs

On the contrary, the Champagne fairs show that the policies and actions undertaken by the public authorities were crucial for the medieval Commercial Revolution (Ogilvie, 2011; Edwards and Ogilvie, 2012b). Between the mid-11th and the late 12th century, the rulers of Champagne guaranteed the property rights of all merchants at the fairs, regardless of their community affiliation (Bautier, 1953, 1970; Bourquelot, 1865). From as early as 1148, the counts of Champagne undertook deliberate and comprehensive action to ensure property rights and personal security for merchants traveling to and from the fairs, and were unusual among medieval fair-authorities in devoting considerable political and military resources to extending such guarantees beyond their territorial boundaries (Bautier, 1953; Laurent, 1935). The counts of Champagne also ensured that the persons and property of visiting merchants were secure at the fairs themselves, enforcing property rights through their own law-courts, employing their own officials to police the streets, and cooperating with municipal and ecclesiastical officials to guarantee security in the fair towns (Bourquelot, 1839–40; Bourquelot, 1865; Laurent, 1935; Terrasse, 2005).

As already mentioned, the public authorities also provided legal contract enforcement at the fairs. The counts of Champagne operated a multitiered system of public law-courts which judged lawsuits and officially witnessed contracts with a view to subsequent enforcement. Cases involving foreign merchants were adjudicated at most levels of this public legal system (Arbois de Jubainville and Pigeotte, 1859–66; Arbois de Jubainville, 1859; Bourquelot, 1839–40; Benton, 1969). By the 1170s, the counts had supplemented ordinary public legal provision at the fairs by appointing the fair-wardens mentioned earlier, who were public officials (Goldschmidt, 1891). Public alternatives to the princely court system also existed, strengthening contract enforcement, since jurisdictional competition created incentives for courts to provide impartial judgments. Three of the Champagne fair towns operated municipal courts which had the right to judge commercial conflicts, derived most of their revenues from doing so, and successfully attracted litigation from international merchants (Bourquelot, 1865; Bautier, 1953; Arbois de Jubainville, 1859; Tardif, 1855; Terrasse, 2005). The church provided an additional set of public law-courts offering contract enforcement to merchants at the fairs, and successfully competed with princely and municipal law-courts in doing so (Bautier, 1953).

The state, in the shape of the counts of Champagne and their administrators, also contributed to the fairs' success institutionally by providing infrastructure and loan guarantees (Bautier, 1953; Bourquelot, 1865; Edwards and Ogilvie, 2012b). The counts built fortifications around the fair towns, roads connecting them, canals from the Seine into the fair town of Troyes, and large buildings to expand accommodation for visiting merchants. They granted tax breaks to other organizations, especially ecclesiastical ones, as incentives for them to provide infrastructure for merchants in the form of accommodation, warehousing, and selling space. The counts encouraged investment in fair infrastructure by granting burghers in the fair towns secure private property rights and free rights to transact in real property (Terrasse, 2005). The counts of Champagne further facilitated the development of the fairs as money markets by guaranteeing loans which merchants made at the fairs to creditors from whom obtaining payment might otherwise be difficult because of high status or privileged legal position—i.e. as rulers they insured lenders against elite confiscation (Bassermann, 1911; Schönfelder, 1988).

Finally, the counts of Champagne created a good institutional environment for commercial growth in their territory by what they did *not* do: they refrained from granting legal privileges to local merchants or other elites that discriminated against foreign merchants (Chapin, 1937; Edwards and Ogilvie, 2012b). Initially, this may have been because the four Champagne fair towns were not great centers of international trade before the fairs arose, and thus did not have powerful, institutionally entrenched guilds of indigenous merchants lobbying for privileges. Then the fairs made the counts wealthy, freeing them from the need to sell privileges to the fair towns and their elites in order to finance princely spending. But the counts also resisted the temptation to sell privileges to special interests, even though these would have brought them short-term gains at the expense of long-term growth. Under the counts, therefore, the Champagne fairs offered the combination of a continuous international trading forum with no institutional discrimination for or against any group of merchants, a combination nearly unique in 13th-century Europe (Alengry, 1915; Chapin, 1937).

The counts of Champagne provide clear evidence of the importance of the political authorities in providing the minimal requirements for market-based economic activity to flourish. They guaranteed personal safety, secure private property rights, and contract enforcement; they built infrastructure, they regulated weights and measures; they supported foreign merchant lenders against politically powerful debtors; and they ensured equal treatment of foreign merchants and locals. The distinguishing characteristic of all these institutional rules was that the counts established them not as particularized privileges granted to specific merchant guilds or communities, but rather as generalized institutional guarantees issued "to all merchants, all merchandise, and all manner of persons coming to the fair" (Alengry, 1915, p. 38). These institutional rules were then maintained and extended by each count in the interests of protecting "his fairs" as a piece of property that delivered a valuable stream of revenues. During this period, from c. 1180 to c. 1300, the Champagne fairs became the fulcrum of European trade, and public-order institutions played a major role in the economic growth that ensued.

But the centrality of public-order institutions to economic growth is a two-edged sword: good public-order institutions can contribute to growth, but bad public-order institutions can harm it. The Champagne fairs provide a clear case of this proposition in action. In 1285, Champagne was annexed by the French crown (Alengry, 1915; Bautier, 1953). The French regime that took over the Champagne fairs gradually ceased to provide the generalized institutional mechanisms that had attracted and sustained international trade since c. 1180 (Laurent, 1935; Bautier, 1953; Strayer, 1980; Boutaric, 1867; Schulte, 1900; Edwards and Ogilvie, 2012b). Security of private property rights, contract enforcement, and access to commercial infrastructure were no longer guaranteed as generalized rules applicable to everyone, but rather became particularized privileges offered (and denied) to specific merchant groups in order to serve the short-term interests of French royal policy. The new public authorities in charge of the fairs no longer guaranteed a level playing-field to all merchants—domestic or foreign, allied or non-allied—but rather granted privileges that favored some groups and discriminated against others (Alengry, 1915; Bourquelot, 1865; Strayer, 1969; Laurent, 1935). The French crown began to tax and coerce particular groups of merchants to serve its fiscal, military, and political ends. By the late 1290s, long-distance trade was deserting Champagne and moving to centers such as Bruges in the southern Netherlands where property rights and contract enforcement were more impartially provided (Schulte, 1900; Bautier, 1953; Munro, 2001; Edwards and Ogilvie, 2012b). The Champagne fairs succeeded as long as the public authorities provided generalized institutional mechanisms applicable to all traders; they declined when the regime switched to particularized institutional privileges which discriminated in favor of (and against) specific groups of merchants (Munro, 1999, 2001; Ogilvie, 2011; Edwards and Ogilvie, 2012b). The Champagne fairs show clearly that by the time of the medieval Commercial Revolution, the policies and actions undertaken by the public authorities were already crucial to economic growth—for good or ill.

What do these findings imply for economic growth more widely? Private-order institutions do not, as is sometimes assumed, have a historical track record of supporting growth by substituting for public-order institutions in guaranteeing property rights or enforcing contracts. This does not exclude a role for private-order institutions in growth, but this role appears to consist in complementing public-order institutions, not substituting for them. For centuries, the public authorities have played a central role in defining the institutional rules of the game for economic activity, for good or ill. There is no historical evidence that private-order institutions have been able to guarantee property rights or enforce contracts independently. This does not mean, however, that public-order institutions always exercise a beneficial impact on economic growth. Public-order institutions that are impartial and generalized are necessary for markets to function. But public-order institutions that are partial and particularized not only fail to support growth but may actively stifle it.

8.3. LESSON 2: STRONG PARLIAMENTS DO NOT GUARANTEE ECONOMIC SUCCESS

This places the spotlight squarely on public-order institutions. As the Champagne fairs show, the public authorities matter for growth, for good or ill. But what characteristics of public-order institutions are good for growth? An idea that has gained considerable traction in the growth literature is that economic growth requires strong parliamentary institutions representing the interests of wealth holders (North and Weingast, 1989; Acemoglu et al. 2005; Acemoglu and Robinson, 2012). For modern poor countries, this implies that strengthening parliaments will ensure the institutional basis for economic success. These are attractive arguments, since there are reasons for believing that representative government is a good thing for its own sake. But does economic history support the idea that strong parliaments are invariably beneficial for economic growth?

This idea was first proposed by North and Weingast (1989), who argued that the Glorious Revolution of 1688 strengthened the English parliament in ways that produced institutions favorable to economic growth. The case of England after 1688, they claimed, provided strong historical support for two theoretical arguments concerning why parliaments are good for growth. First, they argued, a parliament possesses an inherently greater diversity of views than a monarchical government, increasing the costs for special-interest groups of engaging in rent-seeking to secure state regulations favorable to their interests but harmful to wider economic growth (for the initial elaboration of this view, see Ekelund and Tollison (1981, p. 149)). Second, a parliament that represents wealth holders will be one that enforces their interests, which are assumed to include secure private property rights and resistance to rent-seeking by special-interest groups (North and Weingast, 1989, p. 804). Although North and Weingast did not precisely define "wealth holders," their account of 18th-century England portrayed this group as including large landowners, merchants, industrialists, and state creditors (North and Weingast, 1989, pp. 810–12, 815, 817–18). The enhanced influence of these wealth holders via greater parliamentary control over the executive after 1688 is supposed to have caused secure private property rights to emerge for the first time in any society in history and ensured that the economy grew faster and industrialized earlier in England than in comparable Western European societies such as France (North and Weingast, 1989, pp. 830–1). These arguments have influenced the growth literature by apparently providing historical support for the idea that politically inclusive bodies such as parliaments create institutions favorable to growth. In one recent formulation, "the reason that Britain is richer than Egypt is because in 1688, Britain (or England to be exact) had a revolution that transformed the politics and thus the economics of the nation" (Acemoglu and Robinson, 2012, p. 4).

Attractive though these ideas seem, there are both theoretical and empirical problems with them. The theoretical problem is that there is no reason to believe that wealth holders such as large landowners, merchants, or industrialists will necessarily seek policies and institutions that are beneficial for the growth of the whole economy. They may instead seek to establish policies and institutions that benefit themselves, regardless of whether those harm growth. The empirical problem is that historical evidence drawn from a wider sample of economies provides at best mixed support for the idea that control over rulers by parliaments, even when those parliaments represented wealth holders, ensured the creation of favorable property rights, suppressed rent-seeking, or brought about successful economic growth.

8.3.1 Did Strong Parliaments Always Create Good Institutions for Growth?

There were a number of early modern European economies which, like England, had powerful parliaments that were manned by wealth holders, exercised considerable control over the executive, and strongly influenced economic policy, but created institutions and policies that did not favor economic growth.

One example is Poland, a territory well known for the strength of its parliament (the *Sejm*), which was so strong that no ruler of Poland was able to promulgate any legislation or implement any policy without parliamentary consent (Czapliński, 1985; Mączak, 1997; Czaja, 2009). The Polish parliament represented wealth holders, who were made up of the large noble landowners, a group also strongly represented in the English parliament. But the wealth holders represented in the Polish parliament did not manifest a natural diversity of views (Mączak, 1997; McLean, 2004). Rather, they manifested a very homogeneous view, namely that the power of the state should be deployed wherever possible to enforce their own legal privileges over factor and product markets under the second serfdom (Kaminski, 1975; Kula, 1976; Mączak, 1997; Frost, 2006). This gave rise to economic policies that were harmful for economic growth, in two ways. First, the Polish parliament prevented the implementation of many economic policies that were feasible in an early modern European economy and that would have created good incentives for economic agents in the country at large to allocate resources efficiently and undertake productive investments (Topolski, 1974; Kula, 1976; Guzowski, 2013). Second, the Polish parliament successfully promoted economic policies that benefited particular groups in society, specifically the large noble landowners (*szlachta*) who were disproportionately represented in parliament (Kaminski, 1975; Kula, 1976; Mączak, 1997; Frost, 2006).

From the 16th through to the 19th century, Poland was subject to the second serfdom. As we shall see in greater detail in Lesson 8, serfdom was an institutional system that endowed landlords with coercive legal privileges over the economic choices of the vast mass of the rural population and over the operation of factor and product markets in agriculture (Topolski, 1974; Kaminski, 1975; Kula, 1976). Agriculture was the largest sector in all pre-industrial economies, and serfdom constrained agricultural growth. One result of the second serfdom was that per capita GDP was much lower, and grew much more slowly, in Eastern than in Western Europe between c. 1000 and the abolition of serfdom in the later 18th or the early 19th century (Brenner, 1976; Ogilvie, 2013b). The intensity of the second serfdom and its deleterious effects on economic growth varied considerably across Eastern-Central and Eastern Europe, and the balance of power between rulers and parliamentary bodies played a major role in this variation (Brenner, 1976; Harnisch, 1986, 1994; Cerman, 2012; Ogilvie, 2013b). The second serfdom was typically less restrictive in those societies in which the ruler had more power relative to the parliament, since this enabled the ruler to resist extremes of rent-seeking by the noble landowners who were primarily represented in parliaments in those countries (Ogilvie, 2013b; Harnisch, 1986, 1989b). Those Eastern European societies, such as Poland or Mecklenburg, which had very strong parliamentary organs representing the interests of wealth holders, were also those in which the second serfdom was most oppressive and economic growth most stifled, although the existence and direction of a causal connection between strong parliaments and strong second serfdom has not been definitively established (Harnisch, 1986, 1989b; Mączak, 1997; Cerman, 2008, 2012; Ogilvie, 2013b).

The lesson for economic growth is clear. In societies in which the wider institutional system endowed wealth holders with coercive privileges giving them large economic rents, these wealth holders used those rents to obtain representation in parliament. They then used their control over parliament to intensify their own privileges in such a way as to redistribute more wealth toward themselves, even at the expense of the rest of the economy. Under such circumstances, parliamentary control over the executive choked off growth rather than encouraging it.

It might be argued that the problem with the early modern Polish parliament was that the wealth holders it represented were landowners alone, rather than also including the merchants and industrialists emphasized by North and Weingast, and hence that Poland is not a fair test of their theory. But a second example of a European polity with strong parliamentary control over the executive, the German state of Württemberg, is not subject to this objection. Württemberg was a highly democratic German state with strong parliamentary influence over the sovereign from the late 15th century through to the 19th century (Grube, 1954, 1957, 1974; Carsten, 1959; Vann, 1984; Ogilvie, 1999). So widely recognized was the influence of the Württemberg parliament over the crown and the executive arm of government that Charles James Fox famously remarked that there were only two constitutions in Europe, that of Britain and that of Württemberg (Anon. 1818, p. 340). Württemberg also lacked an indigenous landholding nobility, so its parliament was manned almost completely by bourgeois representatives, consisting of substantial businessmen—those active in commerce and industry—selected by the communities of the c. 60 administrative districts from among their own citizenry (Vann, 1984; Ogilvie, 1997, 1999). Thus, Württemberg was a polity with a strong parliament representing bourgeois wealth holders drawn primarily from industrial and commercial occupations, and these parliamentary representatives exercised unusually strong influence over state economic policies (Vann, 1984). But the policies favored by the Württemberg parliament consisted of granting legal monopolies and other exclusive privileges to special-interest groups such as craft guilds, retailers' guilds, and cartellistic companies of merchants and industrial producers (Troeltsch, 1897; Gysin, 1989; Flik, 1990; Dormois, 1994; Medick, 1996; Ogilvie, 1997, 1999, 2004a). So ubiquitous were such privileges, even in the most highly commercialized sectors of the economy, that the Göttingen professor Meiners (1794, p. 292) described how in Württemberg external trade "is constantly made more difficult by the form which it has taken for a long time. The greatest share of trade and manufactures are in the hands of closed and for the most part privileged companies." The entrenched institutional privileges of these traditional interest groups represented in a strong parliament contributed to the stagnation of the Württemberg economy throughout the entire early modern period and its late industrialization compared even to other German territories (Boelcke, 1973, 1984; Schomerus, 1977; Gysin, 1989; Hippel, 1992; Twarog, 1997; Fliegauf, 2007; Burkhardt, 2012; Kollmer-von Oheimb-Loup, 2012).

Again, the lesson for economic growth is clear. The underlying institutions of a society influence whether a strong parliament will foster or stifle growth, since it is they that influence the mechanisms both for becoming wealthy and for getting into parliament, as well as the policies deemed desirable by parliamentary representatives. Strong control over the executive by a parliament manned by wealth holders, even those recruited from industry and commerce, will only encourage growth if the wealth holders in question regard it as in their interest to promote generalized institutional arrangements that benefit the growth of the entire economy rather than particularized institutions that redistribute wealth to themselves. The historical evidence shows that there is no guarantee that they will do so.

More autocratic German states provide a striking contrast to parliamentary Württemberg and cast further doubt on the general validity of the idea that influence over the executive by strong parliaments manned by business interests will inevitably give rise to economic policies that encourage growth. In German states such as Prussia, the sovereign was much stronger relative to the parliament than in Württemberg (Carsten, 1950, 1959; Feuchtwanger, 1970; Koch, 1990; Clark, 2006; Wheeler, 2011). As a result, by the early 19th century the executive arm of government in Prussia became strong enough to withstand much more of the rent-seeking pressure exerted by parliaments manned by representatives of wealth holders. Instead, the Prussian rulers were able to ram through institutional reforms which weakened the privileges of guilds, municipal corporations, and village communities (Rosenberg, 1958; Tipton, 1976; Brophy, 1995; Wheeler, 2011). Prussia abolished its guilds after c. 1808, while Württemberg retained them until 1864. The Prussian state even became strong enough after c. 1808 to abolish serfdom and gradually to restrict many other market-distorting institutional privileges enjoyed by both noble landlords and peasant communes (Schmoller, 1888; Henderson, 1961a,b,c; Tipton, 1976; Sperber, 1985). These state infringements on traditional institutional privileges were not possible in more democratic German territories such as Württemberg, where, although serfdom never existed in the east-Elbian form, the powers of communities over agriculture, guilds over industry, and cartellistic merchant companies over commerce were maintained, with parliamentary support, to a much later date (Tipton, 1976; Schomerus, 1977; Medick, 1996; Ogilvie, 1992, 1999). The economic policies pushed through forcibly against parliamentary protest by the autocratic Prussian state abolished the regime of privileges and rents for special-interest groups, creating better (if not perfect) incentives for the economy at large (Tipton, 1976; Hohorst, 1977). The level of economic development as measured by the best available proxy—the urbanization rate—was much higher in Prussia than in Württemberg over the entire period from 1750 to 1900, and the rate of economic growth was faster in Prussia (Edwards and Ogilvie, 2013).

The Dutch Republic provides a final example of a European society in which a strong parliament manned by wealth holders failed to create the institutional basis for sustained economic growth. From its foundation in 1581 to its dissolution in 1795, the United Provinces of the Netherlands was a republic governed by the States-General, a parliamentary government manned by representatives from each of the seven provinces; each province in turn was governed by the Provincial States, a provincial parliament (Blockmans, 1988; Israel, 1989; Koenigsberger, 2001). The Dutch Republic thus lacked a sovereign altogether and enjoyed parliamentary control over the executive at both the central and the provincial level. So democratic was its government that it strongly influenced the framing of the US Constitution in 1776 (Pocock, 2010). Dutch parliamentary institutions were manned not just by relatively small-scale businessmen such as those in Württemberg, but by large-scale, long-distance traders and industrialists. For the first century of its existence, the Dutch Republic was the miracle economy of early modern Europe, with high agricultural productivity, innovative industries at the forefront of technology, highly competitive global merchants, sophisticated financial markets, high living-standards, and rapid economic growth (DeVries, 1974; Israel, 1989; Bieleman, 1993, 2006, 2010; De Vries and Van der Woude, 1997). But after c. 1670, although the Dutch Republic retained its strong parliamentary institutions, its economy stagnated (De Vries and Van der Woude, 1997; Van Zanden and Van Riel, 2004).[1] This stagnation was caused at least partly by the power of entrenched business elites, whose parliamentary representation was one factor that enabled them to implement institutional arrangements that

[1] Van Zanden and Van Leeuwen (2012) present new macroeconomic estimates suggesting that the province of Holland experienced economic stagnation rather than actual decline between c. 1670 and c. 1800, but their figures refer solely to Holland, by far the most economically successful province of the Netherlands. Even for Holland, they find that industry had a near-zero growth rate between 1665 and 1800 and trade contracted at a rate of 0.13% p.a. between 1720 and 1800 (Tab. 4).

secured rents for themselves at the expense of the wider economy (Mokyr, 1974, 1980; Buyst and Mokyr, 1990; De Vries and Van der Woude, 1997; Van Zanden and Van Riel, 2004). Occupation by French Revolutionary armies enforced institutional reform in the Netherlands after c. 1795, which returned the economy to gradual economic growth, but even then the economy did not industrialize until the later 19th century, very tardily by European standards (Mokyr, 1974, 1980; Buyst and Mokyr, 1990; De Vries and Van der Woude, 1997; Van Zanden and Van Riel, 2004; Van den Heuvel and Ogilvie, 2013). The Dutch Republic thus had all the ingredients emphasized by North and Weingast (1989): executive controlled by strong parliament, parliament manned by wealth holders, wealth holders recruited from big business. But this did not prevent institutional putrefaction and stagnant economic growth after c. 1670.

The forces preventing strong representative institutions manned by wealth holders from giving rise to beneficial economic policies can be seen at work even in 18th-century England. North and Weingast (1989, p. 817) ask what prevented the English parliament from acting as abusively as the crown in passing bad economic regulations that benefited rent-seeking groups. Their answer is "the natural diversity of views in a legislature." Yet the example of other early modern European polities shows that legislatures do not always have a natural diversity of views. And the example of England itself shows that even an English style of parliament does not always pass beneficial economic policies.

Eighteenth-century British policies enforcing the ownership of and trade in slaves are one example of economic policies maintained by a parliament in order to sustain the property rights of the wealth holders whose interests it represented. This was recognized by Adam Smith, who argued (1776, Bk. IV, Ch. 7) that although slavery is both economically inefficient and morally repugnant, it is more difficult to restrict under a parliamentary form of government because slave-owners are represented in the parliamentary assembly and put pressure on magistrates to protect their property rights over their slaves.[2] Slavery, indeed, is an example of how there are types of security of private property rights which can be bad for economic growth, an argument we explore more fully in Lesson 5.

English parliamentary support for the mercantilistic regulations and military activities that defended the English colonies is another example. As early as 1817, the economist

[2] Smith (1776), Bk. IV, Chapter 7 ("Of Colonies"), paras. 76–77: "The law, so far as it gives some weak protection to the slave against the violence of his master, is likely to be better executed in a colony where the government is in a great measure arbitrary than in one where it is altogether free. In every country where the unfortunate law of slavery is established, the magistrate, when he protects the slave, intermeddles in some measure in the management of the private property of the master; and, in a free country, where the master is perhaps either a member of the colony assembly, or an elector of such a member, he dare not do this but with the greatest caution and circumspection. The respect which he is obliged to pay to the master renders it more difficult for him to protect the slave. …That the condition of a slave is better under an arbitrary than under a free government is, I believe, supported by the history of all ages and nations."

Jean-Baptiste Say argued that the costs of maintaining overseas colonies far outweighed the benefits. Colonialism, he argued, was sustained by means of a subsidy, mandated by the government and supported by parliament, which transferred resources from home consumers to the planter and merchant classes.[3] Some modern economic historians have also argued that the colonies cost the British economy more than they benefited it (e.g. Thomas and McCloskey, 1981), although this is contested by others who claim that colonial trade ensured the gainful use of underemployed resources (e.g. O'Brien and Engerman, 1991). O'Rourke et al. (2010) come to the conclusion that the rapid growth of world trade when mercantilist restrictions were removed in the 19th century demonstrates that in the 18th century "a regime of multilateral free trade would have been preferable to mercantilism," although they acknowledge that in a world in which other European powers were also behaving in a mercantilistic way, it may have been essential for each individual country to participate in (and win) mercantilistic conflicts. As this debate illustrates, however, 18th-century English parliamentary support for mercantilism and colonialism was a policy whose effects on the growth of the wider economy were ambiguous, while its benefits in creating rents for plantation-owners and merchants were indisputable.

Another example of an economic policy supported by the English parliament, even though it harmed the economy at large, is provided by the Corn Laws. These were a set of trade laws introduced in 1815 which imposed heavy duties on imported grain (Gash, 1961, 1972; Prest, 1977; Hilton, 1977, 2006; Ward, 2004; Schonhardt-Bailey, 2006). If it had been possible to import cheap grain, agricultural laborers, industrial workers, and manufacturers would have benefited, but landowners, whose interests were strongly represented in the British parliament, would have suffered (Fairlie, 1965, 1969; Vamplew, 1980). The Corn Laws, which increased the profits of landed wealth holders whose interests were represented in Parliament, were only abolished in 1846 under the extraordinary external pressure of harvest failure and famine in Ireland (Gash, 1961; Hilton, 1977, 2006). Even then, the abolition of the Corn Laws was widely opposed in Par-

[3] Say (1817), Bk. I, Ch. XIX, para. 25: "All these losses fall chiefly upon the class of home-consumers, a class of all others the most important in point of number, and deserving of attention on account of the wide diffusion of the evils of any vicious system affecting it, as well as the functions it performs in every part of the social machine, and the taxes it contributes to the public purse, wherein consists the power of the government. They may be divided into two parts; whereof the one is absorbed in the superfluous charges of raising the colonial produce, which might be got cheaper elsewhere; this is a dead loss to the consumer, without gain to any body. The other part, which is also paid by the consumer, goes to make the fortunes of West-Indian planters and merchants. The wealth thus acquired is the produce of a real tax upon the people, although, being centred in few hands, it is apt to dazzle the eyes, and be mistaken for wealth of colonial and commercial acquisition. And it is for the protection of this imaginary advantage, that almost all the wars of the eighteenth century have been undertaken, and that the European states have thought themselves obliged to keep up, at a vast expense, civil and judicial, as well as marine and military, establishments, at the opposite extremities of the globe."

liament on the grounds that repeal would weaken landed wealth holders and empower commercial interests (McCord, 1958; Hilton, 1977, 2006). Abolition required a heroic act of statesmanship by an individual political leader, Robert Peel, which ended his own political career and split his party for a generation (Gash, 1961, 1972), although it had the beneficial effect of reducing grain prices in Britain, increasing market integration across Europe, and favoring economic growth (Semmel, 1970; Peet, 1972; Williamson, 1990; Ward, 2004; Sharp and Weisdorf, 2013). The representation of wealth holders in the English parliament, therefore, did not inevitably result in the passage of economic policies that benefited the growth of the entire economy rather than enhancing the profits of powerful special-interest groups.

It may be true that the English state did not implement as many harmful economic policies favoring special-interest groups as did many continental European states. But this was already the case before 1688 (see, e.g. Archer, 1988; Ogilvie, 1999; Brewer, 1989), and was not necessarily because of the strength of the English parliament. An alternative explanation for the relative paucity of growth-stifling economic policies in early modern England is not so much parliamentary limits on the crown, but rather the absence of a paid local bureaucracy, which made it very difficult to enforce harmful economic policies even when they *were* promulgated by Parliament or executed by the Crown (Brewer, 1989). Most continental European economies experienced an earlier and more extensive growth of state regulation in the hothouse of early modern land-based warfare (Ogilvie, 1992). In these societies, the state appointed paid local personnel, enabling it to grant monopolies and other economic privileges to rent-seeking groups and to offer effective enforcement of these growth-stifling policies (Brewer and Hellmuth, 1999; Ogilvie, 1992, 1999). In England, by contrast, insofar as the Stuart monarchs had managed to put in place the innovation of a centralized administrative apparatus in the early decades of the 17th century, it was destroyed in the 1640s during the Civil War (North and Weingast, 1989, p. 818). Britain did not create a paid local bureaucracy in the 18th century and effective bureaucratic enforcement of regulations in the domestic economy did not begin until after c. 1800 (Brewer, 1989).

These historical findings do not imply that it is unimportant what economic policies a country's parliament is willing to support. Nor do they imply that it is undesirable for a parliament to represent a diversity of views, among which should be those of businessmen and property owners. But the sheer presence of a parliament that represents wealth holders and can influence the executive does not guarantee that a diversity of views will be represented or that growth-favoring economic policies will be implemented. A number of pre-modern European economies had strong parliaments that influenced the executive and were manned by wealth holders, including representatives recruited from commerce and industry. Yet these strong parliaments did not always represent a diversity of views or ensure good economic policies. On the contrary, if the wealth holders that were represented in parliament were themselves agreed that good economic policies were ones that were beneficial to themselves, parliamentary strength could entrench policies that were obstacles to wider economic growth. This is reflected in the fact that a number of European economies with strong parliaments manned by wealth holders remained extremely poor (Poland), experienced long-term stagnation (Württemberg), or moved from growth to stagnation (the Dutch Republic). This was the case whether that economy was located in Eastern Europe under the second serfdom, in Central Europe with strong corporative institutions, or in the comparatively commercialized northwest corner of the continent. The reason was that wealth holders, even ones recruited from big business, did not always know (or care) what economic policies would be best for generalized economic growth rather than their own particularized profits. As a consequence, parliaments manned by business representatives were capable of supporting policies that generated rents for special-interest groups rather than ones that created good incentives for the whole economy. North and Weingast (1989, p. 804) address this crucial issue for England by stating that "the institutional structure that evolved after 1688 did not provide incentives for Parliament to replace the Crown and itself engage in similar 'irresponsible' behavior." But this assertion does not explain what it was about the post-1688 institutional structure in England that prevented this from happening. The historical evidence shows that what matters for growth is not whether a country had a strong parliament (or a weak executive), but what that parliament (or executive) did. Even more important for growth was the underlying institutional framework of the society, which determined how people came to become wealth holders and hence which policies they sought through political action.

8.3.2 Was There a Discontinuity in Institutions and Growth in England after 1688?

A second test of the claim that an increase in parliamentary power in England after 1688 unleashed economic growth is provided by England alone. A more circumscribed version of the theory, after all, might argue that *something* about the style of parliament that emerged in England after 1688 was crucial for growth, even if all the other types of parliament observed in European history were not. Even for England, however, the empirical findings do not support the idea that the Glorious Revolution of 1688 marked an institutional or economic discontinuity.

Extensive parliamentary control over the crown prevailed in England long before 1688 (Goldsworthy, 1999). Since the medieval period, English monarchs had been obliged to get parliamentary consent before levying taxes (Harriss, 1975; Hartley, 1992; Hoyle, 1994). Between 1603 and c. 1642, the early Stuart monarchs (James I and Charles I) sought to restrict this longstanding parliamentary power, and this was one of the major issues underlying the English Civil War (Lambert, 1990; Braddick, 1994). This Civil War, which ended in 1651, established the precedent that the monarch could not govern without the consent of Parliament (Braddick, 1994). The monarchy was restored in 1660, and both Charles II (r. 1660–1685) and James II (r. 1685–1688) attempted to use royal prerogative to pass legislation without parliamentary consent. The Bill of Rights of 1689 explicitly declared passing a bill using royal prerogative to be illegal; but this was simply a reassertion of the English parliament's centuries-old right to veto legislation, although it did extend parliamentary authority to monitor crown spending (Goldsworthy, 1999; Harris, 2004). Although, therefore, the events of 1688 indubitably contributed to enhancing parliamentary authority over the executive, this was in large part a restatement of parliamentary controls over rulers which dated back to at least 1651, which in turn had been a reassertion of the longstanding parliamentary powers that had existed in England between the medieval period and the accession of the Stuarts in 1603 (Harrison, 1990; Goldsworthy, 1999). Only a very few of the parliamentary powers asserted in 1689 were new; most had existed for a long time; and thus the 1689 Bill of Rights must be seen as an incremental component of a longstanding evolutionary development rather than any sort of revolution in the relationship between Parliament and the executive. This casts doubt on the view that the Glorious Revolution of 1688 made a major contribution to early-18th-century economic growth, let alone to the Industrial Revolution, which only began after c. 1760 and involved relatively slow economic growth until c. 1820 (Crafts, 1987; Mokyr, 1987; Williamson, 1987; Broadberry et al. 2013).

Nor did the Glorious Revolution of 1688 mark any *economic* discontinuity. If the style of parliament that emerged in England after 1688 (whatever its features) was crucial for growth, then one should observe a discontinuity in economic growth rates in England before and after 1688. But none of the estimates for the growth rate of the English economy between 1500 and 1820 show any discontinuity around 1688. Maddison (http://www.ggdc.net/MADDISON/oriindex.htm) shows an almost stable growth rate between 1500 and 1820: if anything, growth was slightly faster during the 16th century than it was during the 17th or 18th centuries; his series shows no discontinuity around 1688. Van Zanden (2001) finds rapid growth in England in the second half of the 17th century, but slower growth in the 1700–1820 period; his series also shows no discontinuity around 1700. Broadberry et al. (2011, esp. Table 10) find high per capita GDP growth from the 1650s to the 1690s (0.69% p.a.) but much lower growth from the 1690s to the 1760s (0.27% p.a.). Murrell (2009) examines more than 50 separate data series spanning the period 1688–1701 and estimates the dates of structural breaks: he finds that the entire second half of the 17th century was a period of economic change in England, but that there was no structural break in the years following 1688. Clark (2010) proposes a different data series, which shows real GDP per capita in England hardly changing at all in the 17th century, before increasing modestly in the 18th century and growing strongly in the period 1800–1820. Clark's estimates have been questioned on several grounds, as Broadberry et al. (2011) point out, so it does not seem unreasonable to place most weight on the three broadly similar estimates of Maddison, Van Zanden, Broadberry, Campbell, Klein, Overton and Van Leeuwen. If one does so, evidence that there was a noticeable increase in growth after 1688 is conspicuous only by its absence.

Even for England, therefore, it is not possible to assign an important role to increased parliamentary power after 1688 in any explanation of economic growth or industrialization. There was no discontinuity in the growth of the English economy around 1688. This is not to deny that there may have been institutional causes of the good performance of the English economy in the early modern period. But these must have been institutional arrangements that were already causing the English economy to function well by 1500. Insofar as long-term growth had institutional sources, these resided not in sudden discontinuities but rather in the gradual development of institutional arrangements over the longer term.

What do these historical findings imply for economic growth more widely? Public-order institutions are important for markets to function, but parliaments representing business interests are not their distinguishing feature. Some economies with strong parliaments experience successful historical growth, but others stagnate or even decline, and do so partly because of institutions and policies implemented by their strong parliaments to redistribute resources toward the interests they represent. Other economies with spectacularly weak parliaments achieve successful economic growth over long historical time-spans, partly because of the weakness of those parliaments and their resulting inability to defend entrenched business interests against disruptive innovations. Historical evidence suggests the need to analyze the underlying institutions of each society which influence how wealth holders become wealthy, how they obtain parliamentary representation, and how parliamentary policy concretely affects the economic framework that fosters or stifles growth.

8.4. LESSON 3: THE KEY DISTINCTION IS BETWEEN GENERALIZED AND PARTICULARIZED INSTITUTIONS

Where does that leave us? Lesson 1 taught us that public-order institutions are indispensable for markets. But what exactly is it about public-order institutions that determines growth? In Lesson 2 we reviewed one of the popular answers—parliaments are what makes the difference—and rejected it. So the question remains: what features of public-order institutions influence growth? Economic history does suggest an answer to this question, but it requires that we look at institutions in a somewhat different way than is customary. Rather than looking at the high-profile aspects of government examined by political scientists and political historians, such as parliaments, rulers, power struggles, or revolutions, we focus on how institutions apply to the populations subject to them, and whether that application is uniform or varies systematically by group. When viewed in that perspective, it turns out that generalized institutions—those of more uniform application, i.e. more closely resembling a level playing field among the members of a society—are conducive to growth. Particularized institutions, on the other hand—those whose application varies sharply by group membership, and tilt the playing field in favor of some groups—hinder growth.

The literature has proposed various ways of classifying institutions according to their effects on growth. Some influential recent classification systems have made significant advances by recognizing the importance of political institutions for economic growth and incorporating historical evidence. Thus North et al. (2006, 2009) distinguish open-access social orders, which have benefited growth, and limited-access ones, which have harmed it. Along similar lines, Acemoglu and Robinson (2012) distinguish between inclusive and extractive institutions, where the inclusive systems encourage economic participation by large proportions of people, encourage people to make best use of their skills and choose their own jobs, allow people to make free choices, ensure secure private property, provide unbiased legal judgements, maintain impartial public contracting institutions, and permit entry of new businesses (Acemoglu and Robinson, 2012, pp. 74–75). The existence of inclusive economic institutions, in turn, depends on inclusive political institutions, which are defined more generally as those that are "sufficiently centralized and pluralistic," where centralization means that the state has a monopoly on legal violence and pluralism means that power is broadly distributed in society (Acemoglu and Robinson, 2012, p. 81). Extractive institutions, whether economic or political, are defined as being those that are not inclusive.

These proposed distinctions are useful: they focus on the historical influence of institutions on long-term growth, and they incorporate political and distributional aspects of such institutions. Their usefulness is limited, however, by their vagueness. Both distinctions are extremely broad and leave unclear exactly which aspects of a society's institutional system are critical from the authors' points of view. We believe that the historical research available to date permits the more precise distinction between what we call *generalized* and *particularized* institutions.

Generalized institutions are those whose rules apply uniformly to everyone in a society, regardless of their identity or their membership in particular groups, e.g. a state in which a rule of law is established to some degree, or a competitive market with free entry (Ogilvie, 2005d, 2011; Puttevils, 2009; Hillmann, 2013). The institutional rules of such states and markets apply to any economic agent impartially, without regard to any personal characteristic appertaining to the individual or the group he or she belongs to, rather than the transaction in question (Ogilvie, 2005d, 2011). The rules of particularized institutions, in contrast, apply differentially to different subsets of agents in the economy (Ogilvie, 2005d, 2011; Puttevils, 2009; Hillmann, 2013). Typically, these subsets consist of persons defined according to characteristics that have little or no *prima facie* bearing on the transaction classes in question. These characteristics may be anything, but in practice often include gender, religion, race, parentage, social stratum, group membership, or possession of specific socio-political privileges explicitly entitling their holders to distort markets in their own interest. Particularized institutions include those that favor particular castes, communities, or guilds, as well as systems of serfdom and slavery. Thus, for instance, the rules and entitlements of a medieval guild applied only to its own members, based on their possession of the specific legal privilege of membership, which in turn depended on non-economic criteria such as gender, parentage, religion, and other personal characteristics; non-members of the guild were treated completely differently (Ogilvie, 2005d, 2011). Likewise, as we shall see in Lesson 8, the rules and entitlements of serfdom applied differentially to serf overlords (who were endowed with privileged rights of property and transaction in land, labor, capital, and output), compared to serfs (whose property rights and transactions were institutionally limited). The rules of a guild or the rules of serfdom might guarantee your property rights or enforce your contracts, but only because of your particular identity, rights, and entitlements as a member of a particular subset of economic agents, defined according to transaction-unrelated criteria such as guild membership or serf status (Ogilvie, 2005d, 2011).

In real life there are, of course, no perfectly generalized institutions; even the historical states that best approximated a rule of law often permitted obvious lapses and inconsistencies. It is best to think of the distinction between generalized and particularized institutions as a continuum along which historical institutions are distributed. In addition, the mixture of generalized and particularized institutions is different in each society: this will be discussed in more detail when we consider comprehensive institutional systems in Lesson 7. Generalized and particularized institutions co-exist in all economies, in other words; but historically, those societies in which generalized institutions gradually came to predominate were those where sustained economic growth became possible.

The distinction between the two emerges as central in a number of historical examples of institutional frameworks that fostered—or stifled—long-term growth. To illuminate the precise institutional features and causal mechanisms involved, this section will analyze in detail one historical example widely referred to by economists, that of the institutional framework that fostered growth in long-distance commerce between the medieval period and the Industrial Revolution. Later sections of this chapter then develop the usefulness of this classification system in the context of property rights (Lesson 5), and in the context of serfdom (Lesson 8).

Let us begin, however, by exploring the distinction between generalized and particularized institutions in the growth of international trade. Between c. 1000 and c. 1800, there was a substantial and sustained growth of long-distance trade, first between Europe and its near abroad and after c. 1500 between Europe and other continents. A widely held view in the recent economics literature is that this Commercial Revolution was facilitated by particularized institutions called merchant guilds, corporative associations of wholesale traders (Greif et al. 1994; Greif, 2006c; Ostrom, 1998; Maggi, 1999; Taylor, 2002; Anderson, 2008; Dixit, 2009). Merchant guilds had existed since Greek and Roman antiquity, but became a salient institution in much of Europe between c. 1000 and c. 1500 (Ogilvie, 2011). Although they declined in some societies, particularly the Netherlands and England, from the 16th century on, they survived in many parts of southern, central, Scandinavian, and Eastern Europe into the 18th or early 19th centuries. New merchant guilds (and privileged merchant companies that often resembled guilds) formed in emerging sectors such as proto-industrial exporting and the intercontinental trade until around 1800. Merchant guilds also spread to European colonies, especially to Spanish America, where they were only abolished with independence in the 19th century (Woodward, 2005, 2007).

These particularized institutions thus indisputably *accompanied* the growth of trade in medieval and early modern Europe. But it has recently been urged that they *facilitated* it, by guaranteeing property rights and contract enforcement for long-distance merchants (Greif et al. 1994; Greif, 2006c; Gelderblom and Grafe, 2004; Ewert and Selzer, 2009, 2010; Volckart and Mangels, 1999). Unconvinced, other scholars remark that merchant guilds and associations had been formed by rent-seeking traders for millennia to tilt the playing field in their favor, and that it was, rather, the gradual emergence of more generalized institutional mechanisms that facilitated the growth of trade during the medieval and early modern Commercial Revolution (Boldorf, 1999, 2006, 2009; Dessí and Ogilvie, 2003, 2004; Lindberg, 2008, 2009, 2010; Ogilvie, 2011).

Private property rights are the first sphere in which the distinction between particularized and generalized institutions proves to be central in understanding the basis for commercial growth. In an influential article, Greif et al. (1994) proposed a theoretical model according to which, if merchants belonged to a merchant guild that could make credible collective threats against rulers, this guild could pressure rulers into committing themselves to refrain from attacking the property of guild members and to provide these guilded merchants with adequate levels of security against outside aggressors. This article went on to argue that this was actually why the merchant guild arose and existed in medieval Europe: it was an efficient solution to the problem of guaranteeing security of private property rights for long-distance merchants.

Closer empirical scrutiny, however, casts doubt on the idea that these particularized institutions played a positive role in guaranteeing private property rights during the Commercial Revolution. The enhancements to commercial property rights that merchant guilds might have generated in theory turn out to have been minor in practice; insofar as they existed, they accrued only to guild members, not the economy, or even a local economy, as a whole (Dessí and Ogilvie, 2003, 2004; Ogilvie, 2011, Ch. 6; Lambert and Stabel, 2005; Henn, 1999; Briys and De ter Beerst, 2006; Blondé et al. 2007; Harreld, 2004a,b). Furthermore, merchant guilds also engaged in activities which *reduced* the security of commercial property rights for others, by attacking the trade of rival merchants or lobbying their own governments to do so in order to defend their cartellistic privileges over particular wares, transaction types, and trade routes. These attacks created insecurity of private property rights which not only damaged competitors but spilled over (harmfully) to uninvolved third parties (Barbour, 1911; Katele, 1986; Pérotin-Dumon, 1991; Tai, 1996, 2003a,b; Reyerson, 2003; Ogilvie, 2011).

Historical research shows that it was generalized institutions that improved the security of private property rights during the Commercial Revolution (Lindberg, 2008, 2009, 2010; Ogilvie, 2011, Ch. 6). Princely states and urban governments provided generalized security to all merchants in those times and places at which long-distance trade expanded, as at the Champagne fairs (discussed in Lesson 1). Urban governments and rulers also organized infrastructure such as convoys, fortifications, military defence, and law and order, in order to attract merchants, including those who were not members of guilds (Byrne, 1916; Williams, 1931; Laurent, 1935; Bautier, 1953; Lane, 1963; Lopez, 1987; Doumerc, 1987; Nelson, 1996; Tai, 1996; Dotson, 1999; Stabel, 1999; Laiou, 2001; Middleton, 2005; Ogilvie, 2011; Edwards and Ogilvie, 2012b). Different European societies differed in the precise balance between particularized guarantees of property rights to privileged merchant guilds in return for favors, and generalized guarantees of property rights to all traders in the expectation of being able to tax an expanding trade. But those European polities which followed a more generalized path were those to which long-distance merchants migrated and in which they most vigorously generated gains from trade—Champagne under the counts in the 13th century; Bruges in the 14th; Antwerp in the 15th; Amsterdam in the 16th and early 17th; and London in the 17th and 18th centuries (Ogilvie, 2011; Gelderblom, 2005a, 2013). Long-distance trade expanded more successfully in those periods and locations in which the public authorities guaranteed property rights in a generalized way to all economic agents rather than in a particularized way to members of privileged guilds.

The distinction between particularized and generalized institutions also emerges as central to commercial growth in the evolution of contract enforcement. It has recently been maintained that merchant guilds were also an efficient solution to problems of consistent contract enforcement in international trade. Guild jurisdictions, it is claimed, offered better contract enforcement to merchants than public courts because they had greater commercial expertise, superior information, shared business values, and a special form of law (Milgrom et al. 1990; North, 1991; Benson, 1989). In one variant, merchant guilds are thought to have solved contract enforcement problems by using internal social capital to put pressure on members not to break contracts: if one guild member reneged on a business agreement, information would pass rapidly through the guild and other members would impose social sanctions on him for harming their collective reputation (North, 1991; Benson, 1998, 2002; Grafe and Gelderblom, 2010; Ewert and Selzer, 2009, 2010; Selzer and Ewert, 2005, 2010). In another variant of this claim, merchant guilds are held to have offered an efficient solution to contract enforcement via the kind of reprisals system discussed in Lesson 1: if a member of one guild defaulted on a contract with a member of another, the injured party's guild would impose collective reprisals on all members of the defaulter's guild, giving the latter an incentive to use internal peer pressure or guild courts to penalize the defaulter (Greif, 1997, 2002, 2004, 2006b,c; Boerner and Ritschl, 2005).

Closer empirical scrutiny, however, casts doubt on all variants of the idea that particularized provision of contract enforcement via merchant guilds played an important role in contract enforcement during the growth of long-distance trade. Guild jurisdictions were not universal, those that existed operated under devolved authority from the public legal system, guild tribunals were not capable of resolving complicated business conflicts, many guilded merchants preferred public jurisdictions, and there is no evidence that guild courts applied an autonomous merchant law (Woodward, 2005, 2007; Gelderblom, 2005b; Sachs, 2006; Ogilvie, 2011; Harreld, 2004a,b; Jacoby, 2003; Paravicini, 1992; Lambert and Stabel, 2005; Baker, 1979, 1986; Edwards and Ogilvie, 2012b; Kadens, 2012). Peer pressure left even less empirical trace, with almost no evidence that merchant guilds used it to enforce commercial contracts and several striking cases in which even the most powerful merchant guilds failed to sanction members for defaulting on contracts and had to petition the public authorities for enforcement (Ogilvie, 2011; Sachs, 2006; Gelderblom, 2005b; Ashtor, 1983).

Collective inter-guild reprisals existed, but progressively lost out to superior alternatives, the generalized institutions for commercial contract enforcement which we shall examine shortly. Inter-guild reprisals were widely disliked by medieval merchants themselves, since they harmed entire communities of long-distance merchants and increased the risks of trade for innocent third parties (Wach, 1868; Planitz, 1919; De Roover, 1963; Lloyd, 1977; Lopez, 1987; Tai, 1996; Sachs, 2006). These serious disadvantages were widely recognized by contemporaries, who sought to limit or abolish the reprisals system as soon as trade began to expand after c. 1050 (Mas-Latrie, 1866; Wach, 1868; Goldschmidt, 1891; Del Vecchio and Casanova, 1894; Planitz, 1919; Tai, 1996, 2003a,b; Volckart and Mangels, 1999; Laiou, 2001; Boerner and Ritschl, 2002; Ogilvie, 2011). When collective reprisals were invoked, they were fully embedded into the public legal system as a final stage in a series of formal steps based on consulting written records, mobilizing sureties, invoking arbitration panels, and litigating in public law-courts (Boerner and Ritschl, 2002; Ogilvie, 2011; Edwards and Ogilvie, 2012b). Collective reprisals against the communities of offenders were an ancient practice reaching back into antiquity (Dewey and Kleimola, 1970, 1984; Dewey, 1988). What was new in the medieval Commercial Revolution was the gradual and uneven attempt to circumscribe collective reprisals within formal, public legal proceedings (Mas-Latrie, 1866; Wach, 1868; Goldschmidt, 1891; Planitz, 1919; Cheyette, 1970; Lloyd, 1977; Tai, 1996, 2003a,b; O'Brien, 2002; Boerner and Ritschl, 2002; Fortunati, 2005; Sachs, 2006; Ogilvie, 2011; Edwards and Ogilvie, 2012b).

Peer pressure, reprisals, and rent-seeking corporate groups characterized all ancient and medieval trade, as far as we know, up to the beginning of the Commercial Revolution (Ogilvie, 2011). The new component in many European institutional systems, during that period, was the emergence of generalized institutions whose rules and entitlements applied to all economic agents, not just members of particular groups. A first set of these generalized mechanisms consisted of contractual instruments such as pledges, guarantorship, and cessions of credit (whereby a merchant sold or transferred his rights as creditor to a third party who was better able to enforce them). All three mechanisms were formal, generalized institutional innovations devised by business and legal professionals in the great medieval European trading centers (Szabó, 1983; Reyerson, 1985; Greve, 2001, 2007; González de Lara, 2005; Gelderblom, 2005b; Sachs, 2006). The notarial system of registering contracts in writing, depositing, and storing them, and ultimately certifying them before arbitration panels or in courts of law was another institutional innovation devised in Mediterranean trading centers at the beginning of the Commercial Revolution. Princes and churches had operated notarial systems before, but lay notaries providing services to private individuals emerged in the 11th century and supported the early Commercial Revolution in southern Europe (Doehaerd, 1941; Lopez and Raymond, 1955; Reyerson, 1985; Greve, 2000; Gelderblom, 2005b; Ogilvie, 2011). A little later, the development of municipal offices offering analogous registration, depository, and certification services for long-distance trading contracts in northwest Europe was another institutional innovation which had not been present in the early medieval period (Wach, 1868; Dollinger, 1970; Gelderblom, 2005b; Dijkman, 2007; Ogilvie, 2011). Arbitration panels manned by arbiters appointed from a broad circle of experienced lay judges and neutral merchants, whose decisions were recognized and enforced by public law-courts, constituted a further institutional innovation observable from the early years of the Commercial Revolution (Price, 1991; Epstein, 1996; Basile et al. 1998; Volckart and Mangels, 1999; Gelderblom, 2003, 2005b; Lambert and Stabel, 2005; Sachs, 2006; Aslanian, 2006; Ogilvie, 2011). Finally, if all these mechanisms failed, public law-courts operated by princes, feudal lords, religious institutions, and local municipalities competed to provide justice to international merchants in every locality and time-period in which long-distance trade expanded after c. 1050 (Baker, 1979; Reyerson, 1985; Basile et al. 1998; Boerner and Ritschl, 2002; Gelderblom, 2005b; Munzinger, 2006; Sachs, 2006; Dijkman, 2007; Harreld, 2004a,b; Ogilvie, 2011; Edwards and Ogilvie, 2012b). These generalized alternatives to the traditional patterns, many of them dating from the earliest years of the medieval Commercial Revolution, were consistently successful in promoting growth. Long-distance commerce grew in those places and time-periods in which generalized contracting institutions, provided by the market, the public legal system, the city government, and various other levels of the public authorities, began to offer acceptable contract enforcement which was open to all traders, not just members of particular privileged guilds.

The key feature of these new institutions for guaranteeing property rights and enforcing contracts was not that they were embedded in an open-access social order or that they occurred in polities with sufficient centralization and pluralism: those characteristics were sometimes present, but not always (Ogilvie, 2011, esp. Ch. 5). Rather, it was that these institutions created incentives consistent with economic growth: their rules and entitlements applied impartially to all economic agents rather than only to members of particular groups. Political variables undoubtedly influenced the balance between generalized and particularized institutions in different European societies. But strong representative institutions were neither a necessary nor a sufficient component of such socio-political factors since, as we saw in Lesson 2, representative political institutions could actually help entrench particularized economic institutions such as privileged, cartellistic groups of merchants.

In practice, a range of socio-political factors, in addition to the presence of representative institutions such as parliaments, helped shift the balance toward more generalized institutions in the economy more widely. One strand of research emphasizes the emergence of fiscal systems and financial markets freeing states from financial dependence on granting privileges to special-interest groups (Schofield, 1963, 2004; Elton, 1975; 'T Hart, 1989, 1993; 'T Hart, 1993; Hoyle, 1994; Fritschy, 2003; Davids, 2006). A second strand focuses on the importance of a highly diversified urban system in which towns did not act in concert but rather competed and limited each other's ability to secure privileges from the political authorities (Rabb, 1964; Ashton, 1967; Croft, 1973; Archer, 1988; 'T Hart, 1989; Britnell, 1991; Lis and Soly, 1996; De Vries and Van der Woude, 1997; Harreld, 2004a,b; Van Bavel and Van Zanden, 2004; Gelderblom, 2005a,b; Van Zanden and Prak, 2006; Nachbar, 2005; Price, 2006; Murrell, 2009). A third strand of research emphasizes the importance of having a variegated social structure which included prosperous, articulate and politically influential individuals who wished to engage in entrepreneurial activities but were not members of privileged interest-groups and hence were inclined to object to particularized institutions that imposed barriers to entry (Rabb, 1964; Ashton, 1967; Croft, 1973; De Vries, 1976; De Vries and Van der Woude, 1997). Some subset of these socio-political factors shifting the balance from particularized to generalized economic institutions prevailed in all those medieval and early modern European societies which experienced successful commercial growth. But after c. 1500 these factors coincided in two European polities, the Netherlands and England, where generalized institutions gained ground and where economic growth greatly accelerated (De Vries and Van der Woude, 1997; Ogilvie, 2000, 2011). Generalized and particularized institutions continued to co-exist in all early modern societies, but those where generalized institutions came to dominate enjoyed faster economic growth, not just in trade but also in agriculture and industry, as we shall see in the coming sections.

These historical findings have wider implications for economic growth, not least because of the many potential links between particularized institutions and social capital. Social capital, as is well known, typically involves building institutions whose rules and entitlements are characterized by "closure," i.e. a clear definition of who is a member of a group and who is not (see Coleman, 1988, pp. 104–10; Sobel, 2002, p. 151; Ogilvie, 2005d, 2011; Hillmann, 2013). To generate social capital, institutions need to have closure, information advantages, collective penalties, and commitment devices: that is, they need to be particularized. Once such institutions are formed, though, it is hard to prevent them from being abused to resist changes that threaten existing benefits enjoyed by members of the closed groups enjoying the benefits of closure. Economic history illuminates a darker side of social capital, insofar as it is generated by building particularized institutions whose rules apply exclusively to entrenched groups, rather than generalized institutions whose rules apply to everyone.

8.5. LESSON 4: PROPERTY RIGHTS INSTITUTIONS AND CONTRACTING INSTITUTIONS BOTH MATTER, AND ARE NOT SEPARABLE

Two types of institution that appear to be important for economic growth, as we have seen, are those guaranteeing private property rights and those enforcing contracts. But how precisely do they affect economic growth, and is one more important than the other? Acemoglu and Johnson (2005) have argued that these two types of institution should be strictly distinguished from one another: property rights institutions protect ordinary people against expropriation by the powerful, while contracting institutions enable private contracts between ordinary people. For these reasons, the argument continues, property rights institutions have a first-order effect on long-run economic growth, whereas contracting institutions matter much less. People can find ways of altering the terms of contracts in such a way as to avoid the adverse effects of poor contracting institutions, it is claimed, but cannot do the same against the risk of expropriation by rulers and elites (Acemoglu and Johnson, 2005).

Economic history, however, provides only mixed support for this argument. Historically, there is considerable overlap between contracting institutions and property rights institutions. Indeed, as Lessons 5 and 6 will argue, we need to pay much more analytical attention to the precise characteristics of property rights that matter for growth. But even before embarking on that analysis, the historical evidence suggests strongly that one key characteristic is the degree to which property rights can be freely transferred by contract from one person to another. When people trade, they simultaneously transfer property rights to another person and make a contract. The enforceability of the contract depends on how securely the property rights are defined, and the security of the property rights depends on whether a person is allowed to enter into contracts involving his or her property. Furthermore, rulers and elites intervene not just in property rights (e.g. by expropriating people's property) but also in contracts (e.g. by invalidating agreements, in the interest either of themselves directly or of their clients). In medieval Europe, for instance, property rights governing ownership of many assets (not just land, but also financial assets and moveable goods) were often securely guaranteed in law (Pollock and Maitland, 1895; Campbell, 2005; Clark, 2007; McCloskey, 2010). However, contracts governing transfers of these property rights were sometimes guaranteed very insecurely, particularly if they involved powerful people such as rulers, members of the elite, or people to whom rulers or elites had sold privileges (legal rights to distort markets in the purchasers' interest) (Ogilvie, 2011, 2013b). Historical evidence thus poses difficulties for the idea that one can draw a useful analytical distinction between institutions enforcing contracts and those guaranteeing property rights.

Economic history also casts doubt on the idea that poor contracting institutions do not matter because ordinary people can devise informal substitutes. As Lesson 1 discussed, the two best-known historical cases which are supposed to have demonstrated the success of informal substitutes for poor contracting institutions turn out to be factually wrong. There is no evidence that the 11th-century Maghribi traders operated an informal, private-order coalition to circumvent poor public contract enforcement. Nor is there any evidence that the 12th- and 13th-century Champagne fairs relied on private judges or community-implemented reprisals to circumvent lack of public contract enforcement. It was extremely difficult to circumvent poor contracting institutions with private-order substitutes. Instead, medieval and early modern merchants voted with their feet by moving their business from locations where public-order contract enforcement was inferior to those where it was superior (Ogilvie, 2011; Gelderblom, 2005a, 2013). Economic history does not support the view that it was easy to devise informal substitutes for poor public-order contracting institutions.

The third thing we can learn from economic history is that there are important junctures in long-term economic growth at which property institutions and contracting institutions are jointly essential, in the sense that the growth benefits of one cannot emerge until the other is present. One of the most critical of these is the European agricultural revolution. Agriculture was by far the most important sector of the pre-modern economy, and most economic historians regard a sustained increase in agricultural productivity as an important contributory factor to the European Industrial Revolution. Just such an agricultural revolution began in the Netherlands in the late 15th century, England in the late 16th, parts of France in the 18th, and various territories of German-speaking Europe at different points in the 19th (Mingay, 1963; Chorley, 1981; Bairoch, 1989; Brakensiek, 1991, 1994; Allen, 1992; Overton, 1996a,b; Campbell and Overton, 1998; Kopsidis, 2006; Olsson and Svensson, 2010). For such an increase in agricultural growth to take place, a number of institutional changes were needed—some in property institutions, others in contracting institutions. Until both sets of institutional changes took place, agriculture typically failed to grow.

Secure private property rights in land were almost certainly needed for agricultural growth, although it is important to recognize that there is debate about this issue among economic historians (Allen, 1992, 2004; Neeson, 1993; Overton, 1996a,b; Shaw-Taylor, 2001a,b). Secure private property rights in land existed in most societies in medieval and early modern Europe, as we shall see in Lesson 6. But these private property rights co-existed with and were constrained by other types of property right. The village community often collectively owned a share of the pasture, woods, and wasteland in the village, and constrained the ways in which individuals could use their privately owned arable (crop-growing) fields (Allen, 1992; Neeson, 1993; Brakensiek, 1991; Kopsidis, 2006). The importance of such communal property rights and the constraints they placed on private property rights varied considerably across pre-modern European societies, across regions within the same society, and even from one village to the next (Whittle, 1998, 2000; Campbell, 2005). They also changed over time, with communal property rights gradually being replaced by private property rights in most European societies between c. 1500 and c. 1900 (Overton, 1996a,b; Brakensiek, 1991, 1994; Olsson and Svensson, 2010).

One component of this process (which in England was called the enclosure movement) was the shift from communal to private property rights in pasture. This benefited growth not so much because it solved the tragedy of the commons (Hardin, 1968), since the whole point of community management of collective pasture was to prevent overuse (see Neeson, 1993). In England, in any case, common rights were often owned and traded privately by individuals, typically the largest farmers in the village (Shaw-Taylor, 2001a,b). Instead, the main mechanism by which privatization of common pasture encouraged agricultural growth was by reducing the transaction costs involved in flexibly shifting pasture to alternative uses, which was essential for a number of the new, higher-productivity agricultural techniques that emerged during this period (Slicher van Bath, 1963, 1977; Overton, 1996a,b).

The second component of the enclosure movement affected arable (crop-bearing) land. Typically, each European village divided up all arable land into three large tracts, which were cultivated in three-year rotation to replenish soil nutrients (Slicher van Bath, 1963, 1977; De Vries, 1976). Within each tract, each villager owned and farmed scattered strips, but the village as a whole decided on crops, rotations, and other techniques, and the whole village had collective gleaning and grazing rights on individual arable land after the harvest (Overton, 1996a,b; Brakensiek, 1991, 1994). In different European societies and regions at different dates between c. 1500 and c. 1900, these scattered, open arable strips were reorganized and consolidated to form contiguous holdings over which individuals had exclusive private property rights. This increased scale economies by reducing the time costs involved for each villager in moving from one strip to another, reduced the transaction costs of adopting new arable techniques, and increased individual incentives to invest in productivity improvements (Overton, 1996a,b).

There is considerable debate about the precise growth effects of these changes in property rights. For England, although Allen (1992) contended that such changes in property rights did not increase agricultural productivity, Overton (1996a,b) contested those arguments on grounds of inaccurate periodization, misinterpretation of evidence, and sample selection bias, concluding that improvement in private property rights decreased equity but increased productivity and contributed to faster growth of agriculture. Many German territories experienced similar improvements in agricultural property rights between c. 1770 and c. 1870, often influenced by English and Dutch models, and this German enclosure movement has evoked similar debate (Brakensiek, 1991, 1994; Kopsidis, 2006). The current consensus is that in German societies, as well, replacing communal with private property rights facilitated introducing agricultural innovations, bringing new land under cultivation, shifting lands to new uses, and increasing agricultural growth (Brakensiek, 1991, 1994; Kopsidis, 2006; Fertig, 2007). Improvements in private property rights thus almost certainly did play a role in accelerating agricultural growth.

However, improving private property rights typically did not increase agricultural productivity and growth immediately. Rather, the growth benefits only emerged in the longer term. This was because property rights institutions were not enough in themselves. To have the incentive to increase productivity, the owners of land with better property rights also had to have a reasonable expectation of getting a return for the high investments entailed in introducing innovations. This required *contracting* institutions enabling farmers to obtain the labor and capital they needed, to sell agricultural surpluses, and to purchase other goods which the newly specialized farms no longer produced themselves.

First, the agricultural revolution required contracting institutions enabling the flexible mobilization of the appropriate quantity and quality of labor into the production process (DeVries, 1974, 1976; Overton, 1996a,b; Ogilvie, 2000). The new crops and crop-rotation systems that could be introduced once property rights improved required more intense digging, ploughing, fertilizing, and weeding. Higher grain and milk yields created more work in harvesting, threshing, butter-churning, and cheese-making (Chambers, 1953; Caunce, 1997). Farmers needed to use their own family's labor more intensively and to employ plentiful and flexible supplies of non-familial labor. But contracting in labor markets was often blocked by forced labor extorted from serfs, communal barriers to labor migration, wage ceilings favoring employers, limits on women's work, and other restrictive labor practices reflecting the interests of powerful individuals and groups concerned with distributing larger shares of resources to themselves (DeVries, 1976; Harnisch, 1989a,b; Klein, 2013; Ogilvie, 2004a,b, 2013a,b). Such restrictions on contracting in labor were imposed via particularized institutions such as serfdom, village communities, urban corporations, and craft guilds, whose rules did not treat all economic agents impartially, allowing them to offer and hire labor voluntarily in competitive markets with free entry, but rather differentiated between them according to non-economic criteria such as serf status, gender, religion, ethnicity, community citizenship, and guild membership (Sharpe, 1999; Ogilvie, 1997, 2000, 2004a,b; Ulbrich, 2004; Wiesner, 1989; Wiesner-Hanks, 1996; Wiesner, 2000). Even in comparatively progressive Hanover, as late as 1820, landlords used forced labor from serfs because it was costless to them, although, as the English traveler (Hodgskin, 1820, p. 85) remarked, "If the landlord had to hire laborers, he might have his work tolerably well performed, but it is now shamefully performed, because the people who have it to do have no interest whatever in doing it well and no other wish but to perform as little as possible within the prescribed time." By contrast, in those places in which the agricultural revolution began early (Flanders, the Netherlands, and England), there were good contracting institutions in the labor market, both for farm servants and for migrant agricultural workers. This ensured that the appropriate quantity of skilled and highly motivated labor could be applied at the right intensity at the appropriate point in the agricultural year (De Vries, 1974, 1976; Van Lottum, 2011a,b; Kaal and Van Lottum, 2009; Kussmaul, 1981, 1994).

Contracting institutions governing credit—not high finance in the form of loans to elites and the state, but small investment loans to ordinary people—were also essential for agricultural growth. Changing farming practice always requires at least small investments, as shown by the focus on agricultural micro-credit in modern developing economies (World Bank, 1982) as well as studies of historical European rural economies (De Vries, 1976; Holderness, 1976). Even though the early modern agricultural revolution did not involve machines, it did require capital (Habakkuk, 1994; Holderness, 1976; Lambrecht, 2009; Thoen and Soens, 2009; Van Cruyningen, 2009; Ogilvie et al. 2012). Enclosure of pastures and open fields required fences, hedges, and ditches. New crops required seed purchases. Soil improvement required extra fertilizer, sand, lime, and marl. Heavier harvests required buying more and better draught animals. Farmers and workers had to be supported during the transition to new techniques. Good contracting institutions in the Low Countries and England made it possible for Dutch and English farmers to tap the few sources of capital available in early modern Europe (De Vries and Van der Woude, 1997; Schofield and Lambrecht, 2009). In the Netherlands, capital-rich townsmen invested directly in land and loaned funds to farmers through the country's advanced credit markets (De Vries, 1974, 1976; Van Cruyningen, 2009). In England, landlords had to make their estates pay since they enjoyed few of the privileges to intervene in contracting enjoyed by their Central or Eastern European counterparts. This gave them strong incentives to lend their tenants capital for farm improvements, or even borrow themselves for this purpose in England's financial markets, which were catching up with those of the Netherlands during the 16th and 17th centuries (Holderness, 1976; Muldrew, 1993, 1998, 2003; Spufford, 2000). Good contracting institutions meant that English grain merchants were able and willing to extend credit to farmers, and incidentally to smooth price fluctuations, by speculating on the outcome of the harvest, as described by Defoe (1727, vol. 2, p. 36): "These Corn-Factors in the Country ride about among the Farmers, and buy the Corn, even in the Barn before it is thresh'd, nay, sometimes they buy it in the Field standing, not only before it is reap'd but before it is ripe."

Elsewhere in Europe, the contracting institutions that might have ensured the supply of credit to agriculture developed more slowly. Much of the available capital in the economy was accumulated by rulers through taxes, state loans, and sales of monopolies and offices, then squandered on war or court display (Brewer, 1989; Brewer and Hellmuth, 1999). Another substantial portion of available capital was levied as rents by noble landlords, and then spent on royal offices, monopolies, or conspicuous consumption (Ogilvie, 2000). In many economies—France, Spain, Italy, and many German territories—even commercial and industrial profits tended to flow into landed estates, noble status (conferring tax freedom), bureaucratic office, or legal monopolies over certain lines of business (De Vries, 1976). In societies where the greatest returns and least risk lay in purchasing land or royal favor, poor contracting institutions meant that risky economic projects such as improvement of the land were starved of capital. In many European economies, special-interest groups enjoyed privileged access to contracting institutions governing credit, from which ordinary people, including most peasants in the countryside, were excluded; although peasants were sometimes partly able to circumvent these restrictions by using undocumented and informal lending contracts, these had higher transaction costs (Ogilvie et al. 2012). Part of the delay in introducing the new agricultural techniques outside the Netherlands and England before 1750 resulted from the difficulty of saving or borrowing the requisite capital, especially for ordinary rural people who were making the main agricultural decisions. These restrictive practices in credit markets were often imposed via particularized institutions such as serfdom, village communities, and urban corporations. To give just one example, community institutions in 17th- and 18th-century Germany disallowed loans agreed between willing lenders and willing borrowers on grounds of community membership, wealth, gender, marital status, or whether the borrower was regarded favorably by the headman or village councillors (Sabean, 1990; Ogilvie, 1997; Ogilvie et al. 2012). Restrictive practices in credit markets reflected the interests of powerful individuals and groups who were concerned with redistributing resources to themselves and who made use of favorable institutional arrangements to achieve this end.

Farmers not only needed good contracting institutions to secure the inputs of labor and capital required by new agricultural techniques. They also needed good contracting institutions in output markets so they could sell their farm surpluses profitably, and buy goods they no longer produced themselves (Britnell, 1996; Grantham and Sarget, 1997; Bolton, 2012). But many of the same institutions that hindered contracting in labor and capital also impeded exchanges of food, raw materials and industrial goods. Rulers and town governments in Spain, France, and the Italian and German states often enforced particularized institutional arrangements called "staples," legal rights of prior purchase which they used to force farmers in the surrounding countryside to sell their output in towns at lower-than-market prices (De Vries, 1976; Ogilvie, 2011). This was one of the reasons the highly urbanized regions of northern Italy and southern Germany failed to stimulate an agricultural revolution in the 16th century, in contrast to the Dutch and Flemish cities, where urban consumers had to pay farmers market prices. In Spain, grain price ceilings and other institutional restrictions on contracting in output markets drove peasants off the land, and by 1797 there were almost 1000 deserted villages in rural Castile; grain had to be imported to alleviate famine (De Vries, 1976).

The particularized privileges of towns were not the only barrier to good contracting institutions that would have enabled farmers to profit from investing in the new agricultural techniques. Seigneurial tolls (internal customs barriers) blocked the development of good contracting institutions such as a national grain market in France until 1789, discouraging farmers and worsening famines (Ó Gráda and Chevet, 2002). In Bohemia, Poland, and many eastern German territories, the great landlords forced peasants to sell them grain at fixed (below-market) prices. The landlords exported the grain to Western Europe or used it to brew their own beer in demesne breweries, which they then forced the peasants to buy back from them at fixed (above-market) prices (Cerman, 1996; Ogilvie, 2001, 2005c; Dennison and Ogilvie, 2007). Blocked by poor contracting institutions, peasants could not gain enough profit from grain surpluses for it to be worthwhile investing in new techniques, even where they enjoyed secure private property rights in their land. These restrictive practices in output markets were again often imposed by landlords, village communities, or urban corporations. In early modern Bohemia, for instance, landlords used their institutional powers under serfdom to compel peasants to sell them foodstuffs at below-market prices, penalizing them when they sold grain or livestock outside the estate without first offering it to the manor (Ogilvie, 2001, 2005c). Again, these restrictive practices in output markets reflected the distributional interests of powerful individuals and groups who were concerned with distributing resources to themselves and who made use of institutional privileges to do so.

These differences in contracting institutions thus played a major role, alongside differences in property rights institutions, in deciding whether, when and where agricultural growth could take place in Europe between the 16th and the 19th century. Agricultural growth did not need just secure private property rights. Farmers had to be able to employ laborers readily, borrow money easily, sell profitably to customers, and find cheap supplies of goods they no longer made at home. The Low Countries and England were lucky: they emerged from the medieval period with serfdom weakened or non-existent (as we shall see in Lesson 8), landlords who therefore had economic weight but few legal powers, village communities that were only loosely organized, and town privileges that were poorly enforced and constrained by competition from rival towns within a highly variegated urban system (as we saw in Lesson 3). Some particularized institutions still survived in the Low Countries and England, as we shall see in Lessons 6 and 7. But in the interstices between them, new and more generalized contracting institutions sprang up and grew vigorously in the 16th and 17th centuries, before any interest-group could organize to stop them. In most other parts of Europe, however, landlords, privileged towns, and village communities retained much more extensive rights to intervene in private contracts well into the 18th century, and in some regions long past 1800. Even the abolition of seigneurial privileges in France during the Revolution, and in Prussia and many other German territories after 1808, left many restrictive contracting institutions intact. Not until traditional contracting institutions were broken down, by popular revolution, military defeat, or long and grinding social conflict, could farmers break out of the agricultural productivity trap which had long blocked growth in the largest sector of the economy (Slicher van Bath, 1963, 1977; De Vries, 1976).

Studies of the institutional preconditions for the agricultural revolution in many parts of Europe, even outside England and the Netherlands, explicitly emphasize that improvements in property rights did not in themselves lead to growth. They only did so when they were accompanied by improvements in contracting institutions in the labor market, the credit market, and the output market. Theiller (2009) shows that the emergence of better property rights in land (as evidenced by a rental market) in late medieval Normandy was triggered by the emergence of local market centers enabling and permitting peasants to sell their agricultural surpluses. Serrão (2009) shows how the emergence of urban market demand in Portugal between the 17th and 19th centuries created incentives for farmers to adopt new technologies and invest in their farms, before the liberal reforms to property rights toward the end of that period. Olsson and Svensson (2009)'s analysis of 18th- and 19th-century Sweden shows that the volume of marketable surplus was significantly affected both by privatization of property rights during the radical Swedish enclosures of the early 19th century and by the incentives created by good contracting institutions in markets for agricultural output. For 18th- and 19th-century Germany, special emphasis is placed on the development of market structures and the removal of impediments to trade, enabling the selling of agricultural output at attractive prices and with low transaction costs (Brakensiek, 1991, 1994). Even more substantial German farmers often resisted privatization of commons for an initial period because of the high risks involved and the absence of the well-functioning markets required to secure a return on the non-trivial investments involved. As a result, the reforms to German agricultural property rights proceeded very gradually, over more than a century, from c. 1770 until c. 1890, and their pace and degree varied considerably among territories, regions, and even villages, according to the availability of good contracting institutions as well as the distributional implications of institutional change and the balance of power among state officials, landlords, peasants, and rural laborers (Brakensiek, 1994, p. 139). These findings suggest a strong degree of interlinkage not only between property rights institutions and contracting institutions, but also between both sets of institutions and distributional considerations, a point to which we return in Lessons 7 and 8.

These findings have a number of wider implications for economic growth. First, property rights institutions are not separable from contracting institutions. One measure of security of private property rights is the extent to which those property rights can be securely transferred from one person to another, as we shall see in Lesson 6. This is not a trivial or incidental feature of property rights, but rather central to one of the mechanisms by which secure private property rights can benefit growth, namely by ensuring that resources are allocated to their highest-value uses. If contracting institutions are insecure, an important aspect of how private property rights benefit growth will also be insecure.

Second, property institutions and contracting institutions are jointly essential for economic growth. To unleash the growth benefits of secure private property rights, contracting institutions also have to function well, so as to enable property-owners to save and borrow capital to invest in improving the productivity of their property, employ labor to work on that property, and profitably sell output produced using that property.

Third, it is simplistic to define property rights institutions as those protecting ordinary people against expropriation by rulers and elites and contracting institutions merely as those enabling private contracts between ordinary people. Rulers and elites intervene not just in property rights but also in contracts, refusing to enforce them in their own interests or those of favored groups to whom they have sold privileges. Both property rights institutions and contracting institutions thus involve an economic relationship between ordinary people on the one hand and rulers on the other. Economic history suggests that distributional conflicts and the coercive powers of elites and rulers have always played an important role in contracting institutions, just as they have in the security of private property rights. Poor economies could not improve contracting institutions without dealing with power and distributional conflicts.

Fourth, informal alternatives cannot substitute for poor public contract enforcement. Historically, economic growth occurred when the political authorities improved generalized contract enforcement and ceased supporting particularized interventions by special-interest groups that diminished the security of contracts. Poor economies could not achieve growth by means of informal contracting institutions; they needed to address weaknesses in public-order contract enforcement.

8.6. LESSON 5: PROPERTY RIGHTS ARE MORE LIKELY TO BE BENEFICIAL FOR GROWTH IF THEY ARE GENERALIZED RATHER THAN PARTICULARIZED

Property rights may not be more important than any other type of institution, but there is little doubt that they have major effects on economic growth. It is therefore tempting to regard them as unconditionally beneficial. But the term "property rights" covers a wide variety of arrangements, and historical evidence suggests that only some of these are good for economic growth.

The historical findings, in fact, require us to remind ourselves why property rights are supposed to be good for economic growth. Three answers can be given to this question (De Soto, 1989; Milgrom and Roberts, 1992; Besley and Ghatak, 2010). First, property rights can provide good incentives for assets to be allocated to their most productive uses because property rights motivate the transfer of assets to the people who value them most. Second, property rights can give owners good incentives to devise productive uses for an asset, in order to maintain or increase its value. And third, property rights can make it possible for owners to use an asset as collateral for borrowing funds, which they can use for investments (see esp. De Soto, 2000).

What characteristics do property rights have to have in order to benefit growth via these three mechanisms? One characteristic is that property rights should be well-defined, in the sense that it is clear to everyone in the economy who owns an asset, including how he or she may use it, how and to whom it may be transferred, and what kind of contracts may be concluded concerning it. Well-defined property rights are needed to induce those who value an asset greatly to be willing to pay for its transfer to them, to create good incentives for an asset's current owners to invest in it, and to ensure that an owner can use it as collateral.

A second widely emphasized characteristic is that property rights must be private, in the sense that an asset is held by an individual entity that can exclude others from using it. Private property rights, it is argued, give the individual owner good incentives to use the asset productively, invest to maintain or increase its value, and trade or lease it to other users (Besley and Ghatak, 2010).

A third characteristic is the security of property rights so widely emphasized in the literature (see North and Thomas, 1973; North, 1989, 1991). However, as we shall see in Lesson 6, security of property rights must be broken down into at least three components: security of ownership rights; security of use rights; and security of transfer rights. All three of these are important for ensuring that assets are transferred to the users who value them most, are invested in and used productively, and are available as collateral.

But it is not enough that property rights should be well defined, private, and secure (in all senses of that term). To support growth, property rights must also be generalized, a concept we defined in Lesson 3. That is, ownership, use, and transfer rights in an asset must be available to all agents in the economy, not just to a subset of them. In order for property rights to ensure that an asset passes into the hands of the person who has the highest possible value for it because he or she will use it most productively, the ownership, use, and transfer of that asset must be open to anyone, regardless of their personal characteristics or group affiliation, and transactions involving that asset must be governed by impersonal, voluntary exchange in open and competitive markets rather than by personal characteristics or coercive action. Similarly, to provide incentives to invest in the productive use of the asset, property rights will be more effective if they are generalized, since one incentive for productive use is to maintain the value of the asset with a view to transferring it or renting it to someone else in future. If property rights in that asset are particularized, and are thus restricted to being transferred or rented to a limited circle, this will reduce the incentive for the current owner to maintain its value through productive use. Likewise, the capacity for property rights to support the use of an asset as collateral for investment loans will be limited to the extent that they do not apply to all economic agents and cannot be freely transferred to all economic agents. To the extent that property rights are particularized, therefore, that characteristic will limit all three of the ways in which these rights could in principle support economic growth. Indeed, particularized property rights may positively damage growth by denying ownership, use, and transfer of assets to everyone outside the particular subset of privileged persons, which may comprise large proportions of all agents in the economy (e.g. all women, non-whites, slaves, serfs, non-nobles, non-guild members, etc.).

The possibility that well-defined, private, and secure property rights might not always support growth is occasionally mentioned in some of the literature on institutions and growth in historical perspective. North, for instance, refers to the existence of historical property rights that did not benefit growth because they "redistributed rather than increased income" (1991, p. 110). However, there has been little further analysis of the specific characteristics of property rights that might cause them to redistribute rather than increase income. The full implications of this distinction have not received sufficient emphasis in the literature, which continues to operate on the assumption that the only characteristic of property rights that matters is their security, a concept whose precise characteristics are left quite vague.

Evidence on historical property rights and historical economic growth provides numerous examples of property rights that were clearly defined, were enjoyed privately by individuals, and were perfectly secure against confiscation, but did not benefit growth because they were particularized. That is, the rules establishing and maintaining those property rights circumscribed use of a particular asset to a particular circle of people who were defined according to non-economic criteria or group membership, and limited transfers or contracts involving that asset to that restricted circle. In historical developing economies, such particularized property rights were widespread and various, so much so that they are best analyzed by scrutinizing concrete examples. An excellent context in which to do so is provided by the debate about whether property rights got more or less favorable for growth in Britain in the century before and during the Industrial Revolution.

This issue is no mere historical quibble. Rather, it is central to assessing the historical role of property rights in economic growth, since a number of economic historians have argued that, contrary to the claims of North and Weingast (1989), restrictions on private property rights in England actually increased after 1688, contributing to England's sustained 18th-century growth and to its Industrial Revolution after c. 1780 (Harris, 2004; Hoppit, 2011; Allen, 2011). As summarized by Hoppit (1996, p. 126), "despotic power was only available intermittently before 1688, but it was always available thereafter." Proponents of this view argue that the fact that state restrictions on property rights increased before and during the first Industrial Revolution implies that economic growth does not require secure, well-defined, private property rights, but rather a powerful, interventionist state that is willing and able to take away private individuals' property against their will.

What kind of property rights were the ones that the British state started limiting in the post-1688 period? Hoppit (2011) identifies a whole array of property rights that were restricted or abolished in Britain in the 18th century. After c. 1690, the British government increasingly granted turnpike (toll road) privileges, which empowered their holders to compel land sales, and canal-building permits, which empowered compulsory dissolution of water rights. In 1748, the British government abolished Scottish hereditary jurisdictions—that is, the ownership of particular judicial offices by private individuals who had inherited them from their noble forebears. Between 1787 and 1833, the government first restricted and then abolished property rights in slaves. Between 1825 and 1850, the British government granted charters that empowered railway companies to compel the sale of tens of thousands of acres of private landed property. Between 1750 and 1830, Parliament passed more than 5200 acts of enclosure of open fields, commons, and wastes, redefining and redistributing property rights over some 21% of the land area of England, in many cases against the will of the existing owners.

How is it possible for 18th- and 19th-century Britain to be used to support such diametrically opposed conclusions about whether private property rights are good for growth? The contradiction arises largely from conflating generalized with particularized property rights. The type of property right that is good for growth is a generalized right which allocates clear disposition over an asset to a particular entity, enabling that entity to trade the asset freely and voluntarily in a market. The incentives created by this type of property right ensure that in a market economy, as long as transaction costs are not too high, the asset will be allocated to the user who values it the most, that he or she will then have the incentive to invest in its productive use, and that he or she can use it as collateral to borrow funds for investment purposes. These are the reasons that security of this type of private property right is regarded as being beneficial for economic growth. The property rights that were restricted in 18th-century England, by contrast, were largely particularized ones, which restricted use, transfers, and contracts involving assets to a limited subset of economic agents, who were defined at least partly according to non-economic criteria.

A first set of these particularized private property rights were what might be termed feudal ownership rights, which had been put in place by rulers and elites centuries earlier to generate rents for themselves. Some of these feudal ownership rights limited the freedom of disposition over land so as to maintain concentrated estates that would be large enough to support feudal armies; this applied specifically to noble or gentry land. Other feudal ownership rights assigned use and transfer rights in particular types of land to a subset of economic agents defined according to community membership or social stratum, e.g. membership in the group of substantial farmers in a village (Shaw-Taylor, 2001a,b). Feudal ownership rights also endowed members of particular social strata (e.g. the nobility) with special prerogatives over land owned, held, or used by other social strata. These feudal property rights were attached to personal or group characteristics of their holders and were typically not bought and sold impersonally in markets. As a result, they made it difficult for land to pass into the hands of people who had more productive uses for it.

Many of the salient changes in property rights in 18th- and early 19th-century Britain should not be seen as an attack on security of private property rights, therefore, but rather a reorganization of property rights from particularized to generalized ones. Bogart and Richardson (2011) argue that between 1688 and 1830 the British state did not restrict the security of private property rights, but rather responded to requests from the public to reorganize rights to land and resources in such a way as to enable individuals, families, and communities to exploit new technologies and other opportunities that the inflexible regime of particularized ownership rights inherited from the medieval past could not accommodate. For one thing, much land was held under a legal arrangement called "equitable estate" which limited its mortgage, lease, and sale. For another, many types of land tenure limited the transfer of the affected land to a small subset of persons defined according to their personal identity or membership of the local community. Third, in particular types of village (common-field villages) property rights required owners of land to maintain it in specific traditional uses, made any change in use or ownership subject to agreements with other parties, and subjected such agreements to extensive legal challenges which made them difficult to enforce (Bogart and Richardson, 2011, p. 242). The British state's granting of charters for enclosures, turnpikes, canals and railways thus did not constitute an incursion against the type of generalized private property right which is supposed to encourage growth. Rather, it enhanced the ability to break down particularized ownership rights which meant that assets could only be used by or transferred to a subset of economic agents. These property rights, because of their particularized nature, could not readily be transferred to higher-productivity users and were thus ill suited to allocating assets to their highest-value uses or responding to opportunities offered by technological innovations.

The argument advanced by Bogart and Richardson (2009, 2011) probably overstates the extent to which the reorganization of particularized into generalized property rights was caused by the Glorious Revolution of 1688. The 16th and 17th centuries had already seen extensive reorganizations of particularized ownership rights in England, including the first two waves of enclosures and a number of changes in leases and land tenures (Overton, 1996a,b; Allen, 1999). Although some types of reorganization of particularized ownership rights certainly intensified in the 18th century, many key steps had already taken place long before 1688. Bogart and Richardson (2009, 2011) tacitly acknowledge this fact by arguing that what changed after 1688 was not so much that feudal property rights began for the first time to be reorganized, but rather that the transaction costs of bringing about such reorganization were reduced by a change in the way Parliament and the crown interacted.

A second type of particularized property right which the British state began to restrict during the 18th century was the ownership of public offices. For instance, Scottish hereditary jurisdictions, which the British state abolished in 1748, granted powers of jurisdiction in civil and criminal cases, and could only be used by or transferred to a very small subset of economic agents; in fact, a hereditary jurisdiction was restricted to being owned by the heir of a clan head who had in turn inherited it from his forebears (Chambers, 1869). As Brewer (1989) emphasizes, hereditary ownership of official positions (those of judges, tax-collectors, etc.) remained widespread in many European societies in the 18th century. It contributed to the relative inefficiency of government in those societies compared to Britain, since it ensured a copious stream of rents to owners of the hereditary offices, who had incentives to exploit the coercive powers associated with such offices to obtain profits for themselves. Owners of the office of tax-gatherer skimmed off a share of the taxes collected, while owners of the office of judge demanded fees and bribes from litigants (Brewer, 1989). By abolishing property rights in public offices, the 18th-century British state was not constricting a generalized right which enabled an asset to be allocated to its highest-value use, but rather abolishing a particularized entitlement which enabled entrenched interests to exercise coercion and redistribute resources toward themselves.

The gradual abolition of slavery between 1787 and 1833 must be regarded in a similar light. The ownership of slaves was a coercive right to extort labor and other services from their original owners, the individuals who had been enslaved. Property rights in slaves were not generalized, since they did not apply equally to all economic agents: they could be enjoyed by slave-owners but not by slaves, and the conditions under which they could be transferred from slave-owners to slaves were extremely restrictive. By abolishing property rights in slaves, the state was not limiting a right enabling the asset to be allocated to its highest-value use, but rather abolishing a coercive entitlement maintained as part of a particularized institutional regime whose rules treated slaves and slave-owners completely differently from one another.

The final type of reorganization of property rights that took place in 18th-century Britain relates to the issue of eminent domain, the legal power enjoyed by the state to take private property for public use.[4] Eminent domain represents a conflict between private property rights and the public good which has still not been satisfactorily resolved in modern economies (Fischel, 1995; Benson, 2005, 2008). Sometimes a project which would benefit economic growth (e.g. an infrastructure project) can be blocked by the existence of secure private property rights which cannot be purchased at a competitive price through voluntary exchange because of market failure. Private acquisition of multiple contiguous parcels of land for a road, canal, or railway may be impossible, either because of the transaction costs of negotiating with multiple owners (a coordination problem) or

[4] The term was first used in by the Dutch jurist Grotius (1625), in the following context: "The property of subjects is under the eminent domain [*dominium eminens*] of the state, so that the state or he who acts for it may use and even alienate and destroy such property, not only in the case of extreme necessity, in which even private persons have a right over the property of others, but for ends of public utility, to which ends those who founded civil society must be supposed to have intended that private ends should give way. But it is to be added that when this is done the state is bound to make good the loss to those who lose their property." As quoted in Nowak and Rotunda (2004, p. 263).

because of the thinness of the market which gives owners a monopoly position encouraging them to demand very high prices (a holdout problem). The coordination problem may reinforce the holdout problem. These market failures may create a case for constraining private property rights. This is the only instance of state restrictions on private property rights in 18th-century England which involved an actual conflict between generalized private property rights and economic growth (Bogart and Richardson, 2011). But this type of conflict arises because of market failure, is present in all economies, and is one that modern societies have not yet resolved. It cannot therefore be taken as demonstrating that state restrictions on private property rights are generally beneficial for growth, in the absence of market failures. On the other hand, eminent domain does represent a restriction on generalized property rights which has the potential to benefit economic growth in the presence of market imperfections. This raises an important qualification to the principle that secure and generalized private property rights are invariably good for growth, and it must therefore be taken seriously in thinking about the institutional foundations of economic growth.

What do these findings imply for economic growth more widely? Not all forms of property rights—even if they are well defined, private, secure and transferable—are good for growth. Generalized property rights that can be held by, used by, and transferred to any economic agent, regardless of his or her personal identity or group affiliation, will, if markets are competitive, allocate assets to their most productive uses, give their owners the incentive to use them productively, and enable their owners to employ them as collateral. But particularized property rights that can only be held by, used by, and transferred to a small subset of economic agents, often defined according to non-economic criteria, will limit these growth benefits. People with productive uses for an asset who do not belong to the limited circle of those to whom particularized property rights in that asset apply will not be able to own, use, rent, borrow, or buy that asset. These limits on who may hold or use the asset will reduce incentives for investing in it and reduce its capacity to operate as collateral. Particularized property rights may not only fail to support growth in the ways that generalized ones do, but may positively damage growth by denying use, transfer, or rental of property to everyone outside the particular subset of privileged persons, which may include large proportions of all agents in an economy. Growth will therefore benefit from improvements in the security of *generalized* property rights, but from restrictions on the security of *particularized* property rights.

8.7. LESSON 6: SECURITY OF PRIVATE PROPERTY RIGHTS IS A MATTER OF DEGREE (AND NEEDS CAREFUL ANALYSIS)

The concept of secure private property rights, as we saw in Lesson 5, is not straightforward. Secure private property rights will affect growth differently, depending on whether they are generalized or particularized. But as the present section will argue, even the concept of "security" of property rights needs further analysis before it can be useful. The economics literature on the historical role of institutions in growth emphasizes the importance of property rights that are secure. But as indicated in Lesson 5 above, the understanding of security in this literature appears to involve at least three very different components: security of ownership, security of use, and security of transfer.

Security of ownership means that no one can take an asset away from you arbitrarily: you have a well-defined ownership right that you can reasonably expect to enforce via the legal system or some other institutional mechanism. Security of use means that no one can prevent you from exercising that ownership right by investing in improving the productivity of the asset or altering the way it is used in order to increase its yield. Security of rights of transfer means that no one can intervene to prevent you from transferring that asset temporarily or permanently to someone else by selling, mortgaging, lending, leasing, bequeathing, or otherwise alienating it.

These three components of the security of private property rights, though conflated in the literature, are both analytically and empirically distinct. Analytically, they are distinct because the three types of security are likely to affect growth in different ways and to differing degrees. Empirically, they are distinct because they can occur in different combinations: thus one may have completely secure rights of ownership in one's land but there may (or may not) be limitations on the security of one's right to decide how to use or transfer those ownership rights; likewise, one may have relatively insecure rights of ownership (in the sense that one may have it confiscated by the crown or one's feudal overlord) but be completely secure in how one can use those rights while one has them and in one's right to choose whom to sell, lease, or bequeath them to. From the point of view of the economic effects of property rights, we have already seen that limitations on ownership, use, and transfer of property are important: Lesson 5 showed that particularized property rights imposed one set of limitations; but as we shall see in the present section there are others. For the purposes of the present section, we therefore distinguish between security of ownership, security of use, and security of transfer, while recognizing that any simple classification scheme is subject to the drawback that in practice there is likely to be a continuous range of security, at least within some bounds.

Partly as a result of conflating these three different components of security of private property rights, the economics literature contains two diametrically opposed views of the historical role of secure private property rights in growth. The first assumes that secure private property rights did not exist in Europe in the medieval and early modern period (e.g. North and Weingast, 1989; Olson, 1993; Acemoglu et al. 2005; Acemoglu and Robinson, 2012). Rather, secure private property suddenly came into being in one particular economy, England, at a specific point in time, after the Glorious Revolution of 1688 (North and Weingast, 1989). This sudden and dramatic shift from insecure to secure private property rights is supposed to have enabled England to surpass other European economies and, three quarters of a century later, to become the first country to industrialize (see e.g. North and Weingast, 1989; Olson, 1993; Acemoglu et al. 2005; Acemoglu and Robinson, 2012).

Other economists, however, adopt a diametrically opposed view, arguing that private property rights were completely secure in economies such as England long before 1688, indeed as far back as the records go. Clark (2007), for instance, argues that in 12th-century England security of private property was already good and land markets were already free, so much so that medieval England already satisfied the checklist of good institutions applied to modern developing economies by the World Bank and the IMF. McCloskey (2010), too, points out that England had secure private property rights from at least the 11th century, "that there was little new in British property rights around 1700," and that many other medieval and early modern societies, both inside and outside Europe, also had secure private property rights in the medieval and early modern period (McCloskey, 2010, p. 25).

These divergent accounts of pre-modern English property rights are not just a quibble within a specialized literature. They have much wider implications for the relationship between institutions and economic growth. The view that England moved from insecure to secure private property rights in 1688 is used to argue that property rights play a fundamental role in causing economic growth. Conversely, the view that England already had secure private property rights in the medieval period (or long before) is taken to imply that property rights (and institutions in general) must be irrelevant for economic growth (Clark, 2007, pp. 148ff; McCloskey, 2010, pp. 318ff).

How is it possible for the economic history of medieval and early modern England to be used to support two such contradictory views of the role of property rights in economic growth? To answer this question, it is essential to distinguish between the different components of security (of ownership, of use, and of transfer), and to understand what is known about the historical development of property rights in England. North and Weingast (1989) argued that in 1688 private property rights became secure for three key groups: for owners of land, giving them good incentives to invest; for lenders to the state, encouraging the rise of capital markets; and for taxpayers, protecting them from state rapacity. We begin with land, since agriculture was the most important sector, and hence, property rights in its major input had the greatest potential to affect growth.

8.7.1 Secure Property Rights in Land

In conjunction with their argument regarding the importance of the English parliament after 1688, discussed above (Lesson 2), North and Weingast (1989) also argue that before 1688 landed property in England was deeply insecure even when a stable political regime was in place, because the sovereign was able to redefine property rights at will in his own favor. The Glorious Revolution of 1688, North and Weingast argue, created for the first time in any economy in history institutional limits on a ruler's ability to confiscate private land and capital; this in turn fostered "the ability to engage in secure contracting across time and space" (North and Weingast, 1989, p. 831). Olson (1993, p. 574) echoes this view, asserting that "individual rights to property and contract enforcement" became more secure in England after 1688 than in any other country, and arguing that this was why England industrialized first. Many contributions to the growth literature now accept the view that medieval and early modern Europe failed to experience economic growth because of "lack of property rights for landowners, merchants and proto-industrialists" (Acemoglu et al. 2005, p. 393). This involved insecurity not just of ownership but also of transfer and of use (at least in the sense of investment): "Most land was caught in archaic forms of property rights that made it impossible to sell and risky to invest in. This changed after the Glorious Revolution. … Historically unprecedented was the application of English law to all citizens" (Acemoglu and Robinson, 2012, p. 102). The Glorious Revolution of 1688, therefore, is supposed to have created security of private property rights in all three senses: security of ownership, certainly, but also security of use and transfer.

The empirical findings, however, do not support these claims. Secure private property rights in land existed in England from the 11th century onwards (Smith, 1974; Macfarlane, 1978; Harris, 2004; Campbell, 2005; Clark, 2007; McCloskey, 2010; Bekar and Reed, 2012). Contemporaries, ranging from small farmers to gentry landowners to great nobles to jurists, to the monarch himself, universally regarded property rights as fundamentally secure and not subject to confiscation (Pollock and Maitland, 1895; McCloskey, 2010). Thus individuals had security of ownership, in the sense of protection against arbitrary confiscation by the government or other powerful parties. However, they also had security of transfer, in the sense of having the right to sell, lease, mortgage, bequeath, and otherwise alienate their land. Royal, ecclesiastical, abbatial, and manorial law-courts competed with one another to guarantee and enforce individual ownership and transfer rights even for humble people (Smith, 1974; Macfarlane, 1978; Britnell, 1996; Whittle, 1998, 2000; Campbell, 2005; Clark, 2007; McCloskey, 2010; Briggs, 2013). Security of use rights was somewhat more constrained, for the reasons discussed in Lesson 4: in some regions and localities, village communities had some rights to regulate the ways in which private owners could use their land, specifically via communal regulation of agricultural technology, although in other regions and localities such constraints were very minor.

So overwhelming is the evidence that ownership and transfer rights in private property rights were secure in England from the Middle Ages onwards, that even North and Weingast (1989, p. 831) acknowledge "the fundamental strength of English property rights and the common law that had evolved from the Magna Carta." The Bill of Rights promulgated in 1689 after the Glorious Revolution did not in fact impose any limitation on the English government's ability to confiscate private property and did not require any compensation to be paid when the government did confiscate private property (Harris, 2004, p. 226). Fortunately, however, the English common law had ensured extensive security of ownership and transfer of property in England since the 11th century and, as Harris (2004, p. 228) points out, the judiciary showed far-reaching independence in England long before 1688.

Major political changes took place in England during the 17th century, and these led to some one-off changes in landed property rights. The Stuart monarchs made a series of abortive attempts to introduce absolutist government on the Continental European model in England between 1603 and 1641, and these initiatives involved some insecurity of ownership of landed property for opponents of the Crown. The Civil War of 1642–51 increased insecurity of ownership, as any civil war will, and the Restoration of the monarchy in 1660 resulted in further one-off changes in ownership rights. But, as McCloskey (2010) points out, for investors to have been deterred by such major political changes, they would have had to anticipate their occurrence. In actuality, none of these events were a predictable component of the early modern English property rights regime, so they cannot be viewed as a source of the kind of *ex ante* uncertainty that would have deterred investment. Moreover, the 18th century saw similar insecurity, since the regime in Britain continued to be uncertain: in 1690, serious conflicts between Parliament and Crown caused the king, William of Orange, to go back to the Netherlands; and between then and 1745, a series of Jacobite rebellions aiming to restore the Stuart dynasty operated as an organizing node for opposition to the regime. Insecurity of government always causes some insecurity of private ownership rights, but this operates mainly through expectations. It is unlikely that the regime changes of 17th-century England were expected by investors, and it is unlikely that investors attached no weight to the possibility of a Jacobite overthrow of the monarchy in the first half of the 18th century.

Quantitative analyses also cast doubt on the idea that security of any aspect of landed property rights—ownership, use, or transfer—experienced a discontinuity in England around 1688. Clark (1996) compiled a data series of land rents and land values in England between 1540 and 1750. His analysis finds that neither 1688 nor any of the other political upheavals of the 1540–1750 period marked any discontinuity. From this, he concludes that individuals must have been secure in their property rights over their land from as early as 1540.

This does not, however, mean that property rights were completely static between the medieval period and the industrial revolution, as argued by Clark (2007) or McCloskey (2010). Between c. 1350 and c. 1500, England saw significant changes in the manorial powers of landlords, communal regulation of arable fields and pastures, the costs and impartiality of legal enforcement, and the complexity of tenures and leases (Wrightson, 1982; Wrightson and Levine, 1995; Campbell, 2000, 2005, 2009; Harris, 2004; Briggs, 2009, 2013). As Lesson 4 discussed, additional changes to property rights took place during the agricultural revolution between c. 1550 and c. 1800, during which many communal property rights were restricted or abolished, tenurial forms were simplified, restrictions on alienation imposed by the inheritance system were removed, and the legal enforcement of property conflicts was improved (Overton, 1996a,b; Allen, 1999).

These changes influenced all components of security of property rights. Ownership and transfer rights were particularly strongly affected via changes in the detailed functioning of the legal system, which was a major defence against confiscation or incursion, while use rights were particularly strongly affected via changes in the communal and manorial regulation of agricultural practice, particularly enclosure and changes in leases. Such changes in security of ownership, use, and transfer of land in medieval and early modern England were incremental, did not show any discontinuity around 1688, and continued throughout the 18th century (Neeson, 1984, 1993, 2000; Allen, 1992; Overton, 1996a,b; Shaw-Taylor, 2001a,b). By the 1760s, when the Industrial Revolution was beginning, the complexity of rights governing the ownership, use, and transfer of property in England, and the transaction costs involved in enforcing such rights, had been reduced compared to medieval times. Secure rights of ownership and transfer of property existed in England from at least the 11th century onwards, and even rights of use were fairly secure in many regions. But the way these various rights operated and the economic incentives they created in practice changed over the ensuing centuries, in a gradual and incremental process.

The discussion so far has focused on England, about which the growth literature makes the strongest claims. But similar findings exist for other countries. Secure private rights of ownership, use, and transfer of landed property can be observed in a large number of European economies from the medieval period onwards. Italian economies show secure private ownership rights from the ninth century at latest, and hint strongly at secure rights of transfer in that property as well (Feller, 2004; Van Bavel, 2010, 2011; McCloskey, 2010). The Netherlands had secure private rights of ownership and transfer from the medieval period onwards, which some have argued were more extensive than in England; secure rights of use also became widespread at the latest by the beginning of the Dutch agricultural revolution, in the late 15th century (Van Bavel, 2010, 2011; DeVries, 1974; Bieleman, 2006, 2010). The German territory of Württemberg had secure private ownership and transfer rights in land from at latest the 15th century onwards, which applied to all members of society down to the poorest, included females as well as males, were unrestricted by noble privilege (since Württemberg had no landholding nobility), and included unusually generalized transfer rights such as the right to subdivide all land at will and for women to inherit equally with men; use rights were somewhat less secure because of the strong powers of Württemberg communities, but certain categories of freehold land involved secure use rights in the sense that the owner could redeploy the land to more productive uses such as textile crops (Hippel, 1977; Sabean, 1990; Röhm, 1957). In many societies in Central and Eastern-Central Europe, too, from the medieval period onwards peasants had secure private ownership rights in their land, and secure transfer rights permitting inheritance, sale, rental, and mortgaging; use rights were more restricted, but were nonetheless secure for certain categories of land (Cerman, 2008, 2012; McCloskey, 2010). Again, however, this cannot be taken to imply that property rights in these societies did not change in any way that could affect economic growth between the medieval period and the 19th century. As we saw in Lesson 4, most European societies underwent changes in ownership, use, and transfer rights in land which, together with changes in contracting institutions, contributed to an increase in agricultural productivity and an acceleration in agricultural growth between the late medieval period and the 19th century.

8.7.2 Secure Property Rights for State Creditors

A similar analysis applies to the second set of property rights whose security is emphasized as a basis for economic growth in 18th-century England. North and Weingast (1989) also argue that the Glorious Revolution of 1688, by establishing parliamentary supremacy over public finances, created an environment in which lenders could rely upon the state to meet its financial promises. This, they contend, resulted in investors placing their trust and capital in the British state instead of its foreign rivals, creating the conditions for a financial revolution which greatly improved the sophistication of credit markets, and fuelled the accelerating growth of the British economy between 1688 and 1815. These scholars conclude that the introduction of secure private property rights for state creditors in England after 1688 shows "how institutions played a necessary role in making possible economic growth and political freedom" (North and Weingast, 1989, p. 831). The security of private property for state creditors which this literature regards as being created suddenly in 1688 consists primarily of ownership rights (in the sense that the state could not confiscate creditors' assets by failing to repay), but also extends to transfer rights, since North and Weingast emphasize that security of private property for state creditors also involved "the creation of impersonal capital markets" and "the ability to engage in secure contracting across time and space" (1989, e.g. p. 831). Cameron (1989, p. 155) argues that the Glorious Revolution, by creating security for state creditors, "reacted favorably on private capital markets, making funds available for investment in agriculture, commerce, and industry"—that is,1688 saw an increase in security of both ownership and transfer.

Again, however, the empirical findings indicate that the development of property rights for state creditors was not characterized by a sudden switch from insecurity to security, whether in terms of ownership or transfer. Rather, security of ownership and transfer of assets by state creditors in England fluctuated substantially with political events, while improving incrementally over long periods of time. Analysis of the institutional rules and practices governing taxation and public borrowing in England between c. 1600 and c. 1850 shows that the Civil War (1642-51), although not a sharp break-point, marked a bigger change than 1688 (O'Brien, 2001; Harris, 2004). Overall, however, the development of property rights for state creditors was characterized by very significant elements of continuity between the early 17th and the early 19th century. O'Brien (2001) provides detailed evidence indicating that property rights for lenders to the state were secure in England in the early 17th century, and that in all important ways the institutions necessary for good governance of the public finances were in place prior to the Glorious

Revolution. Harris (2004) argues that there were fundamental institutional continuities after 1688 and that the degree of insecurity, at least of ownership rights, remained quite high, since many institutional tools for effective oversight of the public finances were unavailable to public creditors in England until the 19th century.

This does not mean that security of ownership or transfer rights for creditors of the English state underwent no change over time, and hence can have made no contribution to economic growth. Analysis of interest rates on English government borrowing suggests that property rights for owners of capital developed continuously across the early modern period rather than shifting from insecurity to security suddenly at any particular date. In conjunction with their previously discussed claims, North and Weingast (1989) further asserted that 1688 saw a sharp discontinuity, with a sudden shift from insecure to secure property rights for government creditors causing a sharp decline in the interest rate which the British government had to pay to borrow funds. But Stasavage (2002) tracked interest rates on English government debt in the second half of the 17th century and the first half of the 18th, and concluded that security of private property rights for state creditors was not irrevocably established in 1688 but instead fluctuated between then and 1740. It was particularly violently affected by which political party controlled ministerial posts and the two chambers of parliament. Political events rather than institutional changes most strongly influenced investors' willingness to commit capital to the British state (Stasavage, 2002). Sussman and Yafeh (2006), too, found that interest rates on British government debt did not show any discontinuity after the Glorious Revolution, instead remaining high and volatile for another forty years. The evidence shows neither a sudden switch from insecurity to security around 1688 nor complete stasis between the medieval period and the 19th century.

8.7.3 Secure Property Rights for Taxpayers

Similar issues arise with the third type of property rights which are supposed to have become more secure in England in 1688 and to have contributed to that country's subsequent economic success. Before 1688, it is said, the English Crown frequently engaged in confiscating its subjects' wealth via taxation; through these unconstrained fiscal powers, it is claimed, the sovereign controlled a large fraction of the resources in the English economy and reduced the security with which his subjects could use the remainder (North and Weingast, 1989; Acemoglu et al. 2005, p. 393). The Glorious Revolution of 1688, according to this view, limited the right of the state to demand property from individuals for the first time in any society in history.

These claims sit uneasily, however, with the evidence that after 1688 the British state increased its capacity to raise revenue from individuals through taxation (Harris, 2004). The Bill of Rights promulgated in 1689 made the taxing power of the British state conditional on parliamentary approval, but did not limit Parliament's powers of taxation and did not require any representation or consent from the vast majority of taxpayers who were not represented in Parliament. The landed, financial, and commercial groups in society were over-represented in the British Parliament, while the vast mass of ordinary taxpayers were either under-represented or had no vote at all.

State revenues from taxation and borrowing in England greatly increased between 1689 and 1815, both in absolute terms and as a share of national income (Mathias and O'Brien, 1976, 1978; O'Brien, 1988). Fortunately, state extraction only began to increase in England at the point at which per capita incomes and economic growth had already risen to quite high levels (O'Brien, 1988, pp. 23–4; Brewer, 1989). However, between 1689 and 1815, real gross national product increased by a factor of 3 while real peacetime taxation rose by a factor of 15 (O'Brien, 2001, pp. 8, 10). This huge increase in government control over national resources after 1688 casts serious doubt on the view that 1688 marked an improvement in the security of ownership rights of British taxpayers. Even between 1603 and 1641, when the early Stuart monarchs were trying to introduce continental-style absolutism to England, total government expenditure was a maximum of 1.2–2.4% of national income; after 1688 the state rapidly increased its share of national income to 8–10% (McCloskey, 2010, pp. 318–19). The percentage of English national income over which individuals as opposed to the state enjoyed secure private property rights—whether in terms of ownership or in terms of use—declined after 1688.

In principle, this enhanced state capacity might have supported activities, such as provision of public goods, that indirectly enhanced private property rights or benefited economic growth in other ways. The British state's increased ability to raise funds after 1688 certainly enabled it to undertake a number of new activities. However, these probably did not benefit economic growth. One of the first things the new English king did after public finances were placed on a stable footing in 1688 was to use them to launch a major war against France. This was not a mere blip but the beginning of a long-term trend. The vast majority of English state expenditures after 1688 were not spent on civil government, in the sense of infrastructure, education, or other public goods that might have benefited long-term economic growth. Rather, state expenditures were predominantly allocated to military purposes or to servicing the state debt, which was itself incurred primarily for military purposes (O'Brien, 1988, 2001).

This military spending was not beneficial for the economy. As Williamson (1984, p. 689) shows, British economic growth was modest between 1760 and 1820, both relative to its subsequent performance and relative to modern developing economies, because "Britain tried to do two things at once—industrialize and fight expensive wars, and she simply did not have the resources to do both." Although the precise size of the impact of war on the British economy in the 18th century is debated, most agree that it was negative and non-negligible (Williamson, 1987; Crafts, 1987; Mokyr, 1987). The increased ability of the English state to borrow and tax during the 18th century thus probably did not favor economic growth: the English economy grew in spite of rising state spending, not because of it. Again, however, the evidence shows neither a sudden switch from insecurity to security of taxpayers' ownership, use or transfer rights at any specific date, nor complete stasis between the medieval period and the 19th century.

8.7.4 Security of Property Rights: Analytical Challenges and Open Questions

What do these findings imply for economic growth? The historical findings support neither of the two views prevalent in the growth literature about the relationship between economic growth and security of private property rights, whether these are defined in terms of ownership, use, or transfer. Economic history shows that secure rights of ownership, use, and transfer for landholders, lenders, and taxpayers did not emerge suddenly, recently, or in a single precociously advanced economy from which they subsequently diffused to other backward ones. Secure rights to own, use, and transfer land, capital, and other assets emerged gradually in a large number of European societies over half a millennium or more. None of these societies guaranteed individuals perfectly secure rights of ownership, use, or transfer over property, but none of them lacked such security altogether. Rights of ownership, use, and transfer of private property in most societies in medieval and early modern Europe were neither perfectly insecure nor perfectly secure, but rather changed incrementally over very long periods of time. Economic growth cannot be ascribed to a sudden switch from insecure to secure rights of ownership, use, and transfer; but nor, as we saw in Lesson 4, was growth left wholly unaffected by the incremental changes that did take place in the property rights regime.

These empirical findings from history pose an analytical problem for economics. If rights of ownership, use, and transfer over private property were reasonably secure in England in 1200, but also changed between then and 1800, what do we actually mean by property rights being "secure" enough to lead to economic growth? All economists and historians would probably agree that a necessary condition for economic growth is some degree of security of ownership, in the sense of protection from seizure or taxation of the entirety of what private individuals own or can gain through exchange, investment, and innovation. Most would probably also agree that economic growth also requires some degree of security in the rights to use one's property, whether by investing in improving it or by devising more productive uses for it. And most would agree that economic growth requires some degree of security in the right to transfer assets to other people, whether by selling them, renting them out, or using them as collateral for loans. But which component of security of property matters most for growth? How much security? And how do we measure it?

Without more refined analytical categories than mere security, we are unlikely to achieve a better understanding of the contribution of property rights to growth. Even distinguishing between security of ownership, security of use, and security of unrestricted transfer only takes us so far. The empirical findings from history suggest two directions in which economics must develop its analytical tools for thinking about security of private property rights.

First, constraints on security of private property rights are multifaceted. Constraints on security of ownership rights include such variegated incursions as state confiscation, eminent domain, manorial ejection, rapacious taxation, failure to repay loans, inability to defend one's property title using the legal system, and many more. Constraints on security of use rights are even more varied, and include interlinkage with labor and product markets (e.g. under serfdom), collective usufruct rights, communal regulation of crop rotations, manorial prerogatives, and many more. Constraints on security of unrestricted transfer rights include conditionality of sales; bans on hypothecation; village citizenship requirements; noble entailment; familial redemption rights; limits on female inheritance and marital property; inheritance customs; and many more. These constraints on security of ownership, use, and transfer rights in private property do not necessarily all change at the same time or in the same direction. Furthermore, the intensity of these various constraints on security of private property is not necessarily perfectly correlated across societies. The historical evidence, for instance, suggests that early modern England had, by European standards, strong security of private use rights protecting owners against communal or manorial intervention, but weak security of private ownership and transfer rights for married women; the former type of security of private property right changed significantly during the 18th century, while the latter did not. Similar examples of complex combinations of security and insecurity of different components of property rights can be provided for every pre-modern European economy. Economics needs analytical tools for deciding which of the numerous observed constraints on how people could own, use, and transfer property should be employed as criteria for defining "security" of property rights, and tools for establishing which of these aspects of security were likely to be more or less important for economic growth.

The second need for analytical attention is created by the fact that property rights are only one component of the wider system of institutions in a society. Security of rights of ownership, use, and transfer can be enhanced by other components of the system—for instance, by contracting institutions, as we saw in Lesson 4. But security of rights of ownership, use, and transfer can also be *constrained* by yet other components of the system—for instance, by village communities or the manorial system. As we saw in Lesson 5, the historical evidence suggests that in all pre-industrial European economies, even the most advanced, generalized property rights were constrained by other, more particularized components of the institutional system. Economics therefore needs analytical tools for understanding the interaction between security of private property rights and other components of the broader institutional system.

8.8. LESSON 7. INSTITUTIONS ARE EMBEDDED IN A WIDER INSTITUTIONAL SYSTEM

An understandably widespread assumption is that a particular institution will affect growth similarly in all economies and time-periods. Once secure private property rights are present in an economy, for instance, it is tempting to assume that they will exert an effect on growth that does not depend on the wider environment. But the evidence from historical societies suggests that this assumption is incorrect. Each institution, rather, is embedded in a wider institutional system and is constrained by the other institutions in that system; institutional labels turn out to be approximations which mask important variations that matter for economic growth. While institutional systems are not yet well understood, it is clear that to grasp how these variations affect growth, we must take the rest of the institutional system into account, because the impact of any particular institution on growth is constrained by the entire system in which it is embedded.

This lesson is vividly illustrated by the historical findings about an institution some recent contributions to the growth literature have portrayed as especially important: the family. These contributions claim that early and successful economic growth in the West was favored by a specific family institution called the European Marriage Pattern (EMP), involving late female marriage, high female celibacy, and nuclear rather than extended families. But as we shall see, the apparent relevance to economic growth of historical findings on the institution of the family has been obscured by the failure to take the larger institutional context into account.

Theories of economic growth have increasingly focused on historical demography in recent years, as economists have begun to incorporate fertility decline and population growth rates into their explanations of long-run growth (Galor, 2005a,b; Acemoglu, 2009; Guinnane, 2011). Unified growth theory, in particular, regards falling fertility and slowing population growth as essential preconditions for economies to convert a greater proportion of the yields from factor accumulation and technological innovation into per capita income growth (Galor, 2005a,b, 2012). The central role played by population in recent growth theory raises the question of the determinants of demographic behavior and its relationship with economic growth over the long-term.

One recent approach to this question has sought to ascribe the transition to sustained economic growth in Europe before and during the Industrial Revolution to a specific family institution called the European Marriage Pattern (EMP), involving norms of late female marriage, high female celibacy, and nuclear rather than extended families. This unique family institution, it is claimed, lay behind the early modern divergence in economic growth rates between Europe and the rest of the world, between north-west Europe and the rest of the continent, and between England and everywhere else (Greif, 2006a; Greif and Tabellini, 2010; De Moor, 2008; De Moor and Van Zanden, 2010; Foreman-Peck, 2011; Voigtländer and Voth, 2006, 2010). The EMP is supposed to have favored economic growth by improving women's position, increasing human capital investment, adjusting population growth to economic trends, and sustaining beneficial cultural norms. If these claims were true, they would imply that people in poor economies would have to change very deeply rooted aspects of their private lives before they could enjoy the benefits of economic growth.

Historical demography, however, provides no supporting evidence for the view that the EMP (or any specific type of family institution) influenced economic growth. A metastudy of the historical demography literature (Dennison and Ogilvie, 2013) finds that the three key components of the EMP—late marriage, high celibacy, and nuclear families—were not invariably associated with each other. Where they were associated, they did not invariably lead to economic growth. Indeed, where the components of the EMP did coincide in their most extreme form (German-speaking and Scandinavian Europe), economic growth was slower and industrialization later than in societies such as England and the Netherlands where the EMP was less pronounced. The most rapidly growing European economy, that of England, moved further away from the EMP in the century before and during the Industrial Revolution. The idea that the EMP had a clear causal influence on economic growth is not supported by the evidence.

In each society where the EMP was prevalent, it was embedded in a wider institutional framework. But the wider institutional system differed greatly from one European economy to the next. These wider institutional frameworks, not the institution of the family in isolation from them, influenced whether women enjoyed a good economic position, human capital investment was high, population responded flexibly to economic signals, or specific cultural norms were enforced. It was the institutional system as a whole, not the family or any other institution in isolation, that decided whether an economy grew or stagnated.

This can be seen by examining the institutional determinants of women's position, which is widely regarded as an important contributory factor to economic growth in poor countries (Birdsall, 1988; Dasgupta, 1993; Ray, 1998; Mammen and Paxson, 2000; Ogilvie, 2003, 2004c; Doepke and Tertilt, 2011). Some recent contributions to the literature claim that the EMP contributed to European economic growth by improving women's economic status (De Moor, 2008; De Moor and Van Zanden, 2010; Foreman-Peck, 2011; Voigtländer and Voth, 2006, 2010). However, there is no evidence to support the proposition that women's economic status in early modern Europe was determined solely, or even predominantly, by the institution of the family as opposed to the wider institutional framework (Ogilvie, 2003, 2004b,c, 2013a; Dennison and Ogilvie, 2013). The many empirical studies of women's economic position in pre-modern Europe suggest that women had a good economic position in some societies with the EMP and a bad one in others. England and the Netherlands assigned women a better economic position than other European societies (see the survey in Ogilvie (2003), Ch. 7), and these countries had the most successful economies in early modern Europe. But England and the Netherlands were also distinctive in many other ways: their factor prices, resource endowments, geopolitical position, trade participation, parliaments, legal systems, financial arrangements, and early liberalization of manorial, communal, and corporative institutions, have all been adduced as causes of their early economic success (Mokyr, 1974; De Vries and Van der Woude, 1997; Van Zanden and Van Riel, 2004). There has been much debate about the origins of English and Dutch distinctiveness. It seems obvious, however, that to qualify for consideration, any plausible explanation must invoke factors confined largely to England and the Netherlands—rather than a factor such as the EMP, which England and the Netherlands shared with many other societies in Western, Nordic, Central, and Eastern-Central Europe whose economies grew slowly and industrialized late.

Outside the precociously advanced market economies of England and the Netherlands, women's economic status was much worse. This was not because of the EMP or any other type of family institution, but because of the wider institutional system in which the family was embedded. In Germany, Scandinavia, France, and many other regions, the EMP prevailed but women's participation in industrial and commercial occupations was restricted by guilds of craftsmen, retailers, and merchants (Wiesner, 1989, 2000; Ogilvie, 2003, 2004b,c, 2005d, 2013a; Hafter, 2007). In many regions of Switzerland, Germany, and France, as micro-studies have shown, the EMP prevailed but women's work, wages, property rights, and in some cases even their consumption choices, were limited by local communities—again, by corporative institutions (Wiesner, 1989; Wiesner-Hanks, 1996; Wiesner, 2000; Ogilvie, 2003, 2010, 2013a; Hafter, 2007). In Bohemia (the modern Czech Republic), also characterized by the EMP, female household-headship was low, daughters could not inherit, and communal institutions collaborated with manorial administrators to harass women working independently outside male-headed households (Ogilvie and Edwards, 2000; Ogilvie, 2001, 2005a,b; Velková, 2012; Klein and Ogilvie, 2013). Whether women enjoyed economic autonomy under the EMP (or any type of family institution) depended on the balance of power among other institutions. Strong guilds which succeeded in excluding women from industrial and commercial activities and training existed both in northern Italy (in the absence of the EMP) and in German-speaking Central Europe (in its presence). Much weaker guilds which increasingly failed to exclude women from training and skilled work prevailed both in Eastern Europe (in the absence of the EMP) and in England and the Netherlands (in its presence) (Ogilvie, 2003, 2004b,c, 2005d, 2007b). Other corporative institutions such as village communities were extremely strong both in Russia (non-EMP) and in Germany (EMP) (Ogilvie, 1997, 2003, 2004b, 2006; Dennison and Ogilvie, 2007; Dennison, 2011). Corporative institutions played a central role in lowering women's economic status but show no systematic relationship with the EMP or any other family institution.

The importance of the wider institutional system, as opposed to the institution of the family in isolation, also emerges when we examine human capital investment. The EMP, it is argued, involved lengthy life-cycle phases during which young people were working outside the household, giving them the opportunity and incentive to invest in their human capital. The lower fertility resulting from late marriage and high lifetime celibacy is also claimed to have contributed to a shift from a high quantity of poorly educated offspring to a lower quantity of more highly educated ones, thus improving the quality of their human capital (Foreman-Peck, 2011). But parents will only invest in their offspring's education (as opposed to buying it as a consumption good) if such investment promises a positive return. There are two mechanisms by which this incentive can operate. First, parents may expect to share the returns from their offspring's education via transfers from offspring in adulthood. But this runs counter to a basic characteristic of the EMP, namely that the net intergenerational wealth flow runs from parents to children: offspring leave home early to work in other households, migrate to other localities, form independent households upon marriage, do not reside as adults in the same household (or even the same locality) as their parents, and seldom remit earnings to the parental generation (Caldwell, 1976, 1982). A family system with these characteristics creates disincentives for parents to invest in their offspring's human capital since they cannot expect to share returns when offspring reach adulthood.

The second mechanism by which parents may be motivated to invest in their offspring's education (as opposed to purchasing it as a consumption good) is altruism: their offspring's future well-being increases parents' own well-being. But this incentive will only operate if skilled jobs are open to all members of society. Parents will invest in girls' education only if females are able to take work that requires skills, instead of being restricted to activities which rely on learning-by-doing rather than formal training. Even for boys' education, skilled occupations must be open to all rather than being restricted to members of specific groups. But access to skilled occupations in pre-industrial Europe did not depend solely, or even systematically, on the institution of the family. Rather, it depended on the wider framework of institutions regulating labor markets: craft guilds, merchant associations, urban privileges, village communities, and manorial regulations. Women were allowed access to skilled jobs (e.g. in crafts or commerce) only in some societies with the EMP, specifically the Netherlands and England, and even then not without restrictions (Van Nederveen Meerkerk, 2006a,b, 2010; Van den Heuvel, 2007, 2008; Van der Heijden et al. 2011). In other EMP societies, such as Germany, Scandinavia, and France, craft guilds excluded females (and many "outsider" males) from skilled industrial work, and guilds of merchants and retailers restricted their participation in commerce (Wiesner, 1989; Wiesner-Hanks, 1996; Wiesner, 2000; Hafter, 2007; Ogilvie, 2003; Ogilvie et al. 2011). This reduced the incentive to invest in girls' education, although better-off parents still purchased it as a consumption good. The EMP by itself cannot have been crucial in creating incentives for female education since the EMP existed both in societies where women were more often permitted to do skilled work and those where coercive institutions excluded them. Rather, what decided whether females learned vocational skills was the strength or weakness of barriers to entry imposed by corporative institutions seeking economic rents for insiders by restricting low-cost competitors such as women (Ogilvie, 1986, 2003; Wiesner-Hanks, 1996; Wiesner, 2000; Sanderson, 1996).

Human capital indicators for European economies in the 18th and 19th centuries show that education levels varied hugely across societies with the EMP (Lindert, 2004, pp. 91–2; A'Hearn et al. 2009, p. 801; Reis, 2005, p. 203; Dennison and Ogilvie, 2013, esp. Table 4). This is not surprising, since the family was not the only, or the main institution that affected education levels. Schooling, literacy, and numeracy in early modern Europe were more strongly influenced by other institutions: the market, the church, the state, the local community, the occupational guild (Ogilvie, 1986, 2003; Wiesner-Hanks, 1996; Wiesner, 2000). These non-familial institutions show no significant correlation with the prevalence of the EMP. In some societies, such as Germany and Scandinavia, the church allied with the state and the local community to impose compulsory schooling on children of both sexes, monitor compliance, and penalize violations, leading to very high education levels (Ogilvie, 1986, 2003; Johansson, 1977, 2009). In other societies, such as England, such institutional pressures were absent, leading to much lower levels of school enrolment and literacy. Numeracy was typically learned, to some degree at least, informally in response to market demand in commercialized economies, explaining why England, with its mediocre school enrolment and literacy, had numeracy levels similar to more institutionally regulated societies such as Germany or Scandinavia (A'Hearn et al. 2009).

Historically, human capital investment shows no evidence of having positively affected economic growth in Europe before the late 19th century. England grew fast in the early modern period and industrialized before any other society, yet schooling and literacy stagnated there during the 18th century and were not high by European standards until well into the 19th century. Economic historians who disagree on almost all other issues concur that human capital investment was not important in the English Industrial Revolution (Mokyr, 2009; Allen, 2003). Conversely, other European societies had outstandingly good educational indicators but slow economic growth. The Netherlands had high levels of school enrolment, literacy, and numeracy, but after the end of the Dutch Golden Age in 1670 its economy stagnated and it industrialized very late. German territories had much higher school enrolment and literacy than England and even the Low Countries, but stagnated throughout the early modern period and did not industrialize until after c. 1840. A similar pattern is found in Lutheran Scandinavia, with high school enrolment and literacy rates, but slow growth and late industrialization (Dennison and Ogilvie, 2013).

The available evidence strongly suggests, then, that human capital neither was affected by the EMP nor played any causal role in economic growth before the late 19th century. In many parts of central and northern Europe, school attendance and literacy were imposed and enforced by churches, rulers, landlords, communal officials, and occupational guilds.

These organizations used their institutional powers to impose "social disciplining" on ordinary people for the benefit of elite interests (Ogilvie, 2006). In many societies, education levels were not chosen by ordinary people themselves, for economic or other reasons, but rather imposed on them by elites to serve their own interests, and thus depended on the powers these elites enjoyed via the wider institutional system: the church, the state, serfdom, communities, guilds. This wider institutional system, not the EMP, explains the absence of a systematic relationship between educational indicators and economic growth in Europe before the late 19th century.

In the recent literature on the EMP, yet another pathway has been suggested as a link between the EMP and European economic growth. It has been claimed that England had a particularly extreme version of the EMP, and that the resulting late marriage and high lifetime celibacy ensured that English population growth was uniquely responsive to economic signals. This is supposed to have ensured that in England economic surpluses resulted in capital accumulation, enabling productivity-enhancing innovation and fuelling faster economic growth than in France or China (Voigtländer and Voth, 2006, 2010). However, the historical demography literature does not support the idea that England had an extreme version of the EMP (Dennison and Ogilvie, 2013). Nor does the evidence show higher demographic responsiveness to economic trends in England than elsewhere. An econometric study of French demographic behavior, for instance, found that "at no time between 1670 and 1830 were marriages less responsive to economic conditions in France than in England" and concluded that the origins of the contrast between French and English growth performance "are not to be found in difference of demographic behavior" (Weir, 1984, pp. 43–4). In Germany, too, the elasticity of fertility with respect to economic signals was higher than in England (though slightly lower than in France) throughout the 18th century (Guinnane and Ogilvie, 2008, pp. 23–7). Among the nine European economies studied by Galloway (1988), the responsiveness of fertility to changes in grain prices was weaker in England than in societies where economic growth was much slower (Austria, Sweden, Belgium, the Netherlands) or where the EMP did not prevail (Tuscany). In 18th-century China, where family institutions were also very different from the EMP, recent studies also show fertility rates responding to changes in grain prices (Wang et al. 2010; Campbell and Lee, 2010). For England itself, several analyses have found that preventive checks on population growth weakened or disappeared by c. 1750, indicating that fertility became less responsive to economic signals in England at the precise period when economic growth began to accelerate and to diverge most from growth in other Western European economies (Galloway, 1988; Nicolini, 2007; Crafts and Mills, 2009). Evidence for various European economies suggests that these findings can be explained at least partly in terms of interactions between the family and other components of the institutional system, especially village communities, privileged urban corporations, occupational guilds, and serfdom (Ehmer, 1991; Ogilvie, 1995; Guinnane and Ogilvie, 2008, 2013).

The embeddedness of particular institutions in the broader institutional system also emerges from studying cultural attitudes associated with the EMP. It has been suggested that the EMP caused nuclear families to predominate over wider kinship groups, thereby fostering growth-inducing attitudes, specifically trust beyond the familial group and gender equality. These cultural norms are supposed to have been further propagated by medieval Catholic religious ideology, which is supposed to have compared favorably in this respect to the ideological norms disseminated by non-Christian religions such as Islam (Greif, 2006a; Greif and Tabellini, 2010; De Moor, 2008; De Moor and Van Zanden, 2010). However, these are difficult claims to substantiate empirically. A number of scholars have found that religious attitudes to family and gender issues varied greatly across medieval Catholic Europe, and that this was because they were shaped by a broader framework of social institutions that differed greatly from one Catholic, European society to the next (Biller, 2001; Bonfield, 2001; Donahue, 1983, 2008; Dennison and Ogilvie, 2013). Demographic behavior and family structure also varied enormously across medieval Catholic Europe, with nuclear families dominant in some societies but extended families more important in others, including in strongly Catholic societies such as Italy and Iberia (Smith, 1981a,b; Pérez Moreda, 1997; Reher, 1998a,b; Sonnino, 1997; Micheletto, 2011). It is difficult, therefore, to find empirical support for the notion that the EMP sustained distinctive cultural norms, whether about non-familial trust or gender issues. The widely variegated distribution of European family institutions is not consistently associated with any distinctive set of cultural attitudes, and there is no evidence that such attitudes had a causal effect on European economic growth.

The idea, then, that the emergence of sustained economic growth in early modern Europe was caused by any particular type of family institution is not supported by the historical evidence and, in fact, is refuted by much of it. Whether a society with any given family institution experienced economic growth depended on overall characteristics of its economy and institutional system. In early modern England, the EMP existed within a framework of well-defined, private, transferable and (in most senses) secure property rights; well-functioning factor and product markets; and relatively few particularized institutions constraining female (or male) economic autonomy; economic growth was usually positive and ultimately spectacular. In the early modern Netherlands, the EMP initially existed in a similar framework of property rights, well-functioning markets, and successful economic growth; but after c. 1670 the Dutch economy stagnated and industrialization came late, for reasons that are still vigorously debated but are believed to have included a resurgence of particularized institutional privileges (Mokyr, 1974, 1980; De Vries and Van der Woude, 1997; Van den Heuvel and Ogilvie, 2013). In German-speaking Central Europe, Scandinavia, and the Czech lands, the EMP existed in a more coercive framework of mobility restrictions (including, in some areas, serfdom) and corporative barriers to entry in labor markets (for most women and many men); economic growth remained slow until these institutional obstacles were removed (Ogilvie, 1997, 2003; Dennison and Ogilvie, 2013).

Research in historical demography finds that the institution of the family was interlinked with the wider institutional system in multiple ways (Laslett, 1988; Ehmer, 1991; Solar, 1995; Guinnane and Ogilvie, 2008, 2013). It was these complex interactions among different institutions within an over-arching system, not any single institution in isolation, that affected economic growth itself, as well as influencing potential contributory factors such as women's status, human capital investment, demographic responsiveness, and—to the limited extent that these are empirically observable—cultural attitudes. Current scholarship suggests that the EMP may have required a social framework of strong non-familial institutions that could substitute for familial labor, insurance and welfare which small, nuclear-family households could not provide, and to which large numbers of unmarried individuals did not have access (Laslett, 1988; Solar, 1995; Dennison and Ogilvie, 2013). However, it was not inevitable that this wider framework should be made up of institutions that also happened to benefit economic growth, such as generalized private property rights, well-functioning markets, or impartial legal systems. Instead, this wider framework could as easily have been—and in many cases actually was—made up of particularized institutions with more malign growth effects, including serfdom, guilds, communities, religious bodies, and absolutist states (Ehmer, 1991; Ogilvie, 1995, 2003; Guinnane and Ogilvie, 2008, 2013; Dennison and Ogilvie, 2013). Future research must place at the center of its analysis the wider institutional system that constrained both demographic and economic decisions during European economic growth. No specific type of family institution in isolation can be regarded as necessary, let alone sufficient, for economic growth.

These findings make clear that a specific institution that matters for economic growth will often not operate similarly across different societies and time-periods. Private property rights, for instance, are embedded in broader institutional systems that differ greatly across societies, with the result that they will not affect growth identically everywhere. If they are not embedded in an institutional system containing, for example, accessible and enforceable contracting institutions, they will fail to unleash economic growth, as we saw in Lesson 4. Likewise, the same family institution can exist in different societies characterized by widely differing institutional systems, and will consequently affect economic growth in widely differing ways. The evidence we have shows that the growth effects of any individual institution are constrained by other parts of the institutional system differently in different societies, and that it is the entire institutional system, not any single institution in isolation, that is important for economic growth.

While it is understandable that economists should wish to simplify the analysis of institutions in order to try to get at their essential features, it is important to remember the remark attributed to Einstein to the effect that "everything should be made as simple as possible, but not simpler."[5] While the embeddedness of particular institutions in larger systems undoubtedly adds greatly to the complexity of the analytical (and especially the empirical) task, it seems to be an undeniable fact we cannot simplify away. Institutions just are not easily separable from their contexts and identifiable under the traditional or common-sense headings of conventional labels, but rather have to be analyzed as part of an entire institutional system.

8.9. LESSON 8: DISTRIBUTIONAL CONFLICTS ARE CENTRAL

We have seen in Lessons 1 through 7 that many economists concerned with growth ascribe a major causal role to institutions, whose roots they trace far back in history. But there are also many who challenge the very idea of an institutional system favorable to growth, independent of geographical or cultural context. Some regard institutions essentially as superstructure, with other variables, such as geographical resource endowments or cultural attitudes, as more fundamental causes of economic growth which bring institutions in their wake (e.g. Sachs, 2003). Others hold that a society always has the institutions that are efficient given its endowments, technology, or cultural attitudes (e.g. North and Thomas, 1970, 1973; Greif, 2006c). There are even those who regard both institutions and growth as fundamentally caused by stochastic shocks amplified by subsequent path dependency (e.g. Crafts, 1977; Crafts et al. 1989).

The geographical and efficiency approaches are particularly prominent in the literature on institutions and growth in historical perspective. A number of scholars have sought to explain the historical development of institutions and economic growth in terms of geography and resource endowments. Thus Diamond (1997) explains the last nine thousand years of economic growth and human institutions in terms of geographical characteristics. Pomeranz (2000) accounts for economic divergence between Europe and China since 1750 through coal deposits, disease, ecology, and proximity to exploitable "peripheries." Sachs (2001) argues that tardy growth in modern LDCs derives from their location in tropical zones where agricultural techniques are inherently less productive and the disease burden higher. As we shall see shortly, Domar (1970) explains the economic divergence between Eastern and Western Europe from the medieval period to the 19th century, and serfdom as the central institutional manifestation of that divergence, in terms of the supply of land relative to the supply of labor, which was in turn determined by exogenously occurring population growth and land conquests.

[5] See Calaprice (2011, pp. 384–5, 475), who also reports the following less simple (but probably more accurate) variant of this idea, from Einstein's Herbert Spencer Lecture, "On the Method of Theoretical Physics," delivered in Oxford on 10 June 1933: "It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience."

The efficiency view of institutions and growth is also widespread among economists, as we have seen in earlier lessons. According to this view, the task of the economic historian is not to find out which institutions are most conducive to growth, but to discover how apparently inefficient and growth-discouraging institutions in past societies were actually efficient in their particular natural or cultural context, whatever the appearances. In this spirit, not only the historical institutions we have met in the lessons above, but many others, have been reinterpreted by one economic historian or another in efficiency terms as a beneficial solution to one or more obstacles to possible transactions—merchant guilds (Greif et al. 1994; Greif, 2006c), craft guilds (Hickson and Thompson, 1991; Epstein, 1998; Zanden, 2009), village communities (McCloskey, 1976, 1991; Townsend, 1993; Richardson, 2005), serfdom (North and Thomas, 1970, 1973; Fenoaltea, 1975a,b), the noble feud (Volckart, 2004), vigilante justice (discussed in Little and Sheffield, 1983; Hine, 1998), and lynching (surveyed in Carrigan, 2004), among many others.

If it were true that institutions were always responses to natural endowments or efficient solutions to economic problems, then they would not matter for growth. It is their significance for growth, however, that motivates economists to understand why institutions arise and why they change.

Fortunately, there is an alternative to viewing institutions either as superstructures of more fundamental natural forces, or as efficient responses to such forces. According to this alternative approach, the institutions of a society result partly or wholly from conflicts over distribution (see Knight, 1995; Acemoglu et al. 2005; Ogilvie, 2007b). This conflict view is based on the idea that institutions affect not just the efficiency of an economy but also how its resources are distributed. That is, institutions affect both the size of the total economic pie and who gets how big a slice. Most people in the economy might well want the pie to be as big as possible—hence the assumption of the efficiency theorists. But people will typically disagree about how to share out the slices. Since institutions affect not only the size of the pie (through influencing efficiency) but also the distribution of the slices (through apportioning the output), people typically disagree about which institutions are best. This causes conflict. Some people strive to maintain particular institutions, others merely cooperate, others quietly sabotage them, and still others resist. Individuals struggle over institutions, but so do groups—and some groups organize for that very purpose. Which institution (or system of institutions) results from this conflict will be affected not just by its efficiency but by its distributional implications for the most powerful individuals and groups (Knight, 1995; Acemoglu et al. 2005; Ogilvie, 2007b).

Efficiency theories do sometimes mention that institutions result from conflict. But they seldom incorporate conflict into their explanations. Instead, conflict remains an incidental by-product of institutions portrayed as primarily existing to enhance efficiency. Thus, for instance, North often mentions distributional effects of institutions in his early work, but explains their rise and evolution in terms of economic efficiency (North and Thomas, 1970, 1973; North, 1981). Greif (2006c) also sometimes acknowledges that institutions can have distributional effects, but analyzes the specific institutions he selects—the Maghribi traders' coalition, the European merchant guild—in terms of their efficiency in encouraging medieval commerce and their compatibility with prevailing cultural beliefs. Insofar as rent-seeking is acknowledged, it is characterized as efficient, on the grounds that "monopoly rights generated a stream of rents that depended on the support of other members and so served as a bond, allowing members to commit themselves to collective action" (Greif et al. 1994, p. 749, 758).

Yet a conflict approach which incorporates the distributional activities of institutions into its analysis without assuming such activities to be efficient can explain many facts about pre-modern institutions that efficiency views cannot. One of the frequently cited justifications of the efficiency view is the longevity of the particular institutions it seeks to rediagnose as efficient. If they were not efficient, the challenge goes, why did they last for centuries? Wouldn't they have disappeared much sooner if they had been so bad for output and growth? The conflict view has a powerful explanation for the longevity of institutions that have historically inflicted considerable damage on the growth of the economies in which they prevailed.

For instance, the conflict view would agree that there is a good economic reason why, as we saw in Lesson 3, guild-like merchant associations existed so widely from the 12th to—in some societies—the 19th century. But this reason was not that they increased aggregate output by guaranteeing property rights or contract enforcement. Rather, they limited competition and reduced exchange by excluding craftsmen, peasants, women, Jews, foreigners, and the urban proletariat from most profitable branches of commerce. Merchant guilds and associations were so widespread and so tenacious not because they efficiently solved economic problems, making everyone better off, but because they efficiently distributed resources to a powerful urban elite, with side benefits for rulers (Lindberg, 2009, 2010; Ogilvie, 2011). This rent-seeking agreement between political authorities and economic interest-groups was explicitly acknowledged by contemporaries, as in 1736 when the ruler of the German state of Württemberg described the merchant guild that legally monopolized the national worsted textile proto-industry as "a substantial national treasure" and extended its commercial privileges at the expense of thousands of impoverished weavers and spinners on the grounds that "especially on the occasion of the recent French invasion threat and the military taxes that were supposed to be raised, it became apparent that no just opportunity should be lost to hold out a helping hand to [this merchant guild] in all just matters as much as possible." (Quoted in Troeltsch, 1897, p. 84.)

The conflict approach would also hold that there is a good economic explanation for why craft guilds were widespread in Europe for many centuries. But this is not that they were good for the whole economy. Empirical micro-studies of guilds' actual activities—as opposed to the rhetorical advocacy of their benefits in literature and legislation—show

how they underpaid employees; overcharged customers; stifled competition; excluded women and Jews; and blocked innovation. Guilds were widespread not because were good for everyone, but because they benefited well-organized interest groups. They made aggregate economic output smaller, but dished out large shares of it to established male masters, with fiscal and regulatory side-benefits to town governments and rulers (Ogilvie, 1997, 2003, 2004a,b,c; 2005d; 2007a; 2008).

The conflict view would also agree that there is a good economic explanation for the tenacity of strong peasant communes, which existed in large parts of Europe for centuries, as we saw in Lesson 4. But this is not that they were efficient for the whole economy. Their regulation of land-markets, migration, technology, settlement, and women's work often hindered the allocation of resources, in ways so innumerable that village micro-studies are still uncovering their true extent and implications. This not only diminished aggregate output but brutally narrowed the consumption and production options of poorer social strata, women, minorities, and migrants. Strong communes persisted not because they efficiently maximized the aggregate output of the entire economy, but because they distributed large shares of a much more limited output to village elites (rich peasants, male household heads), with fiscal, military, and regulatory side-benefits to rulers and landlords (Melton, 1990; Ogilvie, 1997, 2005a,b, 2007b; Dennison and Ogilvie, 2007; Dennison, 2011).

Finally, a conflict approach would agree that there is a good economic reason for the long existence of serfdom; but this is not that it efficiently solved market imperfections in public goods, agricultural innovation, or investment. Rather, serfdom created an economy of privileges that hindered efficient resource allocation in land, labor, capital, and output markets. But although serfdom was profoundly ineffective at increasing aggregate output, it was highly effective at distributing large shares to landlords, with fiscal and military side-benefits to rulers and economic privileges for serf elites.

The example of serfdom, in fact, provides an excellent illustration of the superiority of the conflict view of institutions to alternative approaches which explain institutions in terms of geographical resource endowments or economic efficiency. Indeed, economists concerned with institutions and growth have repeatedly turned their attention to serfdom, precisely because it played such a central role in the divergent growth performance of European economies between the Middle Ages and the 19th century. Serfdom set the institutional rules for agriculture, the most important sector of the medieval economy (Campbell, 2000). In the late Middle Ages, serfdom broke down in some European economies (mainly in the west), but intensified or emerged newly in others (mainly in the east), although the chronology and manifestation of this development varied enormously within both zones of the continent (for recent surveys see Cerman, 2013; Ogilvie, 2013b). But through this entire period agriculture remained by far the most important sector even of the most highly developed economies in Europe: it consumed most land, labor and capital; it produced most food and raw materials; and for industry or commerce to grow, inputs and outputs had to be released from farming (DeVries, 1976; Crafts, 1985; Ogilvie, 2000). The survival, breakdown, and intensity of serfdom in different European societies played a fundamental role in their divergent agricultural performance and hence their divergent growth record between the medieval period and the Industrial Revolution.

Because of its central role in long-term growth and stagnation, serfdom has been used as a test case for nearly every possible approach to institutions and growth—in terms of resource endowments (e.g. Postan, 1966; Domar, 1970), economic efficiency (e.g. North and Thomas, 1970, 1973; Fenoaltea, 1975a,b), and distributional conflicts (e.g. Brenner, 1976; Acemoglu and Wolitzky, 2011). The decline of serfdom is widely regarded as a major contributor to the growth of agriculture in Western Europe and its political abolition in Central and Eastern Europe under the impact of the French Revolution is regarded as a major example of institutional effects on growth (Acemoglu et al. 2011). Yet serfdom was not monolithic, it was embedded in the institutional systems of different European economies in different ways, and its growth effects depended, as we shall see, on its interactions with other components of each institutional system. Serfdom therefore provides an excellent context for contrasting different approaches to institutions, illustrating the strengths of the conflict approach, and demonstrating the work that remains to be done in tracing how institutions affected growth in historical perspective.

8.9.1 Resource Endowments, Serfdom, and Growth

Serfdom was an institutional system which obliged a peasant to provide forced labor services to his landlord in exchange for being allowed to occupy land. A serf was legally tied to the landlord in a variety of ways, typically by being prohibited from migrating, marrying, practicing certain occupations, selling certain goods, participating in factor and product markets, or engaging in particular types of consumption without obtaining permission from his landlord. Serfdom was therefore a particularized institution (in the language suggested in Lesson 3) which affected economic growth by restricting access to factor and product markets, preventing allocation of resources to the highest-productivity uses, and creating poor incentives for investment in human capital, land improvements, and technological innovations.

Most economies in Europe were characterized by some version of serfdom between c. 800 and c. 1350. After that date, serfdom began gradually to decline in some societies, such as England, although it survived for longer in others, such as France and western Germany. In the 16th and 17th centuries, some parts of Eastern-Central and Eastern Europe where classic serfdom had either never existed or had declined, including Russia, the Czech lands, Slovakia, Poland, Hungary, and eastern German territories such as Prussia, experienced an intensification of manorial controls by landlords, which has been called the second serfdom. This system remained in force in these economies until its abolition, usually through state action, which occurred in different Central and Eastern European societies at different dates between c. 1760 and c. 1860.

One widely held view within economics is that serfdom was an institutional response to resource endowments, specifically to the relative supply of land and labor. This idea is based on a paper by Domar (1970) arguing that serfdom can be explained as a response to a high land-labor ratio. Labor scarcity created severe competition among employers (landlords) for laborers (peasants) to work their land. Moreover, the abundance of land meant that peasants had attractive options setting up as independent farmers and withdrawing their labor from landlords altogether. This created a strong incentive for landlords to organize an institution to prevent peasants from doing these things, by legally binding them to the estate, forbidding them from migrating to competing employers, and obliging them to deliver a certain quantity of forced labor on the landlord's farm (the demesne). Domar argued that this explains the rise of serfdom in 17th-century Russia: the land-labor ratio rose because of the Muscovite colonial conquests and landlords devised serfdom as a way of protecting their supply of scarce peasant labor.

However, there are many examples of economies in which the land-labor ratio was high, but there was neither serfdom nor slavery. The most striking counterexample to Domar's model of serfdom is Europe after the Black Death. This virulent pandemic greatly increased the land-labor ratio in most parts of Europe by killing off 30–60% of the population between 1348 and 1350. According to Domar's theory, this should have caused serfdom to intensify, or to come into being in societies in which it had not previously existed. However, this did not happen. Instead, many parts of Western Europe saw serfdom break down after the Black Death, and never reappear no matter what happened to the land-labor ratio.

The decline of serfdom in Western Europe after the Black Death had already stimulated Postan (1966) to propose his own theory of serfdom in terms of resource endowments. Postan's theory was diametrically opposed to that of Domar, since it argued that the rising land-labor ratio after the Black Death caused the decline of serfdom because it made landlords compete for peasants by offering better conditions. Postan had only put this forward as an account of the decline of serfdom in Western Europe after the Black Death, not as a general model of serfdom in all societies. Domar (1970) did regard himself as advancing a general model of serfdom in terms of relative resource endowments. But he knew enough about the historical findings to recognize that a high land-labor ratio only provided the incentive for landlords to organize institutions to prevent themselves from losing laborers. Whether they actually did so depended on whether they were able to organize politically, i.e. were powerful enough to coerce peasants and prevent other landlords from competing them away by offering them better conditions (e.g. the freedom to take economic and demographic decisions without landlord permission). So Domar's model is one in which serfdom arises from relative resource endowments plus the political power of different social groups—i.e. it is broadly consistent with the conflict model of serfdom which we shall discuss shortly.

8.9.2 Efficiency, Serfdom, and Growth

Despite the near unanimity among economists and economic historians that serfdom was harmful for growth,[6] it was one of the first institutions to be re-diagnosed as efficient. In the early 1970s, North and Thomas, (1970, 1971, 1973) proposed a model of the "rise of the western world," according to which serfdom was "an efficient solution to the existing problems" in medieval economies, a voluntary contract that committed peasants to provide labor services to lords in exchange for "the public good of protection and justice" (1973, p. 21). North and Thomas explicitly stated that "serfdom in Western Europe was essentially not an exploitative arrangement ... [it] was essentially a contractual arrangement where labor services were exchanged for the public good of protection and justice" (1971, p. 778). The reason serfs had to be forced to render these payments was that protection and justice were non-excludable, so individual serfs had an incentive to free-ride. Serfs were protected from being exploited by the landlord as the monopoly supplier of protection, according to North and Thomas, by institutional rules (the "customs of the manor") and by the fact that they had a low-cost exit option (absconding from their lord). The reason serfs had to be forced to pay in the form of forced labor services rather than cash or kind was uncertainty (the lords could not know *ex ante* how much the serfs were able to produce), transaction costs (the costs incurred by a landlord in reaching a bargain with a large number of peasants), and absence of markets (so that cash or kind would be of no use to the landlord since there was nothing to purchase with them).

The implication of these efficiency theories and others (e.g. Fenoaltea, 1975a,b, 1984) was that serfdom was an efficient institution given the characteristics of the economies in which it occurred, and was therefore beneficial for economic growth until these characteristics changed. But there is little evidence for this. Protection and justice were, in fact, excludable. Protection was provided by the lord's manor house or castle from which serfs could be excluded if they did not pay. Furthermore, the lord's fortifications did not protect serfs against that large proportion of the random violence of medieval society which took the form of unpredictable raids. Justice was also excludable: manorial courts operated by the landlord or his officials could refuse to provide justice to anyone, could strip a serf of legal protection by outlawing him, and could charge court fees to cover the costs of judging legal conflicts. Further doubt is cast on the idea that serfdom was an efficient solution to the provision of justice by the fact that feasible alternatives did exist: the prince, the church, abbeys, and towns all provided law-courts, which offered alternatives to the manorial courts and often did not even acknowledge differences in serf status. Also, neither absconding nor the customs of the manor provided effective protection to serfs against monopolistic landlords. A strong landlord could simply ignore

[6] Revisionist views claiming that serfdom did not harm the economy have been proposed (most recently in Cerman, 2012, 2013), but do not hold up well to empirical scrutiny (see Briggs, 2013; Dennison, 2011, 2013; Guzowski, 2013; Klein, 2013; North, 2013; Ogilvie, 2013b; Rasmussen, 2013; Seppel, 2013).

custom, and many did. Furthermore, absconding was a costly option which required the serf to abandon land, possessions, family, and social capital.

An even more fundamental problem for the efficiency view of serfdom is that much of the insecurity and injustice against which serfs were being "protected" by their landlords was actually produced by feudal landlords themselves. Serfdom was thus much more like a protection racket in which the landlords, as the more powerful party, generated both the problem and the solution. Serfdom did not constitute a bundle of voluntary contracts which contributed to economic efficiency, but rather was a set of rent-seeking arrangements devoted to redistributing resources from peasants to landlords.[7] Moreover, North and Thomas are wrong in claiming that peasants had to pay in the form of labor rather than cash or kind because of absence of markets. Every serf society that has ever been observed had markets for goods as well as for factor inputs, as we shall discuss in greater detail shortly.

The findings for serfdom show clearly the dangers of trying to explain institutions purely as efficient solutions to economic problems. Serfdom, it is clear, also involved coercive power, and some of the problems to which it is supposed to have been a solution were themselves caused by the exercise of this power. This suggests that we cannot assume that any institution we observe, even if it survives for hundreds of years, did so because it was the efficient set of social rules for maximizing aggregate economic output. We have to investigate what effect it had on the distribution of this output (Acemoglu et al. 2005; Ogilvie, 2007b).

8.9.3 Distributional Conflicts, Serfdom, and Growth

A fundamental break from viewing serfdom as resulting from resource endowments or economic efficiency, and thus being neutral or beneficial for economic growth, came with the work of Brenner (1976). Brenner pointed out serious problems with the view that labor scarcity (e.g. in Europe after the Black Death) caused serfdom either to strengthen or to break down. Plague-induced labor scarcity changed the incentives of both serfs and landlords. Certainly, as North and Thomas had argued, labor scarcity increased serfs' incentives to use their increased bargaining position to break down serfdom. But it also increased landlords' incentives to intensify serfdom in order to secure their supply of scarce laborers (the Domar argument). In actual practice, the change in relative supplies

[7] North (1981, p. 131) later conceded that "carrying over the modern-day notion of contract to the serf-lord relationship is imposing a modern-day concept which is misleading. The serf was bound by his lord and his actions and movements were severely constrained by his status; no voluntary agreement was involved. Nevertheless, it is crucial to re-emphasize a key point of our analysis; namely, that it was the changing opportunity cost of lords and serfs at the margin which changed manorialism and eventually led to its demise." However, this does not address all the problems with his model, especially the excludability of the protection and justice services provided by landlords and the fact that landlords themselves generated much of the insecurity and justice they are supposed to have been protecting serfs against.

of land and labor after the Black Death saw serfdom develop in diametrically opposite directions in different European societies. In most Western European economies, serfdom broke down after the Black Death, albeit at different rates and times. In most parts of Eastern Europe, manorial powers survived the Black Death and greatly intensified under the second serfdom.

This was not because serfdom ceased to be efficient and to promote economic growth in the west but continued to be efficient and to promote growth in the east. Rather, which path an economy followed was "a question of power, indeed of force" (Brenner, 1976, p. 51). The outcome in each specific society was determined by the ability of both peasants and landlords to band together collectively with their fellows as well as to ally with the coercive power of the state. In Western Europe, the stronger central state that emerged toward the end of the medieval period pursued policies of "peasant protection" with the motivation of maintaining the peasantry's ability to pay taxes to the state rather than rents and labor services to landlords. In Eastern Europe, by contrast, the state allied with the landlords and enforced their controls over the peasantry in exchange for a share of the spoils. Brenner argued that serfdom was always an exploitative arrangement that redistributed resources from peasants to landlords. He also argued that this redistribution had harmful effects on economic performance: the effect of the second serfdom, in his view, was that "the possibility of …economic growth was destroyed and East Europe consigned to backwardness for centuries" (Brenner, 1976, p. 60).

Acemoglu and Wolitzky (2011) extended Brenner's perspective by proposing a model of labor coercion which sought to combine resource endowments and power. It placed the relative scarcity of labor and land at center stage, but formalized Brenner's point that labor scarcity can have two countervailing effects on serfdom, one intensifying it and one breaking it down. Their model suggests that labor scarcity, via its effect on the price of output and the returns to coercion, tended to intensify serfdom, as argued by Domar (1970). However, their model also suggests that labor scarcity, by improving the outside options of peasants, tended to weaken serfdom, as argued by Postan (1966) and North and Thomas (1971). Acemoglu and Wolitzky argue that what decided whether labor scarcity led serfdom to intensify or alternatively to decline was whether the value of output and the returns to coercion exceeded the value of the outside options of peasants. In Eastern Europe, they argue, missing markets meant that serfs had few external options, so the value of these options was surpassed by the returns to coercion; hence falling population in Eastern Europe intensified serfdom. In Western Europe, by contrast, the existence of markets gave serfs profitable outside options, which exceeded the value of the returns to coercion, so population decrease caused serfdom to decline.

This is a major advance over previous contributions, but leaves out what historical research shows about three important institutions which co-existed with serfdom and affected its operation: the state, the community, and the market. Regarding the state, as Acemoglu and Wolitzky themselves acknowledge (2011, pp. 569–71), their model treats each employer of serfs as an individual rather than recognizing that in practice serf landlords typically exercised coercion collectively and used this collective coercion (often enforced via the state) to regulate serfs' outside options. Although Acemoglu and Wolitzky contend that their argument still holds when the state is included, the fact remains that it fails to address the argument of Brenner (1976), according to which the strongest variable determining whether labor scarcity would strengthen or weaken serfdom was politics, specifically collective action by serfs and landlords and relations between each social group and the state.

Regarding the community, the Acemoglu and Wolitzky model treats each serf employee as an individual, rather than recognizing that in practice serfs formed communities which operated, at least in some ways, as institutional entities. The existence of communal institutions enabled serfs to engage in collective action toward both the landlord and the state. But the serf community also provided an entity with which landlords and the state could bargain in order to help them coerce individual serfs who sought to violate the constraints of serfdom, taxation, or conscription.

Regarding the market, Acemoglu and Wolitzky (2011) simply assume it to be missing in Eastern Europe, rather than recognizing that in practice Eastern European serfs did have access to, and participated in, markets for labor, capital, land, and output. The existence of these markets meant that serfs did have outside options, but the existence of market participation by serfs also offered landlords an additional and highly attractive source of rents. In practice, as we shall see, many landlords used their institutional powers to extract rents from their serfs' participation in markets, the profits from which contributed to their wealth, which they then invested partly in political action to sustain and intensify their own economic privileges under serfdom.

8.9.4 Serfdom and the Institutional System

Closer examination of the variables that created, sustained, and ultimately broke down serfdom strongly supports the view that distributional conflicts and political forces were central. But it also shows the importance of widening our focus beyond one institution in isolation to the wider institutional system. We cannot restrict our attention solely to serfdom, in the sense of the institutional rules governing relations between peasants and landlords. We must also analyze adjacent institutions, particularly those pointed out in the preceding section: the market, the community, and the state.

Markets were neither missing nor irrelevant to peasants' lives in serf societies, whether in medieval Western Europe or in early modern Eastern Europe. In the past few decades, micro-studies have revealed unambiguously that peasants in medieval and early modern serf societies made widespread use of markets. They used markets to buy and sell land (Cerman, 2008, 2012, 2013; Campbell, 2009), to offer and employ labor (Campbell, 2009; Dennison, 2011), to lend and borrow money (Briggs, 2004, 2009; Campbell, 2009; Ogilvie, 2001; Bolton, 2012), and to buy and sell food and craft products (Kaminski, 1975; Smith, 1996; Britnell, 1996; Cerman, 1996; Ogilvie, 2001; Bolton, 2012). Market participation can be widely observed among serfs not just in medieval England, but also in Germany, Switzerland, Austria, Italy, and France in the Middle Ages, as well as many regions of Eastern-Central and Eastern Europe under the early modern second serfdom, including Poland, Hungary, the Czech lands, and Russia (Kaminski, 1975; Dennison, 2011; Cerman, 2012; Ogilvie, 2012). This market participation was not limited to the richest serfs, but extended to all strata of serf society, including women, laborers, landless cottagers, and those subsisting at the edge of starvation (Kaminski, 1975; Cerman, 2012; Ogilvie, 2001, 2012).

Markets were present in serf economies, therefore, and offered attractive outside options for serfs. However, markets also offered attractive options for landlords. The result was that serfs' access to markets was often constrained by landlords' exercise of power in search of further rents. Thus serfs used markets widely to hire out their own labor, to employ the labor of others, and to buy and sell land (Topolski, 1974; Dennison, 2011; Klein, 2013; Ogilvie, 2001, 2005c, 2012, 2013b), although landlords used their powers under serfdom to intervene in both labor and land transactions to obtain rents or when they perceived a benefit to themselves (Harnisch, 1975; Ogilvie, 2001, 2005c, 2012; Dennison and Ogilvie, 2007; Velková, 2012). Serfs bought and sold agricultural and industrial output in markets, even though again landlords used their powers under serfdom to intervene in these markets by obliging serfs to buy licenses, pay arbitrary fees, offer their products first for sale to the landlord at dictated prices, or buy certain products solely from the landlord's own demesne operations (Cerman, 1996; Ogilvie, 2001, 2005c, 2012, 2013b; Klein, 2013). It was not, therefore, that markets were missing in serf societies, and that serfs thus lacked outside options, but rather that landlords intervened in these markets in such a way as to redistribute to themselves part of the profits from serfs' market participation. The interaction with markets entrenched serfdom more deeply and contributed to its longevity by further benefiting landlords at the expense of serfs.

Village communities also played a central role in the existence and survival of serfdom. Scholars such as Brenner (1976) had claimed that, under serfdom, village communities were stifled by landlord oppression. However, subsequent micro-studies have made clear that this was not the case (Wunder, 1978, 1996; Ogilvie, 2005a,b; Dennison and Ogilvie, 2007; Cerman, 2008, 2012). There was no question about the institutional capacity of village communities to operate as autonomous bodies under serfdom (Peters, 1995a,b, 1997; Wunder, 1995). Village communities organized direct resistance against attempts to intensify serfdom, and appealed to princely and urban jurisdictions against the landlord (Harnisch, 1972; Ogilvie, 2005a,b, 2012, 2013b). The strength of serfs' communal institutions and their ability to bargain with outside institutions, such as the state, other landlords, and towns, influenced the extent to which the landlord could intervene in their market transactions.

However, village communities played a complicated role in serfdom; they did not simply operate successfully and single-mindedly to protect serfs' interests. Serf communities were not fully independent of manorial intervention. The top village officers were often selected and appointed by the landlord (Harnisch, 1975; Peters, 1995a,b). Even the communal officials who were selected by serfs themselves were co-opted disproportionately by (and from) the top stratum of rich serfs. This oligarchy ran the village in its own interests and benefitted from communal autonomy (Melton, 1988; Rudert, 1995a,b; Hagen, 2002; Ogilvie, 2005a,b, 2012; Dennison and Ogilvie, 2007). Communal institutions typically implemented the choices of their most powerful members partly by limiting those of the least powerful—big farmers over laborers, men over women, established householders over unmarried youths, insiders over migrants (Ogilvie, 2005a,b, 2012, 2013b; Dennison and Ogilvie, 2007).

These characteristics of serf communities were not merely incidental. Rather, they were central components of how serfdom functioned. In normal times—i.e. except during legal conflicts or revolts of serfs against their landlords—community institutions carried out essential tasks that supported the manorial administration and ensured that serfdom functioned smoothly (Harnisch, 1986, 1989a,b; Dennison and Ogilvie, 2007; Ogilvie, 2012, 2013b). Landlords devolved to communal officers the organization of labor services and the collection of manorial dues (Peters, 1995a,b). They also deployed an elaborate community responsibility system which made the entire serf community responsible for the failings of any individual (Harnisch, 1989b; Peters, 1997). If a serf shirked on his labor services or vacated his farm without permission, his community was institutionally obliged to take up the slack. This created strong incentives for the community to report its delinquent or economically weak members to the manor; such communal reports lay behind many serf expulsions (Harnisch, 1989b). Collective responsibility for rendering forced labor and other payments to the landlord and the state also motivated communities to enforce the mobility restrictions of serfdom, and on many occasions one can observe communal officials pursuing absconding fellow serfs on behalf of the landlord (Peters, 1997). Conversely, staying in the good graces of the communal officials and the village oligarchy was essential if a serf hoped to secure a certificate that he had been a good farmer, which might in turn persuade the landlord to take a positive view of his applications regarding access to land or other resources (Harnisch, 1975; Hagen, 2002; Dennison and Ogilvie, 2007; Ogilvie, 2005a,b, 2012, 2013b). The most powerful stratum of serfs, who typically controlled the serf commune, was given very strong incentives to collaborate with landlord and state (Melton, 1988; Blaschke, 1991; Rudert, 1995a,b; Hagen, 2002; Ogilvie, 2005a,b,c; Dennison, 2011). The serf commune was thus an important component of the institutional system that helped to keep serfdom in being and intensified its negative growth effects while benefiting landlords (Ogilvie, 2005a,b, 2012, 2013b; Dennison and Ogilvie, 2007).

The state, finally, also affected the existence and survival of serfdom. Serfs were the state's main source of tax payments and army conscripts (Harnisch, 1989a,b; Seppel, 2013;

Ogilvie, 2013b). Often serfs were the sole source of tax payments, since the nobility typically used their dominance over parliamentary institutions to free themselves from taxation. This fact gave the state two countervailing incentives vis-à-vis serfdom. On the one hand, fiscal interests motivated the state to compete with landlords for serf money and labor (Hagen, 1989; Cerman, 2012). In a number of early modern Central and Eastern European serf societies, when lords demanded more forced labor, state courts granted redress to serfs, if only to safeguard serfs' fiscal capacities. On the other hand, the costs of maintaining state officials on the ground created strong incentives for the state to devolve tax-collection and conscription to local personnel, which meant collaborating with the landlord's administration and the whole regime of serfdom. The state thus competed with landlords for serf output but collaborated with landlords in the process of extracting that output (Hagen, 1989; Ogilvie, 2005c, 2013b; Cerman, 2008, 2012; Rasmussen, 2013; Seppel, 2013).

The state was also the gatekeeper of serfs' access to the legal system. In most societies under serfdom, the serfs' own village courts enjoyed the lower jurisdiction, which issued decisions on minor offences, neighborly conflicts, and land transactions (Kaak, 1991). But the higher jurisdiction over major offences was exercised in the first instance not by princes' courts but by landlords' courts (Cerman, 2012; Ogilvie, 2013b). Landlords typically secured this jurisdictional control from princes in return for fiscal and political favors, although to varying degrees in different serf societies (Kaak, 1991; Ogilvie, 2013b). In some European serf societies, such as Bohemia and Russia, landlords also successfully secured state legislation restricting serfs' right of appeal to princely courts (Ogilvie, 2005c; Dennison, 2011). But in many others, including Prussia, serfs retained (or were explicitly granted) the institutional entitlement to appeal against their landlords to state courts (Harnisch, 1975, 1989a,b; Hagen, 2002).

The legal balance of power between serfs and their landlords was influenced by the power of the ruler relative to the nobility in each polity (Harnisch, 1989a,b; Cerman, 2012; Ogilvie, 2013b). Where the ruler was weak compared to the nobles, the powers of landlords over serfs tended to be greater. But this did not mean that the state had no effect on serfdom in such societies: where the ruler depended heavily on noble support, he not only refrained from granting redress to serfs but positively supported landlords in most conflicts. Where the ruler lacked alternative sources of financial and political support and needed the support of landlords to obtain grants of taxes and payment of princely debts from the parliament, the ruler was more likely to acquiesce in most noble demands, including intensification of serfdom with state enforcement, as we saw in Lesson 2. Where the ruler had more plentiful alternative sources of revenue (e.g. from taxes on mining) and political support (e.g. from towns), he was able to resist the demands of the nobility (often expressed partly through a parliament) to a greater extent.

Probably the most important role the state played in serfdom was by legislating to shape, sustain, and ultimately abolish the entire system (Harnisch, 1986, 1994; Ogilvie, 2013b). Under serfdom, landlords responded to labor scarcity by using mobility restrictions to prevent serfs from voting with their feet to migrate to better conditions, and by cooperating with other lords to send fugitives back. Like any cartellistic arrangement, this landlord cartel was threatened by free-rider problems: lords collectively benefited from other lords' compliance but individually profited by violating the arrangement. This free-rider problem, as well as the transaction costs of coordinating enforcement across multiple manorial jurisdictions, gave landlords a strong incentive to seek support from the political authorities to enforce the institutional constraints of serfdom (Ogilvie, 2013b). In this way, the state played a fundamental role in sustaining the institution of serfdom.

However, the state also played a fundamental role in the ultimate abolition of serfdom, which took place at different dates in Eastern-Central and Eastern European societies in the course of the 18th and 19th centuries. In a number of serf societies, such as Prussia and Russia, the state reforms that abolished serfdom involved setting up a system of legal obligations requiring former serfs and their descendants to make redemption payments to their former landlords and their descendants as a form of recompense for losing the land, cash rents, and labour services that disappeared with the abolition of serfdom (Harnisch, 1986, 1994). In so doing, the state played a final, essential role in institutional change: mediating an enforceable agreement between serfs and landlords which credibly committed former serfs to reimburse former landlords for the losses caused by the institutional transformation.

The economic history of serfdom thus provides an excellent illustration of the importance to institutional change of dealing with the lack of what Acemoglu (2003) calls a "political Coase theorem." A party that holds (or obtains) some institutional power cannot make a credible commitment to bind its own future actions without an outside agency with the coercive capacity to enforce such a commitment. The absence of a political Coase theorem means that institutional changes that would make an entire economy better off are often blocked by the fact that it is difficult for the potential gainers from institutional reform to commit themselves to reimburse the losers after the latter have lost their institutional powers (Acemoglu, 2003; Acemoglu et al. 2005, p. 436; Ogilvie, 2007, pp. 666–7). The economic history of serfdom provides arguably the best example of this principle influencing the process of institutional change. In societies such as Russia and Prussia, serfdom was only abolished to the extent that the state was able to solve this problem of the missing "political Coase theorem" by mediating and enforcing a commitment for the gainers to compensate the losers. When Prussian serfdom was abolished in 1807, for instance, the state legislated that each former serf was to be allocated a parcel of land and freed from forced labor services, but was also legally obliged to compensate his landlord for the loss of this land and labor by making a series of redemption payments over a period of decades (Knapp, 1887; Harnisch, 1986, 1994). The state thus mediated and enforced a commitment that the serfs, as gainers from the abolition of serfdom, would compensate the landlords, as losers.

Economic history thus provides considerable support for the proposition that institutions are not just a response to resource endowments or efficient solutions to economic problems, in which case they would not matter for growth, but rather that they result partly or wholly from conflicts over distribution and hence have the potential to play a causal role in influencing whether an economy will grow or stagnate. But the growth literature, in pursuing a conflict view of institutions, has not yet made the best use of the historical evidence, and has placed excessive emphasis on high politics and top-down revolutions. The available evidence suggests, rather, that some of the most important institutions that harmed long-term growth in European history—institutions such as serfdom—arose from deep-seated and enduring distributional struggles among special-interest groups, carried out on a local level, far from the noise of parliamentary and ministerial struggles in national capitals, and often outside the formal political arena altogether. Conversely, societies that managed to minimize the influence of such groups over economic policies were the ones that gradually reduced the traction of particularized institutions and increased that of generalized ones, enabling their economies to achieve growth. Economic history thus strongly supports the centrality of socio-political conflict to developing the institutions that affect growth (for good or ill), but suggests that we must widen our definition of conflict from national politics as conventionally conceived, to include lower-level distributional conflicts and slow, gradual, non-revolutionary processes in the provinces.

8.10. ILLUSTRATION OF THE LESSONS: SERFDOM AND GROWTH

Having made it our main illustrative example for Lesson 8, we have now said enough about serfdom that we can further show how it exemplifies each of the eight lessons as well. Serfdom is of some independent interest in any case, as it governed the economic options of a majority of the population in agriculture, by far the largest economic sector in nearly every European economy throughout the medieval period and in many areas until the end of the 19th century. The decline of serfdom in Western Europe and intensification in Eastern Europe after the late medieval period certainly coincided with, and probably contributed to, the significant divergence in the growth of per capita income in the two parts of the continent between then and the 19th century (Ogilvie, 2013b). Understanding serfdom is therefore necessary if one wishes to understand divergence or convergence in the long-term growth performance of European societies between the Middle Ages and the Industrial Revolution.

First, serfdom shows clearly the importance of public-order institutions for economic growth, the argument advanced in Lesson 1. There is no empirical support for the idea that serfdom was an efficient private-order substitute for missing public-order institutions, whether in ensuring private property rights or in guaranteeing contract enforcement

(North and Thomas, 1970, 1971, 1973; Fenoaltea, 1975a,b). The decline of serfdom in Western Europe in the late medieval period was closely related to the unwillingness of the public authorities in those societies to provide support to the landlords in enforcing their institutional privileges over serfs. Conversely, the intensification of serfdom in Eastern-Central and Eastern European societies from the 16th century onwards was only possible because the state provided coercive support to landlords. Finally, the abolition of the second serfdom in Eastern European societies between the 1780s and the 1860s relied upon the public authorities to solve the problem of the missing political Coase theorem.

Second, serfdom shows clearly that a strong parliament, even one representing the interests of wealth holders, is not invariably beneficial for economic growth. In some serf societies, such as Poland, the parliament was extremely strong relative to the ruler. In all serf societies, the parliament represented wealth holders in the shape of the noble landed interests. The stronger the parliament in a serf society, the greater the ability of the landed nobility to hold the state to ransom, demanding that it provide state enforcement to back up the powers of landlords over the rural population, as a precondition for parliament to grant taxes or military support to the ruler. The history of European serfdom shows that economic growth depends not on whether a society has an institution that calls itself a parliament, exercises control over the executive, and represents wealth holders, but rather on the underlying institutions of that society, which determine how people obtain wealth, how wealth holders obtain parliamentary representation, and whether they then use that parliamentary representation to implement institutional rules that redistribute resources to themselves or alternatively ones that enable growth for the entire economy.

Third, serfdom illustrates the centrality of the distinction between generalized and particularized institutions. Serfdom was a completely particularized institution, in the sense that the rules it imposed and the services it provided depended completely on an individual's personal status and privileges as a serf or a non-serf. Access to land, labor, capital, and output under serfdom was not available or transferable to everyone impartially but rather depended upon the identity of the economic agent as a landlord, a freeman, or a serf. Furthermore, most forms of serfdom depended heavily on collaboration with a second particularized institution, that of the village community. The rules of the village community also operated in a particularized way, in the sense that ownership, use, and transfers of inputs and outputs depended upon an individual's personal status and privileges, e.g. as a village member rather than a migrant, a male householder rather than a woman or a dependent male, a substantial farmer rather than a landless laborer. However, in European serf societies, the completely particularized institutions of serfdom and the village community co-existed with the institutions of the state and the market, which were at least partly generalized. The precise balance between particularized and generalized institutions in serf societies determined how long serfdom survived, how much it constrained growth, as well as when and how it would be abolished.

Fourth, serfdom shows how property rights institutions and contracting institutions both matter, and are not separable. When people in serf societies traded, they simultaneously transferred property rights to another person and made a contract. Landlords intervened not just in property rights but also in contracts, by invalidating agreements in their own interests or those of clients to whom they had granted market privileges. Moreover, the abolition of serfdom in Eastern-Central and Eastern Europe often improved the security of private property rights in land, but did not see any improvement in agricultural growth. One reason was that in order for the growth benefits of improved property rights to be unleashed, it was also necessary for contracting institutions to improve so as to provide peasants with incentives to incur the costs and risks of investing in human capital, land improvements, and innovations. That is, the political authorities had to establish not only generalized property rights but also generalized contract enforcement. This required them to stop supporting particularized interventions by special-interest groups that diminished the security of contracts. Only when this was undertaken could the benefits of growth-favorable property rights be unleashed and economic growth quicken. Serfdom shows that distributional conflicts and the coercive powers of elites played a major role in contracting institutions, just as they did in the enforcement of property rights.

Fifth, serfdom shows that secure private property rights can be good or bad for growth, depending on whether they are generalized or particularized. Under serfdom, landlords had very secure, clearly defined, and extensive private property rights. But these were property rights that were particularized, in the sense that they were based on noneconomic characteristics of the owner: his personal status and legal privileges as a noble landlord and his possession of coercive power over his serfs. Transactions involving these secure private property rights were governed by the personal characteristics of the lord, including his coercive capacities. These very secure and well-defined private property rights prevented growth from taking off, by limiting the extent to which resources were allocated to the users that had the highest-productivity uses for them. Instead, the particularized property rights that prevailed under serfdom allocated assets to those with legal privileges and coercive capacities. The particularized nature of private property rights under serfdom limited the extent to which serfs could invest in increasing the productivity of their land, as well as their ability to use it as collateral to obtain loans for investment purposes.

Sixth, serfdom shows that security of private property rights—whether of ownership, use, or transfer—was a matter of degree, rather than presence or absence. In many European serf societies, serfs had rights of ownership over their holdings: in some, it was virtually impossible for a serf to be evicted from his farm by his landlord; in most others, eviction required a legal case to be made that the serf had violated the conditions of his tenure, for instance by failing to pay his rent or labor dues. In most European serf societies that have been studied, there were also secure rights of use, in the sense that serfs can be observed choosing which crops to cultivate (e.g. cash crops such as flax) and investing in their holdings (e.g. by constructing buildings or by manuring fields). In most European serf societies, serfs also bought, sold, and bequeathed their holdings, and were able to lease and rent at least some parcels of land. In principle, a serf required his landlord's permission for all land transfers, but in a majority of cases this was granted virtually automatically. This was certainly the case in England under serfdom, and thus long before 1688, since serfdom declined in England after c. 1350. Moreover, serfs had a considerable (if not perfect) degree of security of ownership and use rights over their property, not just in medieval England but in virtually every other European serf society that has ever been studied. Security of ownership and use over private property existed in nearly every medieval and early modern European society, but their generalized features were often constrained by the operation of adjacent or conflicting particularized institutional arrangements. Serfdom provides a clear example of how security of private property rights is a matter of degree rather than kind. It also illustrates the importance of breaking down the concept of "security" of property rights into its different components, examining each separately, and analyzing how each component influenced economic growth.

Seventh, serfdom shows clearly the importance of recognizing that institutions are embedded in a wider institutional system and are constrained by the other institutions in that system. Behind the facade of serfdom lay a set of institutional arrangements that varied greatly across different European societies and across time-periods. This was because serfdom did not exist in isolation, as a set of institutional rules governing the relationship between peasants and noble landlords. Rather, it was embedded in a wider system of other institutions—the market, the village community, the state, the family, and many others. The functioning of serfdom, its survival, and its impact on growth were all affected by the availability and often the active intervention of these other institutions.

Eighth, serfdom demonstrates the centrality of distributional conflicts to the evolution of institutional systems and their impact on growth. Serfdom survived for centuries in the teeth of changing resource endowments and rampant inefficiency, because it benefited powerful groups: landlords, rulers, and members of the serf oligarchy. But the distributional conflicts that sustained serfdom raged not only, or even predominantly, at the level of high politics. Rather, they consisted of lower-level and longer-lasting distributional struggles among special-interest groups, mostly outside the arena of national politics.

8.11. CONCLUSION

This chapter has sought to bring historical evidence to bear on the question of how institutions affect long-run economic growth. Although we still need to know much more about the institutions that influenced economic success in past centuries, there is much we can say even with the evidence we have, positively and negatively, about the conditions for growth. The growth literature contains a number of strong claims about economic history and institutions. This chapter has shown that some of these claims are not supported by historical research, and must be replaced. Others are controversial, and the evidence surveyed in this chapter has suggested the direction in which they must be revised. Still others are probably right, and this chapter has tried to show how they could be rendered more useful for theory and policy if they made better use of the historical evidence.

We can definitively rule out some very widely held hypotheses which claim that some specific, singular institution played a key causal role in economic growth. Private-order institutions are widely claimed to be capable of substituting for public-order institutions in supporting economic growth. But as we saw in Lessons 1 and 3, the historical examples which are supposed to support this view turn out not to have existed. Private-order institutions can supplement public-order institutions, but cannot substitute for them. Public-order institutions are necessary for markets to function—for good or ill. Parliaments are a second institution widely claimed to play a central role in facilitating economic growth. But, as we saw in Lesson 2, parliaments have a very spotty historical record of supporting growth, and in the few cases they have done so they appear to have required to possess very specific characteristics and to be embedded in a wider system of supporting institutions. Even secure private property rights, widely regarded as a key to economic growth, turn out not to have been invariably beneficial in the historical record. In those cases in which such property rights played an important causal role in growth, as in the European agricultural revolution they needed to possess the special characteristic of being generalized, and they needed also to be supported by other components of the institutional system, especially contracting institutions. These findings enable us to rule out simple institutional recipes, such as focusing solely on building private-order social networks, establishing parliaments, or developing property rights, at the expense of other parts of the institutional system.

A clear corollary emerges from these findings. Institutions do not operate in isolation but as part of a wider system. Property rights institutions are facilitated by contracting institutions and constrained by communal and manorial ones. Contracting institutions operate well or badly depending on public-order institutions; the organizing abilities of urban and rural communes; the privileges of corporative occupational associations; and the powers of landlords under manorial systems such as serfdom. The institution of the family is interdependent with the wider framework of non-familial institutions. Serfdom depended on the state, on peasant communes, and even on markets. Most of the central economic institutions over the past millennium appear to have affected growth only in interaction with other components of the wider institutional system.

The most important lesson from our investigation of institutions and growth in history, however, concerns perspectives for the future. Again and again, the result of our lessons has led us to the remark cited at the end of Lesson 7: "everything should be made as simple as possible, but not simpler." Two apparently opposed kinds of simplification are now particularly conspicuous. One of them tries to find the point at which the indispensable set of institutions came into existence. Since the Glorious Revolution of 1688 occurred conveniently about three generations before the first stirrings of English industrialization, it has been seized upon (as we saw in Lessons 2, 5, and 6) as the turning point of history, at which the institutions essential to growth began. The other, apparently opposite, simplification is that many societies have the right institutions, e.g. secure property rights, without experiencing growth. In particular, it is pointed out that 13th-century England had all the institutions that matter to growth, and yet failed to industrialize.

As we saw in a number of the lessons in this chapter, the apparent disagreement between these two kinds of simplification is superficial. What they agree on is more important—the assumption that institutions can be exhaustively described, in all their implications for growth, by their informal, ordinary-language names such as secure property rights, public-order institutions, or parliament. The assumption is that each such label refers unambiguously to a particular, identifiable social configuration of some kind. This chapter has shown that this assumption is untenable. The reason English economic history can be used to argue both that property rights are essential for growth and that property rights are irrelevant for growth is that property rights encompasses an enormous variety of heterogeneous phenomena. Informal institutional labels, as the historical evidence surveyed in this chapter has shown, are imprecise, they are ambiguous, and in many cases they overlap; none of them has anything like a sharp definition.

A major theme of this chapter has been that the entities referred to by these labels are not well defined—i.e. that the assumption shared by the two apparently opposite kinds of simplification is false. Conventional institutional labels are ill defined in at least three ways: they lack sharp criteria of application (they refer to a large variety of different social configurations); they lack a scale of intensity or degree (they are assumed to be either present or absent, with no gradations in between); and they fail to reflect the interconnections between the configuration they apparently refer to and the entire institutional system of which that configuration is an integral part, let alone to give any hint how the character of that configuration changes as its institutional context and interdependencies change. The historical findings surveyed in this chapter therefore open up three challenges for future research on institutions and growth.

The first challenge is to sharpen the criteria of application of conventional institutional labels. Each institutional label currently used in the analysis of economic growth refers to a large variety of different social configurations. Parliaments, even those representing the interests of wealth holders, as we saw in Lesson 2, can refer to anything from the post-1688 English parliament (relatively pluralistic, if still corrupt), to the 18th-century Württemberg *Landschaft* (the other constitutional monarchy in Europe, but manned by guildsmen and given to granting privileges to rent-seeking corporate groups), to the Polish *Sejm* (much more powerful than the feeble Polish executive, but mainly used to enforce the powers of noble landlords under serfdom). The historical evidence presented in this chapter suggests that economists need to break down the concept of parliament manned by wealth holders analytically by registering how wealth holders obtain their wealth, what kind of wealth it is, how wealth holders obtain representation in parliament, how variegated their economic interests are, and what mechanisms and levers of economic intervention the specific parliamentary institution grants to its members. Likewise, the conventional institutional label "secure property rights" has been applied by respectworthy economists and historians to property regimes as disparate as ninth-century Italy, 13th-century England, 17th-century Germany, and rich Western economies at the beginning of the 21st century. The historical evidence presented in Lessons 5 and 6 suggests that we need to break down the concept of secure private property rights into rights of ownership, use, and transfer; and within each type of right, analyze whether it is a generalized right applying to all economic agents or a particularized right applying only to a privileged subset. It seems likely that other conventional institutional labels—contracting institutions, communities, guilds—would benefit from analytical attention devoted to sharpening the criteria by which they are defined and measured, and the way in which these separate characteristics might be expected to affect economic growth.

The second challenge for future research is to provide a scale of intensity or degree for measuring institutions. The current institutional labels used in the analysis of growth assume those institutions to be either present or absent, with no gradations in between. The growth literature contains too many claims that certain institutions were completely absent or, alternatively, completely present. Public-order institutions are supposed to have been completely absent from the medieval trading world, as we saw in Lessons 1 and 3, implying a major role for private-order substitutes in achieving economic growth—and yet empirical research finds that public-order institutions were present and reveals that they served an important role in commercial growth in those medieval economies, even though they undoubtedly changed over the ensuing centuries, albeit not always in a positive direction. Parliaments are supposed to have had no control over the executive arm of the English government before 1688 and virtually complete control thereafter, as we saw in Lesson 2, implying a major role for democratization in achieving economic growth—and yet the empirical findings reveal that parliamentary powers were usually a matter of incremental changes, except during periods of revolution (and sometimes even then). Property rights, as we saw in Lesson 5, are portrayed as being either completely absent before 1688 or completely present in 1300, implying respectively a major role in economic growth or complete irrelevance to it—and yet the empirical findings reveal that property rights were a matter of degree and incremental change. The historical findings surveyed in this chapter suggest the need for economists to pay much greater analytical attention to devising scales of intensity or degree for conventional institutional labels such as property rights or public-order institutions, preferably for each of the many distinct characteristics of these institutions whose identification is the focus of our first challenge.

Our third challenge for future research is to work out ways of analyzing and measuring the linkages between the configurations to which conventional institutional labels apparently refer—that is, of understanding how institutions interconnect with the wider institutional system. Even very similar property rights regimes, as Lesson 4 showed, could give rise to very different economic outcomes during the agricultural revolution depending on the quality of contracting institutions, which in turn depended on the characteristics of such variegated institutional mechanisms as the village community, serfdom, urban corporations, and the state. As Lesson 7 showed, the apparently identical family institution of the European Marriage Pattern could be associated with widely varying growth outcomes, depending on the rest of the institutional system within which it was embedded, especially corporative institutions such as guilds and communities that influenced women's status, human capital investment, and demographic decisions. Even serfdom, as we saw in Lesson 8, cannot be understood in isolation from the rest of the institutional system—the village community, the state, and the market. The historical evidence surveyed in this chapter suggests that in order to understand institutional influences on long-run growth, economists need ways of characterizing the wider institutional system of which each institution is just one component, and of mapping how the character of that configuration changes as its institutional context and interdependencies change.

This is not to say that any of these challenges will be easy to surmount. But the historical findings surveyed in this chapter show that they will have to be tackled if we are to make further progress. Our best hope of success at this task will be to combine the ability of economics to simplify everything as much as possible, with the ability of history to identify where the complexity of the data resists further simplification and tells us that better analytical tools must be devised.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Jeremy Edwards for his exceptionally stimulating suggestions on several drafts of this chapter; and to Tracy Dennison for her very helpful comments on the final version.

REFERENCES

Acemoglu, D., 2003. Why not a political Coase theorem? Social conflict, commitment and politics. Journal of Comparative Economics 31 (4), 620–652.

Acemoglu, D., 2009. Introduction to Modern Economic Growth. Princeton University Press, Princeton, NJ.

Acemoglu, D., Cantoni, D., Johnson, S., Robinson, J.A., 2011. The consequences of radical reform: the French Revolution. American Economic Review 101 (7), 3286–3307.

Acemoglu, D., Johnson, S.H., 2005. Unbundling institutions. Journal of Political Economy 113 (4), 949–995.

Acemoglu, D., Johnson, S., Robinson, J.A., 2005. Institutions as a fundamental cause of long-run growth. In: Aghion, P., Durlauf, S.N. (Eds.), Handbook of Economic Growth, vol. 1A. Elsevier, Amsterdam/London, pp. 385–472.

Acemoglu, D., Robinson, J.A., 2012. Why Nations Fail: The Origins of Power, Prosperity and Poverty. Crown Publishers, New York.

Acemoglu, D., Wolitzky, A., 2011. The economics of labor coercion. Econometrica 79 (2), 555–600.

Ackerman-Lieberman, P., 2007. A partnership culture: Jewish economic and social life seen through the legal documents of the Cairo Geniza. Columbia University, Ph.D. Dissertation.

Aghion, P., Howitt, P., 1992. A model of growth through creative destruction. Econometrica 60 (2), 323–351.

A'Hearn, B., Baten, J., Crayen, D., 2009. Quantifying quantitative literacy: age heaping and the history of human capital. Journal of Economic History 69 (03), 783–808.

Alengry, C., 1915. Les foires de Champagne: Etude d'histoire économique. Rousseau et Cie, Paris.

Allen, R.C., 1992. Enclosure and the Yeoman: The Agricultural Development of the South Midlands, 1450–1850. Clarendon, Oxford.

Allen, R.C., 1999. Tracking the agricultural revolution in England. Economic History Review 52 (2), 209–235.

Allen, R.C., 2003. Progress and poverty in early modern Europe. Economic History Review 56 (3), 403–443.

Allen, R.C., 2004. Agriculture during the Industrial Revolution, 1700–1850. In: Floud, R., Johnson, P. (Eds.), The Cambridge Economic History of Modern Britain, vol. 1: Industrialisation, 1700–1860. Cambridge University Press, Cambridge, pp. 96–116.

Allen, R.C., 2011. Global Economic History: A Very Short Introduction. Oxford University Press, Oxford.

Anderson, J.E., 2008. Trade and informal institutions. In: Anon. (Ed.), Handbook of International Trade. Blackwell Publishing, Oxford, pp. 279–293.

Anon., 1818. The states of Wirtemberg. Edinburgh Review 29, 337–363.

Aoki, M., 2001. Toward a Comparative Institutional Analysis. MIT Press, Cambridge, MA.

Arbois de Jubainville, M. H. de, 1859. Histoire de Bar-sur-Aube sous les comtes de Champagne. Durand, Dufay-Robert, Jardeaux-Ray, Paris, Troyes, Bar-sur-Aube.

Arbois de Jubainville, M. H. de, Pigeotte, L., 1859–66. Histoire des ducs et des comtes de Champagne. A. Durand, Paris.

Archer, I.W., 1988. The London lobbies in the later sixteenth century. Historical Journal 31 (1), 17–44.

Aron, J., 2000. Growth and institutions: a review of the evidence. The World Bank Research Observer 15 (1), 99–135.

Ashton, R., 1967. The parliamentary agitation for free trade in the opening years of the reign of James I. Past & Present 38, 40–55.

Ashtor, E., 1983. The Levant Trade in the Later Middle Ages. Princeton University Press, Princeton, NJ.

Aslanian, S., 2006. Social capital, trust and the role of networks in Julfan trade: informal and semi-formal institutions at work. Journal of Global History 1 (3), 383–402.

Ba, S., 2001. Establishing online trust through a community responsibility system. Decision Support Systems 31 (3), 323–336.

Bairoch, P., 1989. Les trois révolutions agricoles du monde développé: rendements et productivité de 1800 à 1985. Annales. Histoire, Sciences Sociales 44 (2), 317–353.

Baker, J.H., 1979. The law merchant and the common law before 1700. Cambridge Law Journal 38, 295–322.

Baker, J.H., 1986. The law merchant and the common law. In: Baker, J.H. (Ed.), The Legal Profession and the Common Law: Historical Essays. Hambledon Press, London, pp. 341–386.

Barbour, V., 1911. Privateers and pirates of the West Indies. American Historical Review 16 (3), 523–566.

Bardhan, P., 1996. The nature of institutional impediments to economic development. Center for International and Development Economics Research Papers C96–066.

Basile, M.E., Bestor, J.F., Cocquillette, D.R., Donahue, C. (Eds.), 1998. Lex Mercatoria and Legal Pluralism: A Late Thirteenth-Century Treatise and Its Afterlife. Ames Foundation, Cambridge.

Bassermann, E., 1911. Die Champagnermessen. Ein Beitrag zur Geschichte des Kredits. Mohr, Tübingen.

Bautier, R.-H., 1953. Les foires de Champagne. Recherches sur une évolution historique. Recueils de la Société Jean Bodin 5, 97–147.

Bautier, R.-H., 1970. The fairs of Champagne. In: Cameron, R. (Ed.), Essays in French Economic History. R.D. Irwin, Homewood, IL, pp. 42–63.

Bekar, C.T., Reed, C.G., 2012. Land Markets and Inequality: Evidence from Medieval England. Simon Fraser University Department of Economics Working Papers 12–14.

Benson, B.L., 1989. The spontaneous evolution of commercial law. Southern Economic Journal 55 (3), 644–661.

Benson, B.L., 1998. Law merchant. In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law. Macmillan, London.

Benson, B.L., 2002. Justice without government: the merchant courts of medieval Europe and their modern counterparts. In: Beito, D., Gordon, P., Tabarrok, A. (Eds.), The Voluntary City: Choice, Community and Civil Society. University of Michigan Press, Ann Arbor, pp. 127–150.

Benson, B.L., 2005. The mythology of holdout as a justification for eminent domain and public provision of roads. The Independent Review 10 (2), 165–194.

Benson, B.L., 2008. The evolution of eminent domain: a remedy for market failure or an effort to limit government power and government failure? The Independent Review 12 (3), 423–432.

Benton, J.F., 1969. Philip the Fair and the Jours of Troyes. Studies in Medieval and Renaissance History 6, 281–344.

Bernstein, L., 2001. Private commercial law in the cotton industry: value creation through rules, norms, and institutions. Michigan Law Review 99 (7), 1724–1790.

Besley, T., Ghatak, M., 2010. Property rights and economic development. In: Rodrik, D., Rosenzweig, M.R. (Eds.), Handbook of Development Economics, vol. 5. North Holland, Amsterdam, pp. 4526–4595.

Bieleman, J., 1993. Dutch agriculture in the Golden Age, 1570–1660. Economic and Social History in the Netherlands 4, 159–185.

Bieleman, J., 2006. Dutch agricultural history c. 1500–1950. In: Thoen, E., Van Molle, L. (Eds.), Rural History in the North Sea Area : An Overview of Recent Research, Middle Ages-Twentieth Century. Brepols, Turnhout.

Bieleman, J., 2010. Five Centuries of Farming: A Short History of Dutch Agriculture, 1500–2000. Wageningen Academic Publishers, Wageningen.

Biller, P., 2001. The Measure of Multitude: Population in Medieval Thought. Oxford University Press, Oxford.

Birdsall, N., 1988. Analytical approaches to population growth. In: Chenery, H., Srinivasan, T.N. (Eds.), Handbook of Development Economics, vol. I. North Holland, Amsterdam/New York, pp. 477–542.

Blaschke, K., 1991. Dorfgemeinde und Stadtgemeinde in Sachsen zwischen 1300 und, 1800. In: Blickle, P. (Ed.), Landgemeinde und Stadtgemeinde in Mitteleuropa. Ein struktureller Vergleich. Oldenbourg, Munich, pp. 119–143.

Blockmans, W.P., 1988. Alternatives to monarchical centralization. In: Koenigsberger, H.G. (Ed.), Republik und Republikanismus in Europea der Frühen Neuzeit. Oldenbourg, Munich.

Blondé, B., Gelderblom, O., Stabel, P., 2007. Foreign merchant communities in Bruges, Antwerp and Amsterdam, c. 1350–1650. In: Calabi, D., Christensen, S.T. (Eds.), Cultural Exchange in Early Modern Europe, vol. 2: Cities and Cultural Exchange in Europe, 1400–1700. Cambridge University Press, Cambridge, pp. 154–174.

Boelcke, W.A., 1973. Wege und Probleme des industriellen Wachstums im Königreich Württemberg. Zeitschrift für Württembergische Landesgeschichte 32, 436–520.

Boelcke, W.A., 1984. Industrieller Aufstieg im mittleren Neckarraum zwischen Konjunktur und Krise. Zeitschrift für Württembergische Landesgeschichte 43, 287–326.

Boerner, L., Ritschl, A., 2002. Individual enforcement of collective liability in premodern Europe. Journal of Institutional and Theoretical Economics 158, 205–213.

Boerner, L., Ritschl, A., 2005. Making financial markets: contract enforcement and the emergence of tradable assets in late medieval Europe. Society for Economic Dynamics 2006 Meeting Papers 884.

Bogart, D., Richardson, G., 2009. Making property productive: reorganizing rights to real and equitable estates in Britain, 1660–1830. European Review of Economic History 13 (1), 3–30.

Bogart, D., Richardson, G., 2011. Property rights and parliament in industrializing Britain. Journal of Law and Economics 54 (2), 241–274.

Boldorf, M., 1999. Institutional Barriers to Economic Development: The Silesian Linen Proto-industry (17th to 19th Century). Institut für Volkswirtschaftslehre und Statistik, Universität Mannheim, Working Papers 566–99.

Boldorf, M., 2006. Europäische Leinenregionen im Wandel. Institutionelle Weichenstellungen in Schlesien und Irland (1750–1850). Böhlau, Cologne/Weimar/Vienna.

Boldorf, M., 2009. Socio-economic institutions and transaction costs: merchant guilds and rural trade in eighteenth-century Lower Silesia. European Review of Economic History 13 (2), 173–198.

Bolton, J.L., 2012. Money in the Medieval English Economy 973–1489. Manchester University Press, Manchester.

Bonfield, L., 2001. Developments in European family law. In: Kertzer, D.I., Barbagli, M. (Eds.), The History of the European Family, vol. 1: Family Life in Early Modern Times, 1500–1789. Yale University Press, New Haven, CT, pp. 87–124.

Bourquelot, F., 1839–40. Histoire de Provins. Lebeau, Paris.

Bourquelot, F., 1865. Études sur les foires de Champagne, sur la nature, l'étendue et les règles du commerce qui s'y faisait aux XIIe, XIIIe et XIVe siècles. L'Imprimerie Impériale, Paris.

Boutaric, E.P., 1867. Actes du Parlement de Paris, Première série: De l'an 1254 à l'an 1328. H. Plon, Paris.

Braddick, M.J., 1994. Parliamentary Taxation in Seventeenth-Century England: Local Administration and Response. Royal Historical Society, Woodbridge, Suffolk.

Brakensiek, S., 1991. Agrarreform und Ländliche Gesellschaft: die Privatisierung der Marken in Nordwestdeutschland 1750–1850. Schöningh, Paderborn.

Brakensiek, S., 1994. Agrarian individualism in North-Western Germany, 1770–1870. German History 12 (2), 137–179.

Brenner, R., 1976. Agrarian class structure and economic development in pre-industrial England. Past & present 70, 30–75.

Brewer, J., 1989. The Sinews of Power: War, Money and the English State, 1688–1783. Unwin Hyman, London.

Brewer, J., Hellmuth, E. (Eds.), 1999. Rethinking Leviathan: The Eighteenth-Century State in Britain and Germany. Oxford University Press, Oxford.

Briggs, C., 2004. Empowered or marginalized? Rural women and credit in later thirteenth- and fourteenth-century England. Continuity and Change 19 (1).

Briggs, C., 2009. Credit and Village Society in Fourteenth-Century England. Oxford University Press, Oxford.

Briggs, C., 2013. English serfdom, c. 1200 - c. 1350: towards an institutional analysis. In: Cavaciocchi, S. (Ed.), Schiavitu e servaggio nell'economia europea. Secc. XI-XVIII./Slavery and Serfdom in the European Economy from the 11th to the 18th Centuries. XLV settimana di studi della Fondazione istituto internazionale di storia economica F. Datini, Prato 14–18 April 2013. Firenze University Press, Florence.

Britnell, R.H., 1991. The towns of England and Northern Italy in the early fourteenth century. Economic History Review 44 (1), 21–35.

Britnell, R.H., 1996. The Commercialisation of English Society, 1000–1500. Manchester University Press, Manchester.

Briys, E., De ter Beerst, D.J., 2006. The Zaccaria deal: contract and options to fund a Genoese shipment of alum to Bruges in 1298. Paper presented at the XIV International Economic History Congress, Helsinki, August.

Broadberry, S., Campbell, B., Klein, A., Overton, M., et al., 2011. British economic growth, 1270–1870: an output-based approach. University of Kent Department of Economics Studies in Economics 1203.

Broadberry, S., Campbell, B., Van Leeuwen, B., 2013. When did Britain industrialise? The sectoral distribution of the labor force and labor productivity in Britain, 1381–1851. Explorations in Economic History 50 (1), 16–27.

Brophy, J.M., 1995. Salus publica suprema lex: Prussian Businessmen in the New Era and Constitutional Conflict. Central European History 28 (2), 122–151.

Burkhardt, M., 2012. Zentren und Peripherie zu Beginn der Industriellen Revolution in Württemberg - ein kartografischer Nachtrag. Zeitschrift für Württembergische Landesgeschichte 71, 479–480.

Buyst, E., Mokyr, J., 1990. Dutch manufacturing and trade during the French Period (1795–1814) in a long-term perspective. In: Aerts, E., Crouzet, F. (Eds.), Economic Effects of the French Revolutionary and Napoleonic Wars, Leuven University Press, Leuven, pp. 64–78.

Byrne, E.H., 1916. Commercial contracts of the Genoese in the Syrian trade of the twelfth century. Quarterly Journal of Economics 31 (1), 128–170.

Byrne, E.H., 1930. Genoese Shipping in the Twelfth and Thirteenth Centuries. The Medieval Academy of America, Cambridge, MA.

Calaprice, A. (Ed.), 2011. The Ultimate Quotable Einstein. Princeton University Press, Princeton, NJ.

Caldwell, J.C., 1976. Toward a restatement of demographic transition theory. Population and Development Review 2 (3/4), 321–366.

Caldwell, J.C., 1982. Theory of Fertility Decline. Academic Press, London, New York.

Cameron, R.E., 1989. A Concise Economic History of the World: From Paleolithic Times to the Present. Oxford University Press, Oxford.

Campbell, B.M.S., 2000. English Seigniorial Agriculture, 1250–1450. Cambridge University Press, Cambridge.

Campbell, B.M.S., 2005. The agrarian problem in the early fourteenth century. Past & Present 188, 3–70.

Campbell, B.M.S., 2009. Factor markets in England before the Black Death. Continuity and Change 24 (1), 79–106.

Campbell, C., Lee, J.Z., 2010. Demographic impacts of climatic fluctuations in northeast China, 1749–1909. In: Kurosu, S., Bengtsson, T., Campbell, C. (Eds.), Demographic Responses to Economic and Environmental Crisis. Reitaku University Press, Kashiwa, pp. 107–132.

Campbell, B.M.S., Overton, M., 1998. L'histoire agraire anglaise jusqu'en 1850: revue historiographique sur l'état actuel de la recherche'. Histoire et Sociétés rurales 9 (1), 77–105.

Carrigan, W.D., 2004. The Making of a Lynching Culture: Violence and Vigilantism in Central Texas, 1836–1916. University of Illinois Press, Urbana/Chicago, IL.

Carsten, F.L., 1950. The Great Elector and the foundation of the Hohenzollern despotism. English Historical Review 65 (255), 175–202.

Carsten, F.L., 1959. Princes and Parliaments in Germany. Clarendon, Oxford.

Caunce, S., 1997. Farm servants and the development of capitalism in English agriculture. Agricultural History Review 45 (1), 49–60.

Cerman, M., 1996. Proto-industrialisierung und Grundherrschaft. Ländliche Sozialstruktur, Feudalismus und Proto-industrielles Heimgewerbe in Nordböhmen vom 14. bis zum 18. Jahrhundert (1381–1790). Ph.D. Dissertation, Vienna.

Cerman, M., 2008. Social structure and land markets in late medieval central and east-central Europe. Continuity and Change 23 (1), 55–100.

Cerman, M., 2012. Villagers and Lords in Eastern Europe, 1300–1800. Palgrave Macmillan, Houndmills/New York.

Cerman, M., 2013. Seigniorial systems in east-central and eastern Europe, 1300–1800: regional realities. In: Cavaciocchi, S. (Ed.), Schiavitu e servaggio nell'economia europea. Secc. XI-XVIII./Slavery and Serfdom in the European Economy from the 11th to the 18th Centuries. XLV settimana di studi della Fondazione istituto internazionale di storia economica F. Datini, Prato 14–18 April 2013. Firenze University Press, Florence.

Chambers, R., 1869. History of the Rebellion of 1745–6, W. & R. Chambers, London.

Chambers, J.D., 1953. Enclosure and labour supply in the Industrial Revolution. Economic History Review 5 (3), 319–343.

Chapin, E., 1937. Les villes de foires de Champagne des origines au début du XIVe siècle. Champion, Paris.

Cheyette, F.L., 1970. The sovereign and the pirates, 1332. Speculum 45 (1), 40–68.

Chorley, G.P.H., 1981. The agricultural revolution in northern Europe, 1750–1880: nitrogen, legumes, and crop productivity. Economic History Review 34 (1), 71–93.

Clark, G., 1996. The political foundations of modern economic growth: England, 1540–1800. Journal of Interdisciplinary History 26 (4), 563–588.

Clark, C.M., 2006. Iron Kingdom: The Rise and Downfall of Prussia, 1600–1947. Allen Lane, London.

Clark, G., 2007. A Farewell to Alms: A Brief Economic History of the World. Princeton University Press, Princeton, NJ.

Clark, G., 2010. The macroeconomic aggregates for England, 1209–1869. Research in Economic History 27, 51–140.

Clay, K., 1997. Trade without law: private-order institutions in Mexican California. Journal of Law, Economics and Organization 13 (1), 202–231.

Cohen, M.R., 2013. A partnership gone bad: a letter and a power of attorney from the Cairo Geniza, 1085. In: Wasserstein, D., Ghanaim, M. (Eds.), The Sasson Somekh Festschrift [not yet titled], Tel Aviv.

Coleman, J.S., 1988. Social capital in the creation of human capital. American Journal of Sociology 94, S95–S120.

Court, R., 2004. "Januensis ergo mercator": trust and enforcement in the business correspondence of the Brignole family. Sixteenth Century Journal 35 (4), 987–1003.

Crafts, N., 1977. Industrial Revolution in Britain and France: some thoughts on the question, "Why was Britain first?" Economic History Review 2nd ser. 30, 429–441.

Crafts, N., 1985. British Economic Growth during the Industrial Revolution. Clarendon, Oxford.

Crafts, N.F.R., 1987. British economic growth, 1700–1850; some difficulties of interpretation. Explorations in Economic History 24 (3), 245–268.

Crafts, N., Mills, T.C., 2009. From Malthus to Solow: how did the Malthusian economy really evolve? Journal of Macroeconomics 31 (1), 68–93.

Crafts, N., Leybourne, S.J., Mills, T.C., 1989. Trends and cycles in British industrial production, 1700–1913. Journal of the Royal Statistical Society Series A (Statistics in Society) 152 (1), 43–60.

Croft, P., 1973. Introduction: the revival of the Company, 1604–6. In: Croft, P. (Ed.), The Spanish Company. London Record Society, London, pp. xxix-li.

Czaja, R., 2009. Die Entwicklung der ständischen Versammlungen in Livland, Preußen und Polen im Spätmittelalter. Zeitschrift für Ostmitteleuropa-Forschung 58 (3), 312–328.

Czapliński, W., 1985. The Polish Parliament at the summit of its development. Zakład Narodowy Imienia Ossolińskich, Wrocław.

Dahl, G., 1998. Trade, Trust and Networks: Commercial Culture in Late Medieval Italy. Nordic Academic Press, Lund.

Dasgupta, P., 1993. An Inquiry into Well-Being and Destitution. Clarendon Press, Oxford.

Dasgupta, P., 2000. Economic progress and the idea of social capital. In: Dasgupta, P., Serageldin, I. (Eds.), Social Capital: A Multifaceted Perspective. World Bank, Washington, pp. 325–424.

Davids, K., 2006. Monasteries, economies and states: the dissolution of monasteries in early modern Europe and T'ang China. Paper presented at the Global Economic History Network (GEHN) Conference 10, Washington, 8–10 September 2006.

Davidsohn, R., 1896–1901. Forschungen zur Geschichte von Florenz. E.S. Mittler und Sohn, Berlin.

Davidson, J., Weersink, A., 1998. What does it take for a market to function? Review of Agricultural Economics 20 (2), 558–572.

Defoe, D., 1727. The Complete English Tradesman. Charles Rivington, London.

Del Vecchio, A., Casanova, E., 1894. Le rappresaglie nei comuni medievali e specialmente in Firenze. C. e G. Zanichelli, Bologna.

De Moor, T., 2008. The silent revolution: a new perspective on the emergence of commons, guilds, and other forms of corporate collective action in western Europe. International Review of Social History 53, 179–212.

De Moor, T., Van Zanden, J.L., 2010. Girlpower: the European Marriage Pattern and labour markets in the North Sea region in the late medieval and early modern period. Economic History Review 63 (1), 1–33.

Dennison, T., 2011. The Institutional Framework of Russian Serfdom. Cambridge University Press, Cambridge.

Dennison, T., 2013. The institutional framework of serfdom in Russia: the view from 1861. In: Cavaciocchi, S. (Ed.), Schiavitu e servaggio nell'economia europea. Secc. XI-XVIII./Slavery and Serfdom in the European Economy from the 11th to the 18th Centuries. XLV settimana di studi della Fondazione istituto internazionale di storia economica F. Datini, Prato 14–18 April 2013. Firenze University Press, Florence.

Dennison, T., Ogilvie, S., 2007. Serfdom and social capital in Bohemia and Russia. Economic History Review 60 (3), 513–544.

Dennison, T., Ogilvie, S., 2013. Does the European Marriage Pattern Explain Economic Growth? CESifo Working Paper 4244.

De Roover, R., 1948. The Medici Bank: Its Organization, Management, Operations and Decline. New York University Press, New York.

De Roover, R., 1963. The Rise and Decline of the Medici Bank, 1397–1494. Harvard University Press, Cambridge, MA.

De Soto, H., 1989. The Other Path: The Invisible Revolution in the Third World. Harper & Row, New York.

De Soto, H., 2000. The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else. Basic Books, New York.

Dessí, R., Ogilvie, S., 2003. Social Capital and Collusion: The Case of Merchant Guilds. CESifo Working Papers 1037.

Dessí, R., Ogilvie, S., 2004. Social Capital and Collusion: The Case of Merchant Guilds (Long Version). Cambridge Working Papers in economics 0417.

De Vries, J., 1974. The Dutch Rural Economy in the Golden Age, 1500–1700. Yale University Press, New Haven, CT.

De Vries, J., 1976. The Economy of Europe in an Age of Crisis, 1600–1750. Cambridge University Press, Cambridge.

De Vries, J., Van der Woude, A., 1997. The First Modern Economy: Success, Failure, and Perseverance of the Dutch Economy, 1500–1815. Cambridge University Press, Cambridge.

Dewey, H.W., 1988. Russia's debt to the Mongols in suretyship and collective responsibility. Comparative Studies in Society and History 30 (2), 249–270.

Dewey, H.W., Kleimola, A.M., 1970. Suretyship and collective responsibility in pre-Petrine Russia. Jahrbücher für Geschichte Osteuropas 18, 337–354.

Dewey, H.W., Kleimola, A.M., 1984. Russian collective consciousness: the Kievan roots. Slavonic and East European Review 62 (2), 180–191.

Diamond, J., 1997. Guns, Germs and Steel. W.W. Norton, New York, NY.

Dijkman, J., 2007. Debt litigation in medieval Holland, c. 1200 – c. 1350. Paper presented at the GEHN conference, Utrecht, 20–22 September 2007.

Dixit, A.K., 2004. Lawlessness and Economics: Alternative Modes of Governance. Princeton University Press, Princeton, NJ.

Dixit, A.K., 2009. Governance institutions and economic activity. American Economic Review 99 (1), 5–24.

Doehaerd, R., 1941. Les relations commerciales entre Gênes, la Belgique, et l'Outremont d'après les archives notariales gênoises aux XIIIe et XIVe siècles. Palais des académies, Brussels.

Doepke, M., Tertilt, M., 2011. Does Female Empowerment Promote Economic Development? World Bank Policy Research Working Paper 5714.

Dollinger, P., 1970. The German Hansa. Macmillan, London.

Domar, E.D., 1970. The causes of slavery or serfdom: a hypothesis. Journal of Economic History 30 (1), 18–32.

Donahue, C., 1983. The canon law on the formation of marriage and social practice in the later Middle Ages. Journal of Family History 8 (2), 144–158.

Donahue, C., 2008. Law, Marriage, and Society in the Later Middle Ages: Arguments about Marriage in Five Courts. Cambridge University Press, Cambridge.

Dormois, J.-P., 1994. Entwicklungsmuster der Protoindustrialisierung im Mömpelgarder Lande während des 18. Jahrhunderts. Zeitschrift für Württembergische Landesgeschichte 53, 179–204.

Dotson, J.E., 1999. Fleet operations in the first Genoese-Venetian war, 1264–1266. Viator: Medieval and Renaissance Studies 30, 165–180.

Doumerc, B., 1987. Les Vénitiens à La Tana (Azov) au XVe siècle. Cahiers du monde russe et soviétique 28 (1), 5–19.

Edwards, J., Ogilvie, S., 2008. Contract Enforcement, Institutions and Social Capital: The Maghribi Traders Reappraised. CESifo Working Papers 2254.

Edwards, J., Ogilvie, S., 2012a. Contract enforcement, institutions, and social capital: the Maghribi traders reappraised. Economic History Review 65 (2), 421–444.

Edwards, J., Ogilvie, S., 2012b. What lessons for economic development can we draw from the Champagne fairs? Explorations in Economic History 49 (2), 131–148.

Edwards, J., Ogilvie, S., 2013. Economic growth in Prussia and Württemberg, c. 1750 - c. 1900. Unpublished paper, University of Cambridge, June 2013.

Ehmer, J., 1991. Heiratsverhalten, Sozialstruktur und ökonomischer Wandel. England und Mitteleuropa in der Formationsperiode des Kapitalismus. Vandenhoeck & Ruprecht, Göttingen.

Ekelund, R.B., Tollison, R.D., 1981. Mercantilism as a Rent-Seeking Society: Economic Regulation in Historical Perspective. Texas A&M University Press, College Station, TX.

Elton, G.R., 1975. Taxation for war and peace in early Tudor England. In: Winter, J.M. (Ed.), War and Economic Development: Essays in Memory of David Joslin. Cambridge University Press, Cambridge.

Epstein, S.A., 1996. Genoa and the Genoese, 958–1528. University of North Carolina Press, Chapel Hill, NC.

Epstein, S.R., 1998. Craft guilds, apprenticeship, and technological change in preindustrial Europe. Journal of Economic History 58, 684–713.

Ewert, U.-C., Selzer, S., 2009. Building bridges, closing gaps: the variable strategies of Hanseatic merchants in heterogeneous mercantile environments. In: Murray, J.M., Stabel, P. (Eds.), Bridging the Gap: Problems of Coordination and the Organization of International Commerce in Late Medieval European Cities. Brepols, Turnhout.

Ewert, U.-C., Selzer, S., 2010. Wirtschaftliche Stärke durch Vernetzung. Zu den Erfolgsfaktoren des hansischen Handels. In: Häberlein, M., Jeggle, C. (Eds.), Praktiken des Handels: Geschäfte und soziale Beziehungen europäischer Kaufleute in Mittelalter und früher Neuzeit. UvK Verlag, Konstanz, pp. 39–70.

Faille, C., 2007. Trading on reputation. Reason (January, 2007).

Fairlie, S., 1965. The nineteenth-century corn law reconsidered. Economic History Review 18 (3), 562–575.

Fairlie, S., 1969. The Corn Laws and British wheat production, 1829–76. Economic History Review 22 (1), 88–116.

Feller, L., 2004. Quelques problèmes liés à l'étude du marché de la terre durant le Moyen Âge. In: Cavaciocchi, S. (Ed.), Il mercato della terra: secc. XIII-XVIII: atti della trentacinquesima Settimana di studi, 5–9 maggio 2003, Le Monnier, Florence, pp. 21–47.

Fenoaltea, S., 1975a. Authority, efficiency, and agricultural organization in medieval England and beyond: a hypothesis. Journal of Economic History 35 (3), 693–718.

Fenoaltea, S., 1975b. The rise and fall of a theoretical model: the manorial system. Journal of Economic History 35 (2), 386–409.

Fenoaltea, S., 1984. Slavery and supervision in comparative perspective: a model. Journal of Economic History 44, 635–668.

Fertig, G., 2007. Äcker, Wirte, Gaben. Ländlicher Bodenmarkt und liberale Eigentumsordnung im Westfalen des 19. Jahrhunderts. Akademie Verlag, Berlin.

Feuchtwanger, E.J., 1970. Prussia: Myth and Reality. The Role of Prussia in German History. Wolff, London.

Fischel, W.A., 1995. Regulatory Takings: Law, Economics, and Politics. Harvard University Press, Cambridge, MA.

Fliegauf, U., 2007. Die Schwäbischen Hüttenwerke zwischen Staats- und Privatwirtschaft. Zur Geschichte der Eisenverarbeitung in Württemberg (1803–1945). Thorbecke, Ostfildern.

Flik, R., 1990. Die Textilindustrie in Calw und in Heidenheim 1705–1870. Eine regional vergleichende Untersuchung zur Geschichte der Frühindustrialisierung und Industriepolitik in Württemberg. Steiner, Stuttgart.

Foreman-Peck, J., 2011. The Western European marriage pattern and economic development. Explorations in Economic History 48 (2), 292–309.

Fortunati, M., 2005. The fairs between lex mercatoria and ius mercatorum. In: Piergiovanni, V. (Ed.), From Lex Mercatoria to Commercial Law. Duncker & Humblot, Berlin, pp. 143–164.

Friedman, M.A., 2006. Qusayr and Geniza documents on the Indian Ocean trade. Journal of the American Oriental Society 126 (3), 401–409.

Fritschy, W., 2003. A "financial revolution" reconsidered: public finance in Holland during the Dutch Revolt, 1568–1648. Economic History Review 56 (1), 57–89.

Frost, R.I., 2006. The nobility of Poland-Lithuania, 1569–1795. In: Scott, H.M. (Ed.), European Nobilities in the 17th and 18th Centuries: Northern, Central and Eastern Europe. Palgrave-Macmillan, New York, NY, pp. 266–310.

Galloway, P.R., 1988. Basic patterns in annual variations in fertility, nuptiality, mortality, and prices in pre-industrial Europe. Population Studies 42 (2), 275–303.

Galor, O., 2005a. The demographic transition and the emergence of sustained economic growth. Journal of the European Economic Association 3 (2/3), 494–504.

Galor, O., 2005b. From stagnation to growth: unified growth theory. In: Aghion, P., Durlauf, S.N. (Eds.), Handbook of Economic Growth, vol. 1, Part A. Elsevier, Amsterdam/London, pp. 171–293.

Galor, O., 2012. The demographic transition: causes and consequences. Cliometrica 6 (1), 1–28.

Gash, N., 1961. Mr Secretary Peel: The Life of Sir Robert Peel to 1830. Longman, London.

Gash, N., 1972. Sir Robert Peel : The life of Sir Robert Peel After 1830. Longman, Harlow.

Gelderblom, O., 2003. The governance of early modern trade: the case of Hans Thijs (1556–1611). Enterprise and Society 4 (4), 606–639.

Gelderblom, O. 2005a. The decline of fairs and merchant guilds in the Low Countries, 1250–1650. Economy and Society of the Low Countries Working Papers 2005–1.

Gelderblom, O. 2005b. The Resolution of Commercial Conflicts in Bruges, Antwerp, and Amsterdam, 1250–1650. Economy and Society of the Low Countries Working Papers 2005–2.

Gelderblom, O., 2013. Cities of Commerce: The Institutional Foundations of International Trade in the Low Countries, 1250–1650. Princeton University Press, Princeton, NJ.

Gelderblom, O., Grafe, R., 2004. The costs and benefits of merchant guilds, 1300–1800: position paper. Paper presented at the Fifth European Social Science History Conference, Berlin, 24–27 March 2004.

Gil, M., 2003. The Jewish merchants in the light of eleventh-century Geniza documents. Journal of the Economic and Social History of the Orient 46 (3), 273–319.

Gil, M., 2004a. Institutions and events of the eleventh century mirrored in Geniza letters (Part I). Bulletin of the School of Oriental and African Studies 67 (2), 151–167.

Gil, M., 2004b. Jews in Islamic Countries in the Middle Ages. Brill, Leiden.

Goitein, S.D., 1966. Studies in Islamic History and Institutions. Brill, Leiden.

Goitein, S.D., 1967/93. A Mediterranean Society: The Jewish Communities of the Arab World as Portrayed in the Documents of the Cairo Geniza. University of California Press, Berkeley/Los Angeles.

Goitein, S.D., Friedman, M.A., 2007. India Traders of the Middle Ages: Documents from the Cairo Geniza ("India Book"). Brill, Leiden/Boston.

Goldberg, J., 2005. Geographies of trade and traders in the eleventh-century Mediterranean: A study based on documents from the Cairo Geniza. Columbia University, Ph.D. Dissertation.

Goldberg, J., 2012a. Trade and Institutions in the Medieval Mediterranean: The Geniza Merchants and their Business World. Cambridge University Press, Cambridge.

Goldberg, J., 2012b. The use and abuse of commercial letters from the Cairo Geniza. Journal of Medieval History 38 (2), 127–154.

Goldberg, J.L., 2012c. Choosing and enforcing business relationships in the eleventh-century mediterranean: reassessing the "Maghribī traders". Past & Present 216 (1), 3–40.

Goldschmidt, L., 1891. Handbuch des Handelsrechts. Enke, Stuttgart.

Goldsworthy, J.D., 1999. The Sovereignty of Parliament: History and Philosophy. Clarendon Press, Oxford.

Goldthwaite, R.A., 1987. The Medici Bank and the world of Florentine capitalism. Past & Present 114, 3–31.

González de Lara, Y., 2005. The Secret of Venetian Success: The Role of the State in Financial Markets. Instituto Valenciano de Investigaciones Económicas (IVIE) Working Paper WP-AD 2005–28.

Grafe, R., Gelderblom, O., 2010. The rise and fall of the merchant guilds: re-thinking the comparative study of commercial institutions in pre-modern Europe. Journal of Interdisciplinary History 40 (4), 477–511.

Grantham, G.W., Sarget, M.-N., 1997. Espaces privilégiés: Productivité agraire et zones d'approvisionnement des villes dans l'Europe préindustrielle. Annales. Histoire, Sciences Sociales 52(3), 695–725.

Greif, A., 1989. Reputation and coalitions in medieval trade: evidence on the Maghribi traders. Journal of Economic History 49 (4), 857–882.

Greif, A., 1993. Contract enforceability and economic institutions in early trade: the Maghribi traders' coalition. American Economic Review 83 (3), 525–548.

Greif, A., 1994. Cultural beliefs and the organization of society: a historical and theoretical reflection on collectivist and individualist societies. Journal of Political Economy 102 (5), 912–950.

Greif, A., 1997. On the Social Foundations and Historical Development of Institutions that Facilitate Impersonal Exchange: From the Community Responsibility System to Individual Legal Responsibility in Pre-modern Europe. Stanford University Working Papers 97–016.

Greif, A., 2002. Institutions and impersonal exchange: from communal to individual responsibility. Journal of Institutional and Theoretical Economics 158 (1), 168–204.

Greif, A., 2004. Impersonal exchange without impartial law: the community responsibility system. Chicago Journal of International Law 5 (1), 109–138.

Greif, A., 2006a. Family structure, institutions, and growth: the origins and implications of western corporations. American Economic Review: Papers and Proceedings 96 (2), 308–312.

Greif, A., 2006b. History lessons. The birth of impersonal exchange: the community responsibility system and impartial justice. Journal of Economic Perspectives 20 (2), 221–236.

Greif, A., 2006c. Institutions and the Path to the Modern Economy: Lessons from Medieval Trade. Cambridge University Press, Cambridge.

Greif, A., 2012. The Maghribi traders: a reappraisal? Economic History Review 65 (2), 445–469.

Greif, A., Milgrom, P., Weingast, B., 1994. Coordination, commitment, and enforcement: the case of the merchant guild. Journal of Political Economy 102 (4), 912–950.

Greif, A., Tabellini, G., 2010. Cultural and institutional bifurcation: China and Europe compared. American Economic Review: Papers and Proceedings 100 (2), 135–140.

Greve, A., 2000. Brokerage and trade in medieval Bruges: regulation and reality. In: Stabel, P., Blondé, B., Greve, A. (Eds.), International Trade in the Low Countries 14th-16th Centuries. Garant, Leuven/Apeldoorn, pp. 37–44.

Greve, A., 2001. Die Bedeutung der Brügger Hosteliers für hansische Kaufleute im 14. und 15. Jahrhundert. Jaarboek voor middeleeuwse geschiedenis 4, 259–296.

Greve, A., 2007. Hansen, Hosteliers und Herbergen: Studien zum Aufenthalt hansischer Kaufleute in Brügge im 14. und 15. Jahrhundert. Brepols, Turnhout.

Grossman, G.M., Helpman, E., 1991. Innovation and Growth in the Global Economy. The MIT Press, Cambridge, MA.

Grotius, H., 1625. De jure belli ac pacis libri tres, in quibus jus naturae et gentium, item juris publici praecipua explicantur. Buon, Paris.

Grube, W., 1954. Dorfgemeinde und Amtsversammlung in Altwürttemberg. Zeitschrift für Württembergische Landesgeschichte 13, 194–219.

Grube, W., 1957. Der Stuttgarter Landtag, 1457–1957. Ernst Klett Verlag, Stuttgart.

Grube, W., 1974. Stadt und Amt in Altwürttemberg. In: Maschke, E., Sydow, J. (Eds.), Stadt und Umland: Protokoll der X. Arbeitstagung des Arbeitskrieses für südwestdeutsche Stadtgeschichtsforschung, Calw, 12.-14. November 1971, Kohlhammer, Stuttgart, pp. 20–28.

Guinnane, T.W., 2011. The historical fertility transition: A guide for economists. Journal of Economic Literature 49 (3), 589–561.

Guinnane, T.W., Ogilvie, S., 2008. Institutions and demographic responses to shocks: Württemberg, 1634–1870. Yale University Economic Growth Center Discussion Paper 962.

Guinnane, T.W., Ogilvie, S., 2013. A Two-Tiered Demographic System: "Insiders" and "Outsiders" in Three Swabian Communities, 1558–1914. Yale University Economic Growth Center Discussion Paper 1021.

Guzowski, P., 2013. The role of enforced labour in the economic development of church and royal estates in 15th and 16th-century Poland. In: Cavaciocchi, S. (Ed.), Schiavitu e servaggio nell'economia europea. Secc. XI-XVIII./Slavery and Serfdom in the European economy from the 11th to the 18th Centuries. XLV settimana di studi della Fondazione istituto internazionale di storia economica F. Datini, Prato 14–18 April 2013. Firenze University Press, Florence.

Gysin, J., 1989. "Fabriken und Manufakturen" in Württemberg während des ersten Drittels des 19. Jahrhunderts, Scripta Mercaturae Verlag, St. Katharinen.

Habakkuk, J., 1994. Marriage, Debt, and the Estates System: English Landownership 1650-1950. Clarendon, Oxford.

Hafter, D.M., 2007. Women at Work in Pre-industrial France. Penn State Press, University Park, PA.

Hagen, W.W., 1989. Seventeenth-century crisis in Brandenburg: the Thirty Years War, the destabilization of serfdom, and the rise of absolutism. American Historical Review 94 (2), 302–325.

Hagen, W.W., 2002. Ordinary Prussians. Brandenburg Junkers and Villagers 1500–1840. Cambridge University Press, Cambridge.

Harbord, D., 2006. Enforcing Cooperation among Medieval Merchants: the Maghribi Traders Revisited. Munich Personal Repec Archive Working Paper.

Hardin, G., 1968. The tragedy of the commons. Science 162 (3859), 1243–1248.

Harnisch, H., 1972. Zur Herausbildung und Funktionsweise von Gutswirtschaft und Gutsherrschaft. Eine Klageschrift der Bauern der Herrschaft Neugattersleben aus dem Jahre 1610. Jahrbuch für Regionalgeschichte 4, 179–199.

Harnisch, H., 1975. Klassenkämpfe der Bauern in der Mark Brandenburg zwischen frühbürgerlicher Revolution und Dreißigjährigem Krieg. Jahrbuch für Regionalgeschichte 5, 142–172.

Harnisch, H., 1986. Peasants and markets: the background to the agrarian reforms in feudal Prussia east of the Elbe, 1760–1807. In: Evans, R.J., Lee, W.R. (Eds.), The German Peasantry: Conflict and Community in Rural Society from the Eighteenth to the Twentieth Centuries. Croom Helm, London, pp. 37–70.

Harnisch, H., 1989a. Bäuerliche Ökonomie und Mentalität unter den Bedingungen der ostelbischen Gutsherrschaft in den letzten Jahrzehnten vor Beginn der Agrarreformen. Jahrbuch für Wirtschaftsgeschichte 1989 (3), 87–108.

Harnisch, H., 1989b. Die Landgemeinde in der Herrschaftsstruktur des feudalabsolutistischen Staates. Dargestellt am Beispiel von Brandenburg-Preussen. Jahrbuch für Geschichte des Feudalismus 13, 201–245.

Harnisch, H., 1994. Der preußische Absolutismus und die Bauern. Sozialkonservative Gesellschaftspolitik und Vorleistung zur Modernisierung. Jahrbuch für Wirtschaftsgeschichte 1994 (2), 11–32.

Harreld, D.J., 2004a. High Germans in the Low Countries: German Merchants and Commerce in Golden Age Antwerp. Brill, Leiden.

Harreld, D.J., 2004b. Merchant and guild: the shift from privileged group to individual entrepreneur in sixteenth-century Antwerp. Paper delivered at the Fifth European Social Science History Conference. Berlin, 24–27 March 2004.

Harris, R., 2004. Government and the economy, 1688–1850. In: Floud, R., Johnson, P. (Eds.), The Cambridge Economic History of Modern Britain, vol. 1: Industrialisation, 1700–1860. Cambridge University Press, Cambridge, pp. 204–237.

Harrison, G., 1990. Prerogative revolution and Glorious Revolution: political proscription and parliamentary undertaking, 1687–1688. Parliaments, Estates and Representation 10 (1), 29–43.

Harriss, G.L., 1975. King, Parliament, and Public Finance in Medieval England to 1369. Clarendon Press, Oxford.

Hartley, T.E., 1992. Elizabeth's Parliaments: Queen, Lords, and Commons, 1559–1601. Manchester University Press, Manchester.

Helpman, E., 2004. The Mystery of Economic Growth. Harvard University Press, Cambridge, MA.

Henderson, W.O., 1961a. Die Struktur der preußischen Wirtschaft um 1786. Zeitschrift für die Gesamte Staatswissenschaft 117, 292–319.

Henderson, W.O., 1961b. The Industrial Revolution on the Continent: Germany, France, Russia, 1800–1914. F. Cass, London.

Henderson, W.O., 1961c. The rise of the metal and armament industries in Berlin and Brandenburg, 1712–1795. Business History 3 (2), 63–74.

Henn, V., 1999. Der "dudesche kopman" zu Brügge und seine Beziehungen zu den "nationes" der übrigen Fremden im späten Mittelalter. In: Jörn, N., Kattinger, D., Wernicke, H. (Eds.), "Kopet uns werk by tyden": Beiträge zur hansischen und preussischen Geschichte. Walter Stark zum 75. Geburtstag. Thoms Helms Verlag, Schwerin, pp. 131–142.

Hickson, C.R., Thompson, E.A., 1991. A new theory of guilds and European economic development. Explorations in Economic History 28, 127–168.

Hillmann, H., 2013. Economic institutions and the state: insights from economic history. Annual Review of Sociology 39 (1), 215–273.

Hilton, B., 1977. Corn, Cash, Commerce: The Economic Policies of the Tory Governments, 1815–1830. Oxford University Press, Oxford.

Hilton, B., 2006. Mad, Bad, and Dangerous People? England, 1783–1846. Clarendon Press, Oxford.

Hine, K.D., 1998. Vigilantism revisited: an economic analysis of the law of extra-judicial self-help or why can't Dick shoot Henry for stealing Jane's truck. American University Law Review 47, 1221–1255.

Hippel, W. von, 1977. Die Bauernbefreiung im Königreich Württemberg. Harald Boldt, Boppard am Rhein.

Hippel, W. von, 1992. Wirtschafts- und Sozialgeschichte 1800 bis 1918. In: Schwarzmaier, H., Fenske, H., Kirchgässner, B., Sauer, P., Schaab, M. (Eds.), Handbuch der baden-württembergischen Geschichte: vol. 3: Vom Ende des Alten Reiches bis zum Ende der Monarchien. Klett-Cotta, Stuttgart, pp. 477–784.

Hodgskin, T., 1820. Travels in the North of Germany: Describing the Present State of the Social and Political Institutions, the Agriculture, Manufactures, Commerce, Education, Arts and Manners in that Country Particularly in the Kingdom of Hannover. A. Constable, Edinburgh.

Hohorst, G., 1977. Wirtschaftswachstum und Bevölkerungsentwicklung in Preußen 1816 bis 1914. Arno, New York.

Holderness, B.A., 1976. Credit in English rural society before the nineteenth century, with special reference to the period 1650–1720. Agricultural History Review 24, 97–109.

Hoppit, J., 1996. Patterns of parliamentary legislation, 1660–1800. The History Journal 39, 109–131.

Hoppit, J., 2011. Compulsion, compensation and property rights in Britain, 1688–1833. Past & Present 210, 93–128.

Hoyle, R.W., 1994. Parliament and taxation in sixteenth-century England. English Historical Review 109 (434), 1174–1196.

Israel, J., 1989. Dutch primacy in world trade, 1585–1740. Clarendon, Oxford.

Jacoby, D., 2003. Foreigners and the urban economy in Thessalonike, ca. 1150-ca. 1450. Dumbarton Oaks Papers 57, 85–132.

Johansson, E., 1977. The history of literacy in Sweden, in comparison with some other countries. Educational Reports, Umeå 12, 2–42.

Johansson, E., 2009. The history of literacy in Sweden, in comparison with some other countries. In: Graff, H.J., Mackinnon, A., Sandin, B., Winchester, I. (Eds.), Understanding Literacy in its Historical Contexts: Socio-cultural History and the Legacy of Egil Johansson. Nordic Academic Press, Lund, pp. 28–59.

Kaak, H., 1991. Die Gutsherrschaft: theoriegeschichtliche Untersuchungen zum Agrarwesen im ostelbischen Raum. Walter de Gruyter, Berlin/New York.

Kaal, H., Van Lottum, J., 2009. Immigrants in the Polder. Rural-rural long distance migration in north-western Europe: the case of Watergraafsmeer. Rural History 20, 99–117.

Kadens, E., 2012. The myth of the customary law merchant. Texas Law Review 90 (5), 1153–1206.

Kaminski, A., 1975. Neo-serfdom in Poland-Lithuania. Slavic Review 34 (2), 253–268.

Katele, I.B., 1986. Captains and corsairs: Venice and piracy, 1261–1381. University of Illinois at Urbana-Champaign, Ph.D. Dissertation.

Katz, A., 1996. Taking private ordering seriously. University of Pennsylvania Law Review 144 (5), 1745–1763.

Katz, E.D., 2000. Private order and public institutions: comments on McMillan and Woodruff's "Private order under dysfunctional public order". Michigan Law Review 98 (8), 2481–2493.

Klein, A., 2013. The institutions of the second serfdom and economic efficiency: review of the existing evidence for Bohemia. In: Cavaciocchi, S. (Ed.), Schiavitu e servaggio nell'economia europea. Secc. XI-XVIII./Slavery and Serfdom in the European Economy from the 11th to the 18th Centuries. XLV settimana di studi della Fondazione istituto internazionale di storia economica F. Datini, Prato 14–18 April 2013. Firenze University Press, Florence.

Klein, A., Ogilvie, S., 2013. Occupational Structure in the Czech Lands under the Second Serfdom. CESifo Working Papers.

Knapp, G.F., 1887. Die Bauernbefreiung und der Ursprung der Landarbeiter in den älteren Theilen Preußens. Duncker und Humblot, Leipzig.

Knight, J., 1995. Models, interpretations, and theories: constructing explanations of institutional emergence and change. In: Knight, J., Sened, I. (Eds.), Explaining Social Institutions. University of Michigan Press, Ann Arbor, MI, pp. 95–119.

Koch, H.W., 1990. Brandenburg-Prussia. In: Miller, J. (Ed.), Absolutism in Seventeenth-Century Europe. Macmillan, Basingstoke, pp. 123–155.

Koenigsberger, H.G., 2001. Monarchies, States Generals and Parliaments: The Netherlands in the Fifteenth and Sixteenth Centuries. Cambridge University Press, Cambridge.

Kollmer-von Oheimb-Loup, G., 2012. Die Entwicklung der Wirtschaftsstruktur am Mittleren Neckar 1800 bis 1950. Zeitschrift für Württembergische Landesgeschichte 71, 351–383.

Kopsidis, M., 2006. Agrarentwicklung: historische Agrarrevolutionen und Entwicklungsökonomie. Steiner, Stuttgart.

Kula, W., 1976. An Economic Theory of the Feudal System: Towards a Model of the Polish Economy. NLB, London.

Kussmaul, A., 1981. Servants in Husbandry in Early Modern England. Cambridge University Press, Cambridge.

Kussmaul, A., 1994. The pattern of work as the eighteenth century began. In: Floud, R., McCloskey, D.N. (Eds.), The Economic History of Britain Since 1700, vol. 1. Cambridge University Press, Cambridge, pp. 1–11.

Laiou, A.E., 2001. Byzantine trade with Christians and Muslims and the Crusades. In: Laiou, A.E., Mottahedeh, R.P. (Eds.), The Crusades from the Perspective of Byzantium and the Muslim World. Dumbarton Oaks Research Library and Collection, Washington, DC, pp. 157–196.

Lambert, S., 1990. Committees, religion, and parliamentary encroachment on royal authority in early Stuart England. English Historical Review 105 (414), 60–95.

Lambert, B., Stabel, P., 2005. Squaring the circle: merchant firms, merchant guilds, urban infrastructure and political authority in late medieval Bruges. Paper presented at the Workshop on Mercantile Organization in Pre-Industrial Europe. Antwerp, 18–19 November 2005.

Lambrecht, T., 2009. Rural credit and the market for annuities in eighteenth-century Flanders. In: Schofield, P.R., Lambrecht, T. (Eds.), Credit and the Rural Economy in North-Western Europe, c. 1200-c.1850. Brepols, Turnhout, pp. 75–98.

Lane, F.C., 1963. Venetian merchant galleys, 1300–1334: private and communal operation. Speculum 38, 179–205.

Laslett, P., 1988. The European family and early industrialization. In: Baechler, J., Hall, J.A., Mann, M. (Eds.), Europe and the Rise of Capitalism. Basil Blackwell, Oxford, pp. 234–242.

Laurent, H., 1935. Un grand commerce d'exportation au moyen âge: la draperie des Pays Bas en France et dans les pays mediterranéens, XIIe - XVe siècle. E. Droz, Paris.

Lewis, W.A., 1954. Economic development with unlimited supplies of labour. Manchester School of Economics and Social Studies 22, 139–191.

Lewis, W.A., 1958. Unlimited labour: further notes. Manchester School of Economics and Social Studies 26, 1–32.

Lindberg, E., 2008. The rise of Hamburg as a global marketplace in the seventeenth century: a comparative political economy perspective. Comparative Studies in Society and History 50 (3), 641–662.

Lindberg, E., 2009. Club goods and inefficient institutions: why Danzig and Lübeck failed in the early modern period. Economic History Review 62 (3), 604–628.

Lindberg, E., 2010. Merchant guilds in Hamburg and Königsberg: a comparative study of urban institutions and economic development in the early modern period. Journal of European Economic History 39 (1), 33–66.

Lindert, P.H., 2004. Growing Public: Social Spending and Economic Growth Since the Eighteenth Century. Cambridge University Press, Cambridge.

Lis, C., Soly, H., 1996. Ambachtsgilden in vergelijkend perspectief: de Noordelijke en de Zuidelijke Nederlanden, 15de–18de eeuw. In: Lis, C., Soly, H. (Eds.), Werelden van verschil: ambachtsgilden in de Lage Landen. Brussels, pp. 11–42.

Little, C.B., Sheffield, C.P., 1983. Frontiers and criminal justice: English private prosecution societies and American vigilantism in the eighteenth and nineteenth centuries. American Sociological Review 48 (6), 796–808.

Lloyd, T.H., 1977. The English Wool Trade in the Middle Ages. Cambridge University Press, Cambridge.

Lopez, R.S., 1987. The trade of medieval Europe: the south. In: Postan, M.M., Miller, E. (Eds.), The Cambridge Economic History of Europe, vol. 3: Economic Organization and Policies in the Middle Ages. Cambridge University Press, Cambridge, pp. 306–401.

Lopez, R.S., Raymond, I.W., 1955. Medieval Trade in the Mediterranean World. Columbia University Press, New York.

Macaulay, S., 1963. Non-contractual relations in business: a preliminary study. American Sociological Review 28 (1), 55–67.

Macfarlane, A., 1978. The Origins of English Individualism: the Family, Property and Social Transition. Blackwell, Oxford.

Mączak, A., 1997. Polen-Litauen als Paradoxon: Erwägungen über die Staatlichkeit des frühmodernen Polen. In: Lubinski, A., Rudert, T., Schattkowsky, M. (Eds.), Historie und Eigen-Sinn. Festschrift für Jan Peters zum 65. Geburtstag, Böhlau, Weimar, pp. 87–92.

Maggi, G., 1999. The role of multilateral institutions in international trade cooperation. American Economic Review 89 (1), 190–214.

Mammen, K., Paxson, C., 2000. Women's work and economic development. Journal of Economic Perspectives 14, 141–164.

Margariti, R.E., 2007. Aden and the Indian Ocean Trade: 150 Years in the Life of a Medieval Arabian Port. University of North Carolina Press, Chapel Hill, NC.

Mas-Latrie, R. de, 1866. Du droit de marque ou droit de représailles au Moyen Âge [premier article]. Bibliothèque de l'école des chartes 27, 529–577.

Mathias, P., O'Brien, P., 1976. Taxation in Britain and France, 1715–1810: a comparison of the social and economic incidence of taxes collected for central government. Journal of European Economic History 5, 601–650.

Mathias, P., O'Brien, P., 1978. The incidence of taxes and the burden of proof. Journal of European Economic History 7, 211–213.

McCloskey, D., 1976. English open fields as behavior towards risk. Research in Economic History 1, 124–170.

McCloskey, D., 1991. The prudent peasant: new findings on open fields. Journal of Economic History 51 (2), 343–355.

McCloskey, D., 2010. Bourgeois Dignity: Why Economics Can't Explain the Modern World. University of Chicago Press, Chicago.

McCord, N., 1958. The Anti-Corn Law League, 1838–1846. Allen & Unwin, London.

McLean, P.D., 2004. Widening access while tightening control: office-holding, marriages, and elite consolidation in early modern Poland. Theory and Society 33 (2), 167–212.

McLean, P., Padgett, J.F., 1997. Was Florence a perfectly competitive market? Transactional evidence from the Renaissance. Theory and Society 26 (2–3), 209–244.

McMillan, J., Woodruff, C., 2000. Private order under dysfunctional public order. Michigan Law Review 98 (8), 2421–2458.

Medick, H., 1996. Weben und Überleben in Laichingen 1650–1900. Untersuchungen zur Sozial-, Kultur- und Wirtschaftsgeschichte aus der Perspektive einer lokalen Gesellschaft im frühneuzeitlichen Württemberg. Vandenhoeck & Ruprecht, Göttingen.

Meiners, C., 1794. Bemerkungen auf einer Herbstreise nach Schwaben. Geschrieben im November 1793. In: Meiners (Ed.), Kleinere Länder- und Reisebeschreibungen, vol. 2, Spener, Berlin, pp. 235–380.

Melton, E., 1988. Gutsherrschaft in East Elbian Germany and Livonia, 1500–1800: a critique of the model. Central European History 21 (4), 315–349.

Melton, E., 1990. Enlightened seigniorialism and its dilemmas in serf Russia, 1750–1830. Journal of Modern History 62 (4), 675–708.

Micheletto, B.Z., 2011. Reconsidering the southern Europe model: dowry, women's work and marriage patterns in pre-industrial urban Italy (Turin, second half of the 18th century). The History of the Family 16 (4), 354–370.

Middleton, N., 2005. Early medieval port customs, tolls and controls on foreign trade. Early Medieval Europe 13 (4), 313–358.

Miguel, E., Gertler, P., Levine, D., 2005. Does social capital promote industrialization? Evidence from a rapid industrializer. Review of Economics and Statistics 87 (4), 754–762.

Milgrom, P.R., Roberts, J.F., 1992. Economics, Organization and Management. Prentice-Hall, Englewood Cliffs, NJ.

Milgrom, P.R., North, D.C., Weingast, B.R., 1990. The role of institutions in the revival of trade: the medieval law merchant, private judges and the Champagne fairs. Economics and Politics 2 (1), 1–23.

Mingay, G.E., 1963. The agricultural revolution in English history: a reconsideration. Agricultural History 37 (3), 123–133.

Mokyr, J., 1974. The Industrial Revolution in the Low Countries in the first half of the nineteenth century: a comparative case study. Journal of Economic History 34 (2), 365–391.

Mokyr, J., 1980. Industrialization and poverty in Ireland and the Netherlands. Journal of Interdisciplinary History 10 (3), 429–458.

Mokyr, J., 1987. Has the Industrial Revolution been crowded out? Some reflections on Crafts and Williamson. Explorations in Economic History 24 (3), 293–319.

Mokyr, J., 2009. The Enlightened Economy: An Economic History of Britain, 1700–1850. Princeton University Press, Princeton, NJ.

Muldrew, C., 1993. Credit and the courts: debt litigation in a seventeenth-century urban community. Economic History Review 46 (1), 23–38.

Muldrew, C., 1998. The Economy of Obligation: The Culture of Credit and Social Relations in Early Modern England. St. Martin's Press, New York/Basingstoke.

Muldrew, C., 2003. "A mutual assent of her mind"? Women, debt, litigation and contract in early modern England. History Workshop Journal 55 (1), 47–71.

Munro, J., 1999. The Low Countries' export trade in textiles with the Mediterranean Basin, 1200–1600: a cost-benefit analysis of comparative advantages in overland and maritime trade routes. International Journal of Maritime History 11 (2), 1–30.

Munro, J., 2001. The "new institutional economics" and the changing fortunes of fairs in medieval and early modern Europe: the textile trades, warfare, and transaction costs. Vierteljahrschrift für Sozial- und Wirtschaftsgeschichte 88 (1), 1–47.

Munzinger, M.R., 2006. The profits of the Cross: merchant involvement in the Baltic Crusade (c. 1180–1230). Journal of Medieval History 32 (2), 163–185.

Murrell, P., 2009. Design and Evolution in Institutional Development: The Insignificance of the English Bill of Rights. University of Maryland Department of Economics Working Paper, 13 December 2009.

Nachbar, T.B., 2005. Monopoly, mercantilism, and the politics of regulation. Virginia Law Review 91 (6), 1313–1379.

Neeson, J.M., 1984. The opponents of enclosure in eighteenth-century Northamptonshire. Past & Present 105, 114–139.

Neeson, J.M., 1993. Commoners: Common Right, Enclosure and Social Change in England, 1700–1820. Cambridge University Press, Cambridge.

Neeson, J.M., 2000. English enclosures and British peasants: current debates about rural social structure in Britain c. 1750–1870. Jahrbuch für Wirtschaftsgeschichte 2000 (2), 17–32.

Nelson, L.H. (Ed.), 1996. Liber de restauratione Monasterii Sancti Martini Tornacensis: the restoration of the Monastery of Saint Martin of Tournai, by Herman of Tournai. Catholic University of America Press, Washington, DC.

Nicolini, E.A., 2007. Was Malthus right? A VAR analysis of economic and demographic interactions in pre-industrial England. European Review of Economic History 11 (1), 99–121.

North, D.C., 1981. Structure and Change in Economic History. Norton, New York/London.

North, D.C., 1989. Institutions and economic growth: an historical introduction. World Development 17 (9), 1319–1332.

North, D.C., 1991. Institutions, transaction costs, and the rise of merchant empires. In: Tracy, J.D. (Ed.), The Political Economy of Merchant Empires: State Power and World Trade, 1350–1750. Cambridge University Press, Cambridge, pp. 22–40.

North, M., 2013. Serfdom and corvée labour in the Baltic area 16th-18th centuries. In: Cavaciocchi, S. (Ed.), Schiavitu e servaggio nell'economia europea. Secc. XI-XVIII./Slavery and serfdom in the European economy from the 11th to the 18th centuries. XLV settimana di studi della Fondazione istituto internazionale di storia economica F. Datini, Prato 14–18 April 2013. Firenze University Press, Florence.

North, D.C., Thomas, R.P., 1970. An economic theory of the growth of the western world. Economic History Review 2nd Ser. 23, 1–18.

North, D.C., Thomas, R.P., 1971. The rise and fall of the manorial system: a theoretical model. Journal of Economic History 31 (4), 777–803.

North, D.C., Thomas, R.P., 1973. The Rise of the Western World. Cambridge University Press, Cambridge.

North, D.C., Weingast, B.R., 1989. Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century England. Journal of Economic History 49 (4), 803–832.

North, D.C., Wallis, J.J., Weingast, B.R., 2006. A Conceptual Framework for Interpreting Recorded Human History. NBER Working Papers 12795.

North, D.C., Wallis, J.J., Weingast, B.R., 2009. Violence and Social Orders. A Conceptual Framework for Interpreting Recorded Human History. Cambridge University Press, Cambridge.

Nowak, J.E., Rotunda, R.D., 2004. Constitutional Law. Thomson/West, St. Paul, MI.

O'Brien, J.G., 2002. In defense of the mystical body: Giovanni da Legnano's theory of reprisals. Roman Legal Tradition 1, 25–55.

O'Brien, P.K., 1988. The political economy of British taxation, 1660–1815. Economic History Review 41 (1), 1–32.

O'Brien, P.K., 2001. Fiscal Exceptionalism: Great Britain and its European Rivals from Civil War to Triumph at Trafalgar and Waterloo. LSE Department of Economic History Working Paper 65/01.

O'Brien, P.K., Engerman, S.L., 1991. Exports and the growth of the British economy from the Glorious Revolution to the Peace of Amiens. In: Solow, B.L. (Ed.), Slavery and the Rise of the Atlantic Systems. Cambridge University Press, Cambridge, pp. 177–209.

O'Driscoll, G.P., Hoskins, L., 2006. The case for market-based regulation. Cato Journal 26, 469–487.

Ogilvie, S., 1986. Coming of age in a corporate society: capitalism, Pietism and family authority in rural Württemberg, 1590–1740. Continuity and Change 1 (3), 279–331.

Ogilvie, S., 1992. Germany and the seventeenth-century crisis. Historical Journal 35, 417–441.

Ogilvie, S., 1995. Population Growth and State Policy in Central Europe Before Industrialization. Centre for History and Economics Working Paper.

Ogilvie, S., 1997. State Corporatism and Proto-industry: The Württemberg Black Forest, 1580–1797. Cambridge University Press, Cambridge.

Ogilvie, S., 1999. The German state: a non-Prussian view. In: Hellmuth, E., Brewer, J. (Eds.), Rethinking Leviathan: The Eighteenth-Century State in Britain and Germany. Oxford University Press, Oxford, pp. 167–202.

Ogilvie, S., 2000. The European economy in the eighteenth century. In: Blanning, T.W.C. (Ed.), The Short Oxford History of Europe, vol. XII: The Eighteenth Century: Europe 1688–1815. Oxford University Press, Oxford, pp. 91–130.

Ogilvie, S., 2001. The economic world of the Bohemian serf: economic concepts, preferences and constraints on the estate of Friedland, 1583–1692. Economic History Review 54, 430–453.

Ogilvie, S., 2003. A Bitter Living: Women, Markets, and Social Capital in Early Modern Germany. Oxford University Press, Oxford.

Ogilvie, S., 2004a. Guilds, efficiency and social capital: evidence from German proto-industry. Economic History Review 57 (2), 286–333.

Ogilvie, S., 2004b. How does social capital affect women? Guilds and communities in early modern Germany. American Historical Review 109 (2), 325–359.

Ogilvie, S., 2004c. Women and labour markets in early modern Germany. Jahrbuch für Wirtschaftsgeschichte 2004 (2), 25–60.

Ogilvie, S., 2005a. Communities and the second serfdom in early modern Bohemia. Past & Present 187, 69–119.

Ogilvie, S., 2005b. Staat und Untertanen in der lokalen Gesellschaft am Beispiel der Herrschaft Frýdlant (Böhmen). In: Cerman, M., Luft, R. (Eds.), Untertanen, Herrschaft und Staat in Böhmen und im "Alten Reich". Sozialgeschichtliche Studien zur Frühen Neuzeit. Oldenbourg, Munich, pp. 51–86.

Ogilvie, S., 2005c. The use and abuse of trust: the deployment of social capital by early modern guilds. Jahrbuch für Wirtschaftsgeschichte 2005 (1), 15–52.

Ogilvie, S., 2005d. Village community and village headman in early modern Bohemia. Bohemia 46 (2), 402–451.

Ogilvie, S., 2006. "So that every subject knows how to behave": social disciplining in early modern Bohemia. Comparative Studies in Society and History 48 (1), 38–78.

Ogilvie, S., 2007a. Can We Rehabilitate the Guilds? A Sceptical Re-appraisal. Cambridge Working Papers in Economics 0745.

Ogilvie, S., 2007b. "Whatever is, is right"? Economic institutions in pre-industrial Europe. Economic History Review 60 (4), 649–684.

Ogilvie, S., 2008. Rehabilitating the guilds: a reply. Economic History Review 61 (1), 175–182.

Ogilvie, S., 2010. Consumption, social capital, and the "industrious revolution" in early modern Germany. Journal of Economic History 70 (2), 287–325.

Ogilvie, S., 2011. Institutions and European Trade: Merchant Guilds, 1000–1800. Cambridge University Press, Cambridge.

Ogilvie, S., 2012. Choices and Constraints in the Pre-industrial Countryside. Cambridge Working Papers in Economic and Social History (CWPESH) 0001.

Ogilvie, S., 2013a. Married women, work and the law: evidence from early modern Germany. In: Beattie, C., Stevens, M. (Eds.), Married Women and the Law in Northern Europe c.1200-1800. Boydell and Brewer, Woodbridge, pp. 213–239.

Ogilvie, S., 2013b. Serfdom and the institutional system in early modern Germany. In: Cavaciocchi, S. (Ed.), Schiavitu e servaggio nell'economia europea. Secc. XI-XVIII./Slavery and Serfdom in the European Economy from the 11th to the 18th Centuries. XLV settimana di studi della Fondazione istituto internazionale di storia economica F. Datini, Prato 14–18 April 2013. Firenze University Press, Florence.

Ogilvie, S., Edwards, J.S.S., 2000. Women and the "second serfdom": evidence from early modern Bohemia. Journal of Economic History 60 (4), 961–994.

Ogilvie, S., Küpker, M., Maegraith, J., 2011. Krämer und ihre Waren im ländlichen Württemberg zwischen 1600 und 1740. Zeitschrift für Agrargeschichte und Agrarsoziologie 59 (2), 54–75.

Ogilvie, S., Küpker, M., Maegraith, J., 2012. Household debt in early modern Germany: evidence from personal inventories. Journal of Economic History 72 (1), 134–167.

Ó Gráda, C., Chevet, J.M., 2002. Famine and market in *ancien régime* France. Journal of Economic History 62 (3), 706–733.

Olson, M., 1993. Dictatorship, democracy, and development. American Political Science Review 87 (3), 567–576.

Olsson, M., Svensson, P., 2009. Peasant economy - markets and agricultural production in southern Sweden 1711–1860. In: Pinilla Navarro, V. (Ed.), Markets and Agricultural Change in Europe from the 13th to the 20th Century. Brepols, Turnhout, pp. 75–106.

Olsson, M., Svensson, P., 2010. Agricultural growth and institutions: Sweden, 1700–1860. European Review of Economic History 14 (2), 275–304.

O'Rourke, K.H., Prados de la Escosura, L., Daudin, G., 2010. Trade and empire. In: Broadberry, S., O'Rourke, K.H. (Eds.), The Cambridge Economic History of Modern Europe, vol. 1: 1700–1870. Cambridge University Press, Cambridge, pp. 96–121.

Ostrom, E., 1998. A behavioral approach to the rational choice theory of collective action: presidential address, American Political Science Association, 1997. American Political Science Review 92 (1), 1–22.

Overton, M., 1996a. Agricultural Revolution in England: The Transformation of the Agrarian Economy 1500–1850. Cambridge University Press, Cambridge.

Overton, M., 1996b. Re-establishing the English agricultural revolution. Agricultural History Review 43 (1), 1–20.

Paravicini, W., 1992. Bruges and Germany. In: Vermeersch, V. (Ed.), Bruges and Europe. Mercatorfonds, Antwerp, pp. 99–128.

Parente, S.L., Prescott, E.C., 2000. Barriers to Riches. MIT Press, Cambridge, MA.

Parente, S.L., Prescott, E.C., 2005. A unified theory of the evolution of international income levels. In: Aghion, P., Durlauf, S.N. (Eds.), Handbook of Economic Growth, vol. 1, Part B. Elsevier, Amsterdam/London, pp. 1371–1416.

Peet, R., 1972. Influences of the British market on agriculture and related economic development in Europe before 1860. Transactions of the Institute of British Geographers 56, 1–20.

Pérez Moreda, V., 1997. La péninsule Ibérique: I. La population espagnole à l'époque moderne (XVIe-XVIIIe siècle). In: Bardet, J.-P., Dupâquier, J. (Eds.), Histoire des populations de l'Europe, vol. 1. Fayard, Paris, pp. 463–479.

Pérotin-Dumon, A., 1991. The pirate and the emperor: power and the law on the seas. In: Tracy, J.D. (Ed.), The Political Economy of Merchant Empires: State Power and World Trade, 1350–1750. Cambridge University Press, Cambridge, pp. 196–227.

Peters, J., 1995a. Inszenierung von Gutsherrschaft im 16. Jahrhundert: Matthias v. Saldern auf Plattenburg-Wilsnack (Prignitz). In: Peters, J. (Ed.), Konflikt und Kontrolle in Gutsherrschaftsgesellschaften: über Resistenz- und Herrschaftsverhalten in ländlichen Sozialgebilden der frühen Neuzeit. Vandenhoeck & Ruprecht, Göttingen, pp. 248–286.

Peters, J. (Ed.), 1995b. Konflikt und Kontrolle in Gutsherrschaftsgesellschaften: über Resistenz- und Herrschaftsverhalten in ländlichen Sozialgebilden der frühen Neuzeit. Vandenhoeck & Ruprecht, Göttingen.

Peters, J., 1997. Die Herrschaft Plattenburg-Wilsnack im Dreißigjährigen Krieg – eine märkische Gemeinschaft des Durchkomments. In: Beck, F., Neitmann, K. (Eds.), Brandenburgische Landesgeschichte und Archivwissenschaft: Festschrift für Lieselott Enders zum 70. Geburtstag. Verlag Hermann Böhlaus Nachfolger, Weimar, pp. 157–170.

Planitz, H., 1919. Studien zur Geschichte des deutschen Arrestprozesses, II. Kapital: Der Fremdenarrest. Zeitschrift der Savigny-Stiftung für Rechtsgeschichte, Germanistische Abteilung 40, 87–198.

Pocock, J.G.A., 2010. The Atlantic republican tradition: the republic of the seven provinces. Republics of Letters: A Journal for the Study of Knowledge, Politics, and the Arts 2 (1), 1–10.

Pollock, F., Maitland, F.W., 1895. The History of English Law Before the Time of Edward I. Cambridge University Press, Cambridge.

Pomeranz, K., 2000. The Great Divergence: Europe, China, and the Making of the Modern World Economy. Princeton University Press, Princeton, NJ.

Postan, M.M., 1966. Medieval agrarian society in its prime: England. In: Postan, M.M. (Ed.), The Cambridge Economic History of Europe, vol. 1: The Agrarian Life of the Middle Ages. Cambride University Press, Cambridge, pp. 548–632.

Prest, J., 1977. Politics in the Age of Cobden. Macmillan, London.

Price, J.M., 1991. Transaction costs: a note on merchant credit and the organization of private trade. In: Tracy, J.D. (Ed.), The Political Economy of Merchant Empires: State Power and World Trade, 1350–1750. Cambridge University Press, Cambridge, pp. 276–297.

Price, W.H., 2006. The English Patents of Monopoly. Harvard University Press, Cambridge, MA.

Puttevils, J., 2009. Relational and institutional trust in the international trade of the Low Countries, 15th – 16th centuries. Paper presented at the N.W. Posthumus Institute work in progress seminar. Amsterdam, 16–17 April 2009.

Rabb, T.K., 1964. Sir Edwyn Sandys and the parliament of 1604. American Historical Review 69 (3), 646–670.

Ranis, G., Fei, J.C.H., 1961. A theory of economic development. American Economic Review 51 (4), 533–565.

Rasmussen, C.P., 2013. Forms of serfdom and bondage in the Danish monarchy. In: Cavaciocchi, S. (Ed.), Schiavitu e servaggio nell'economia europea. Secc. XI-XVIII./Slavery and Serfdom in the European Economy from the 11th to the 18th Centuries. XLV settimana di studi della Fondazione istituto internazionale di storia economica F. Datini, Prato 14–18 April 2013. Firenze University Press, Florence.

Ray, D., 1998. Development Economics. Princeton University Press, Princeton, NJ.

Reher, D.S., 1998a. Family ties in Western Europe: persistent contrasts. Population and Development Review 24 (2), 203–234.

Reher, D.S., 1998b. Le Monde ibérique: I. L'Espagne. In: Bardet, J.-P., Dupâquier, J. (Eds.), Histoire des populations de l'Europe, vol. 2. Fayard, Paris, pp. 533–553.

Reis, J., 2005. Economic growth, human capital formation and consumption in western Europe before 1800. In: Allen, R.C., Bengtsson, T., Dribe, M. (Eds.), Living Standards in the Past: New Perspectives on Well-being in Asia and Europe. Oxford University Press, Oxford, pp. 195–225.

Reyerson, K.L., 1985. Business, Banking and Finance in Medieval Montpellier. Pontifical Institute of Mediaeval Studies, Toronto.

Reyerson, K.L., 2003. Commercial law and merchant disputes: Jacques Coeur and the law of marque. Medieval Encounters 9 (2–3), 244–255.

Richardson, G., 2005. The prudent village: risk pooling institutions in medieval English agriculture. Journal of Economic History 65 (2), 386–413.

Richman, B.D., 2004. Firms, courts, and reputation mechanisms: towards a positive theory of private ordering. Columbia Law Review 104 (8), 2328–2368.

Röhm, H., 1957. Die Vererbung des landwirtschaftlichen Grundeigentums in Baden-Württemberg. Bundesanstalt für Landeskunde, Remagen am Rhein.

Romer, P.M., 1987. Growth based on increasing returns due to specialization. American Economic Review 77 (2), 56–62.

Romer, P.M., 1990. Endogenous technological change. Journal of Political Economy 98 (5), S71–S102.

Rosenberg, H., 1958. Bureaucracy, Aristocracy and Autocracy: The Prussian Experience, 1600–1815. Harvard University Press, Cambridge, MA.

Rudert, T., 1995a. Gutsherrschaft und Agrarstruktur: der ländliche Bereich Mecklenburgs am Beginn des 18. Jahrhunderts. P. Lang, Frankfurt am Main/New York.

Rudert, T., 1995b. Gutsherrschaft und ländliche Gemeinde. Beobachtungen zum Zusammenhang von gemeindlicher Autonomie und Agrarverfassung in der Oberlausitz im 18. Jahrhundert. In: Peters, J. (Ed.), Gutsherrschaft als soziales Modell. Vergleichende Betrachtungen zur Funktionsweise frühneuzeitlicher Agrargesellschaften. Oldenbourg, Munich, pp. 197–218.

Sabean, D.W., 1990. Property, Production and Family in Neckarhausen, 1700–1870. Cambridge University Press, Cambridge.

Sachs, J.D., 2001. Tropical Underdevelopment. NBER Working Paper 8119.

Sachs, J.D., 2003. Institutions Don't Rule: Direct Effects of Geography on Per Capita Income. NBER Working Paper 9490.

Sachs, S.E., 2006. From St. Ives to cyberspace: the modern distortion of the medieval "Law Merchant". American University International Law Review 21 (5), 685–812.

Sanderson, E.C., 1996. Women and Work in Eighteenth-Century Edinburgh. Macmillan, Basingstoke.

Say, J.B., 1817. Traité d'économie politique, ou, Simple exposition de la manière dont se forment, se distribuent et se consomment les richesses. Déterville, Paris.

Schmoller, G. von, 1888. Die Einführung der französischen Regie durch Friedrich den Großen 1766. Sitzungsberichte der preußischen Akademie der Wissenschaften 1, 63–79.

Schofield, P.R., Lambrecht, T., 2009. Introduction: credit and the rural economy in north-western Europe, c. 1200-c. 1800. In: Schofield, P.R., Lambrecht, T. (Eds.), Credit and the Rural Economy in North-Western Europe, c. 1200-c.1850. Brepols, Turnhout, pp. 1–18.

Schofield, R.S., 1963. Parliamentary Lay Taxation, 1485–1547. University of Cambridge, Ph.D. Dissertation.

Schofield, R.S., 2004. Taxation under the Early Tudors: 1485–1547. Blackwell, Oxford.

Schomerus, H., 1977. Die Arbeiter der Maschinenfabrik Esslingen. Forschungen zur Lage der Arbeiterschaft im 19. Jahrhundert. Ernst Klett Verlag, Stuttgart.

Schönfelder, A., 1988. Handelsmessen und Kreditwirtschaft im Hochmittelalter. Die Champagnemessen. Verlag Rita Dadder, Saarbrücken-Scheidt.

Schonhardt-Bailey, C., 2006. From the Corn Laws to Free Trade: Interests, Ideas, and Institutions in Historical Perspective. MIT Press, Cambridge, MA.

Schulte, A., 1900. Geschichte des mittelalterlichen Handels zwischen Westdeutschland und Italien mit Ausschluss von Venedig. Duncker und Humblot, Leipzig.

Selzer, S., Ewert, U.-C., 2005. Die neue Institutionenökonomik als Herausforderung an die Hanseforschung. Hansische Geschichtsblätter 123, 7–29.

Selzer, S., Ewert, U.C., 2010. Netzwerke im europäischen Handel des Mittelalters. Konzepte - Anwendungen - Fragestellungen. In: Fouquet, G., Gilomen, H.-J. (Eds.), Netzwerke im europäischen Handel des Mittelalters. Thorbecke, Ostfildern, pp. 21–48.

Semmel, B., 1970. The Rise of Free Trade Imperialism: Classical Political economy, the Empire of Free Trade and Imperialism, 1750–1850. Cambridge University Press, Cambridge.

Seppel, M., 2013. The growth of the state and its consequences on the structure of serfdom in the Baltic provinces, 1550–1750. In: Cavaciocchi, S. (Ed.), Schiavitu e servaggio nell'economia europea. Secc. XI-XVIII./Slavery and Serfdom in the European Economy from the 11th to the 18th Centuries. XLV settimana di studi della Fondazione istituto internazionale di storia economica F. Datini, Prato 14–18 April 2013. Firenze University Press, Florence.

Serrão, J.V., 2009. Land management responses to market changes. Portugal, seventeenth-nineteenth centuries. In: Pinilla Navarro, V. (Ed.), Markets and Agricultural Change in Europe from the 13th to the 20th Century. Brepols, Turnhout, pp. 47–74.

Sharp, P., Weisdorf, J., 2013. Globalization revisited: Market integration and the wheat trade between North America and Britain from the eighteenth century. Explorations in Economic History 50 (1), 88–98.

Sharpe, P., 1999. The female labour market in English agriculture during the Industrial Revolution: expansion or contraction? Agricultural History Review 47 (2), 161–181.

Shaw-Taylor, L., 2001a. Labourers, cows, common rights and parliamentary enclosure: the evidence of contemporary comment c. 1760–1810. Past & Present 171, 95–126.

Shaw-Taylor, L., 2001b. Parliamentary enclosure and the emergence of an English agricultural proletariat. Journal of Economic History 61 (3), 640–662.

Slicher van Bath, B.H., 1963. The Agrarian History of Western Europe, A.D. 500–1850. E. Arnold, London.

Slicher van Bath, B.H., 1977. Agriculture in the vital revolution. In: Rich, E.E., Wilson, C.H. (Eds.), The Cambridge Economic History of Europe: vol. 5: The Economic Organization of Early Modern Europe. Cambridge University Press, Cambridge, pp. 42–132.

Smith, A., 1776. An Inquiry into the Nature and Causes of the Wealth of Nations. W. Strahan and T. Cadell, London.

Smith, R.M., 1974. English peasant life-cycles and socio-economic networks: a quantitative geographical case study. University of Cambridge, Ph.D. Dissertation.

Smith, R.M., 1981a. Fertility, economy and household formation in England over three centuries. Population and Development Review 7 (4), 595–622.

Smith, R.M., 1981b. The people of Tuscany and their families in the fifteenth century: medieval or Mediterranean? Journal of Family History 6, 107–128.

Smith, R.M., 1996. A periodic market and its impact upon a manorial community: Botesdale, Suffolk, and the manor of Redgrave, 1280–1300. In: Smith, R.M. (Ed.), Razi, Z. Medieval Society and the Manor Court. Clarendon Press, Oxford, pp. 450–481.

Sobel, J., 2002. Can we trust social capital? Journal of Economic Literature 40 (1), 139–154.

Solar, P.M., 1995. Poor relief and English economic development before the Industrial Revolution. Economic History Review NS 48 (1), 1–22.

Sonnino, E., 1997. L'Italie: II. Le tournant du XVIIe siècle. In: Bardet, J.-P., Dupâquier, J. (Eds.), Histoire des populations de l'Europe, vol. 1. Fayard, Paris, pp. 496–508.

Sperber, J., 1985. State and civil society in Prussia: thoughts on a new edition of Reinhart Koselleck's "Preussen zwischen Reform und Revolution". Journal of Modern History 57 (2), 278–296.

Spufford, P., 2000. Long-term rural credit in sixteenth- and seventeenth-century England: the evidence of probate accounts. In: Arkell, T., Evans, N., Goose, N. (Eds.), When Death Do Us Part: Understanding and Interpreting the Probate Records of Early Modern England. Oxford University Press, Oxford, pp. 213–228.

Stabel, P., 1999. Venice and the Low Countries: commercial contacts and intellectual inspirations. In: Aikema, B., Brown, B.L. (Eds.), Renaissance Venice and the North: Crosscurrents in the Time of Bellini, Dürer and Titian, London, pp. 31–43.

Stasavage, D., 2002. Credible commitment in early modern Europe: North and Weingast revisited. Journal of Law, Economics and Organization 18 (1), 155–186.

Stillman, N.A., 1970. East-West relations in the Islamic Mediterranean in the early eleventh century: a study in the Geniza correspondence of the house of Ibn 'Awkal. University of Pennsylvania, Ph.D. Dissertation.

Stillman, N.A., 1973. The eleventh century merchant house of Ibn 'Awkal (a Geniza study). Journal of the Economic and Social History of the Orient 16 (1), 15–88.

Strayer, J.R., 1969. Italian bankers and Philip the Fair. In: Herlihy, D., Lopez, R.S., Slessarev, V. (Eds.), Economy, Society and Government in Medieval Italy: Essays in Memory of Robert L. Reynolds. Kent State University Press, Kent, OH, pp. 239–247.

Strayer, J.R., 1980. The Reign of Philip the Fair. Princeton University Press, Princeton.

Sussman, N., Yafeh, Y., 2006. Institutional reforms, financial development and sovereign debt: Britain 1690–1790. Journal of Economic History 66 (4), 906–935.

Swedberg, R., 2003. The case for an economic sociology of law. Theory and Society 32 (1), 1–37.

Szabó, T., 1983. Xenodochia, Hospitäler und Herbergen - kirchliche und kommerzielle Gastung im mittelalterlichen Italien (7. bis 14. Jahrhundert). In: Peyer, H.C., Müller-Luckner, E. (Eds.), Gastfreundschaft, Taverne und Gasthaus im Mittelalter. R. Oldenbourg, Munich/Vienna, pp. 61–92.

Tai, E.S., 1996. Honor among thieves: piracy, restitution, and reprisal in Genoa, Venice, and the Crown of Catalonia-Aragon, 1339–1417. Harvard University, Ph.D. Dissertation.

Tai, E.S., 2003a. Marking water: piracy and property in the pre-modern West. Paper presented at the conference on Seascapes, Littoral Cultures, and Trans-Oceanic Exchanges, Library of Congress, Washington DC, 12–15 February.

Tai, E.S., 2003b. Piracy and law in medieval Genoa: the *consilia* of Bartolomeo Bosco. Medieval Encounters 9 (2–3), 256–282.

Tardif, J., 1855. Charte française de 1230 conservée aux archives municipales de Troyes. Bibliothèque de l'école des chartes 16, 139–146.

Taylor, A.M., 2002. Globalization, Trade, and Development: Some Lessons from History. NBER Working Paper w9326.

Terrasse, V., 2005. Provins: une commune du comté de Champagne et de Brie (1152–1355). L'Harmattan, Paris.

'T Hart, M.C., 1989. Cities and statemaking in the Dutch republic, 1580–1680. Theory and Society 18 (5), 663–687.

'T Hart, M.C., 1993. The Making of a Bourgeois State: War, Politics and Finance during the Dutch Revolt. Manchester University Press, Manchester.

Theiller, I., 2009. Markets as agents of local, regional and interregional trade. Eastern Normandy at the end of the Middle Ages. In: Pinilla Navarro, V. (Ed.), Markets and Agricultural Change in Europe from the 13th to the 20th Century, Brepols, Turnhout, pp. 37–46.

Thoen, E., Soens, T., 2009. Credit in rural Flanders, c. 1250–c.1600: its variety and significance. In: Schofield, P.R., Lambrecht, T. (Eds.), Credit and the Rural Economy in North-Western Europe, c. 1200–c.1850. Brepols, Turnhout, pp. 19–38.

Thomas, R.P., McCloskey, D.N., 1981. Overseas trade and empire 1700–1860. In: Floud, R., McCloskey, D. (Eds.), The Economic History of Britain Since 1700, vol. 1. Cambridge University Press, Cambridge.

Tipton, F.B., 1976. Regional Variations in the Economic Development of Germany during the Nineteenth Century. Wesleyan University Press, Middletown, CT.

Toch, M., 2010. Netzwerke im jüdischen Handel des Früh- und Hochmittelalters?. In: Fouquet, G., Gilomen, H.-J. (Eds.), Netzwerke im europäischen Handel des Mittelalters. Thorbecke, Ostfildern, pp. 229–244.

Topolski, J., 1974. The manorial-serf economy in central and eastern Europe in the 16th and 17th centuries. Agricultural History 48 (3), 341–352.

Townsend, R.M., 1993. The Medieval Village Economy: A Study of the Pareto Mapping in General Equilibrium Models. Princeton University Press, Princeton, NJ.

Trivellato, F., 2009. The Familiarity of Strangers: The Sephardic Diaspora, Livorno, and Cross-cultural Trade in the Early Modern Period. Yale University Press, New Yaven, CT.

Troeltsch, W., 1897. Die Calwer Zeughandlungskompagnie und ihre Arbeiter. Studien zur Gewerbe- und Sozialgeschichte Altwürttembergs. Gustav Fischer, Jena.

Twarog, S., 1997. Heights and living standards in Germany, 1850–1939: the case of Württemberg. In: Steckel, R.H., Floud, R. (Eds.), Health and Welfare during Industrialization. University of Chicago Press, Chicago.

Udovitch, A.L., 1977a. Formalism and informalism in the social and economic institutions of the medieval Islamic world. In: Banani, A., Vryonis, S. (Eds.), Individualism and Conformity in Classical Islam. Undena Publications, Wiesbaden, pp. 61–81.

Udovitch, A.L., 1977b. A tale of two cities: commercial relations between Cairo and Alexandria during the second half of the eleventh century. In: Miskimin, H.A., Herlihy, D., Udovitch, A.L. (Eds.), The Medieval City. Yale University Press, New Haven, CT, pp. 143–162.

Ulbrich, C., 2004. Shulamit and Margarete: Power, Gender, and Religion in a Rural Society in Eighteenth-Century Europe. Brill Academic Publishers, Boston.

Vamplew, W., 1980. The protection of English cereal producers: the Corn Laws reassessed. Economic History Review 33 (3), 382–395.

Van Bavel, B.J.P., 2010. Manors and Markets: Economy and Society in the Low Countries, 500–1600. Oxford University Press, Oxford.

Van Bavel, B.J.P., 2011. Markets for land, labor, and capital in northern Italy and the Low Countries, twelfth to seventeenth centuries. Journal of Interdisciplinary History 41 (4), 503–531.

Van Bavel, B.J.P., Van Zanden, J.L., 2004. The jump-start of the Holland economy during the late-medieval crisis, c. 1350–c. 1500. Economic History Review 57, 503–532.

Van Cruyningen, P., 2009. Credit and agriculture in the Netherlands, eighteenth - nineteenth centuries. In: Schofield, P.R., Lambrecht, T. (Eds.), Credit and the Rural Economy in North-Western Europe, c. 1200-c. 1850. Brepols, Turnhout, pp. 99–108.

Van den Heuvel, D., 2007. Women and Entrepreneurship: Female Traders in the Northern Netherlands, c. 1580–1815. Aksant, Amsterdam.

Van den Heuvel, D., 2008. Partners in marriage and business? Guilds and the family economy in urban food markets in the Dutch Republic. Continuity and Change 23 (2), 217–236.

Van den Heuvel, D., Ogilvie, S., 2013. Retail development in the Consumer Revolution: The Netherlands, c. 1670–c. 1815. Explorations in Economic History 50 (1), 69–87.

Van der Heijden, M., Van Nederveen Meerkerk, E., Schmidt, A., 2011. Women's and children's work in an industrious society: The Netherlands, 17th-19th centuries. In: Ammannati, F. (Ed.), Religione e istituzioni religiose nell'economia Europea. 1000–1800/Religion and religious institutions in the European economy, 1000–1800. Atti della Quarantatreesima Settimana di Studi 8–12 maggio 2011. Firenze University Press, Florence, pp. 543–562.

Van Doosselaere, Q., 2009. Commercial Agreements and Social Dynamics in Medieval Genoa. Cambridge University Press, Cambridge.

Van Lottum, J., 2011a. Labour migration and economic performance: London and the Randstad, c. 1600–1800. Economic History Review 64 (2), 531–570.

Van Lottum, J., 2011b. Some considerations about the link between economic development and migration. Journal of Global History 6 (2), 339–344.

Vann, J.A., 1984. The Making of a State: Württemberg, 1593–1793. Cornell University Press, Ithaca, NY.

Van Nederveen Meerkerk, E., 2006a. De draad in eigen handen. Vrouwen in loonarbeid in de Nederlandse textielnijverheid, 1581–1810. Vrije Universiteit, Amsterdam.

Van Nederveen Meerkerk, E., 2006b. Segmentation in the pre-industrial labour market: women's work in the Dutch textile industry, 1581–1810. International Review of Social History 51, 189–216.

Van Nederveen Meerkerk, E., 2010. Market wage or discrimination? The remuneration of male and female wool spinners in the seventeenth-century Dutch Republic. Economic History Review 63 (1), 165–186.

Van Zanden, J.L., 2001. Early modern economic growth: a survey of the European economy, 1500–1800. In: Prak, M. (Ed.), Early Modern Capitalism: Economic and Social Change in Europe 1400–1800. Routledge, London, pp. 69–87.

Van Zanden, J.L., 2009. The Long Road to the Industrial Revolution: The European Economy in a Global Perspective, 1000–1800. Brill, Leiden.

Van Zanden, J.L., Prak, M., 2006. Towards an economic interpretation of citizenship: the Dutch Republic between medieval communes and modern nation-states. European Review of Economic History 10 (2), 11–147.

Van Zanden, J.L., Van Leeuwen, B., 2012. Persistent but not consistent: the growth of national income in Holland 1347–1807. Explorations in Economic History 49 (2), 119–130.

Van Zanden, J.L., Van Riel, A., 2004. The Strictures of Inheritance: The Dutch Economy in the Nineteenth Century. Princeton University Press, Princeton, NJ.

Velková, A., 2012. The role of the manor in property transfers of serf holdings in Bohemia in the period of the "second serfdom". Social History 37 (4), 501–521.

Verlinden, C., 1965. Markets and fairs. In: Postan, M.M., Rich, E.E., Miller, E. (Eds.), The Cambridge Economic History of Europe, vol. 3: Economic Organization and Policies in the Middle Ages. Cambridge University Press, Cambridge, pp. 119–153.

Voigtländer, N., Voth, H.-J., 2006. Why England? Demographic factors, structural change and physical capital accumulation during the Industrial Revolution. Journal of Economic Growth 11 (4), 319–361.

Voigtländer, N., Voth, H.-J., 2010. How the West "Invented" Fertility Restriction. NBER Working Paper 17314.

Volckart, O., 2004. The economics of feuding in late medieval Germany. Explorations in Economic History 41, 282–299.

Volckart, O., Mangels, A., 1999. Are the roots of the modern *lex mercatoria* really medieval? Southern Economic Journal 65 (3), 427–450.

Wach, A., 1868. Der Arrestprozess in seiner geschichtlichen Entwicklung. 1. Teil: der italienischen Arrestprozess. Hässel, Leipzig.

Wang, F., Campbell, C., Lee, J.Z., 2010. Agency, hierarchies, and reproduction in northeastern China, 1749–1840. In: Tsuya, N.O., Wang, F., Alter, G., Lee, J.Z. (Eds.), Prudence and Pressure: Reproduction and Human Agency in Europe and Asia, 1700–1900. MIT Press, Cambridge, MA, pp. 287–316.

Ward, T., 2004. The Corn Laws and English wheat prices, 1815–1846. Atlantic Economic Journal 32 (3), 245–255.

Weir, D.R., 1984. Life under pressure: France and England, 1670–1870. Journal of Economic History 44 (1), 27–47.

Wheeler, N.C., 2011. The noble enterprise of state building: reconsidering the rise and fall of the modern state in Prussia and Poland. Comparative Politics 44 (1), 21–38.

Whittle, J., 1998. Individualism and the family-land bond: a reassessment of land transfer patterns among the English peasantry. Past & Present 160, 25–63.

Whittle, J., 2000. The Development of Agrarian Capitalism: Land and Labour in Norfolk, 1440–1580. Clarendon, Oxford.

Wiesner, M.E., 1989. Guilds, male bonding and women's work in early modern Germany. Gender & History 1 (1), 125–137.

Wiesner, M.E., 2000. Women and Gender in Early Modern Europe. Cambridge University Press, Cambridge.

Wiesner-Hanks, M.E., 1996. Ausbildung in den Zünften. In: Kleinau, E., Opitz, C. (Eds.), Geschichte der Mädchen- und Frauenbildung, vol. I: Vom Mittelalter bis zur Aufklärung. Campus Verlag, Campus, Frankfurt/New York, pp. 91–102.

Williams, D.T., 1931. The maritime relations of Bordeaux and Southampton in the thirteenth century. Scottish Geographical Journal 47 (5), 270–275.

Williamson, J.G., 1984. Why was British growth so slow during the Industrial Revolution? Journal of Economic History 44, 687–712.

Williamson, J.G., 1987. Debating the British Industrial Revolution. Explorations in Economic History 24 (3), 269–292.

Williamson, J.G., 1990. The impact of the Corn Laws just prior to repeal. Explorations in Economic History 27 (2), 123–156.

Woodward, R.L., 2005. Merchant guilds. In: Northrup, C.C. (Ed.), Encyclopedia of World Trade from Ancient Times to the Present, vol. 3. M.E. Sharpe, New York, pp. 631–638.

Woodward, R.L., 2007. Merchant guilds (*Consulados de Comercio*) in the Spanish world. History Compass 5 (5), 1576–1584.

World Bank, 1982. World Development Report 1982: Agriculture and Economic Development. Oxford University Press, Oxford.

World Bank, 2002. World Development Report 2002: Building Institutions for Markets. Oxford University Press, Oxford.

Wrightson, K., 1982. English Society 1580–1680. Hutchinson, London.

Wrightson, K., Levine, D., 1995. Poverty and Piety in an English Village: Terling, 1525–1700. Clarendon, Oxford.

Wunder, H., 1978. Peasant organization and class conflict in east and west Germany. Past & Present (78), 47–55.

Wunder, H., 1995. Das Selbstverständliche denken. Ein Vorschlag zur vergleichenden Analyse ländlicher Gesellschaften in der Frühen Neuzeit, ausgehend vom "Modell ostelbische Gutsherrschaft". In: Peters, J. (Ed.), Gutsherrschaft als soziales Modell. Vergleichende Betrachtungen zur Funktionsweise früh-neuzeitlicher Agrargesellschaften. Oldenbourg, Munich, pp. 23–49.

Wunder, H., 1996. Agriculture and agrarian society. In: Ogilvie, S. (Ed.), Germany: A New Social and Economic History, vol. II: 1630–1800. Edward Arnold, London, pp. 63–99.

CHAPTER ONE

What Do We Learn From Schumpeterian Growth Theory?

Philippe Aghion[*], **Ufuk Akcigit**[†], and **Peter Howitt**[‡]
[*]Harvard University, NBER, and CIFAR, USA
[†]University of Pennsylvania and NBER, USA
[‡]Brown University and NBER, USA

Abstract

Schumpeterian growth theory has operationalized Schumpeter's notion of creative destruction by developing models based on this concept. These models shed light on several aspects of the growth process that could not be properly addressed by alternative theories. In this survey, we focus on four important aspects, namely: (i) the role of competition and market structure; (ii) firm dynamics; (iii) the relationship between growth and development with the notion of appropriate growth institutions; and (iv) the emergence and impact of long-term technological waves. In each case, Schumpeterian growth theory delivers predictions that distinguish it from other growth models and which can be tested using micro data.

Keywords

Creative destruction, Entry, Exit, Competition, Firm dynamics, Reallocation, R&D, Industrial policy, Technological frontier, Schumpeterian wave, General-purpose technology

JEL Classification Codes

O10, O11, O12, O30, O31, O33, O40, O43, O47

1.1. INTRODUCTION

Formal models allow us to make verbal notions operational and confront them with data. The Schumpeterian growth theory surveyed in this paper has "operationalized" Schumpeter's notion of creative destruction—the process by which new innovations replace older technologies—in two ways. First, it has developed models based on creative destruction that shed new light on several aspects of the growth process. Second, it has used data, including rich micro data, to confront the predictions that distinguish it from other growth theories. In the process, the theory has improved our understanding of the underlying sources of growth.

Handbook of Economic Growth, Volume 2B
ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53540-5.00001-X

Over the past 25 years,[1] Schumpeterian growth theory has developed into an integrated framework for understanding not only the macroeconomic structure of growth but also the many microeconomic issues regarding incentives, policies, and organizations that interact with growth: who gains and who loses from innovations, and what the net rents from innovation are. These ultimately depend on characteristics such as property right protection; competition and openness; education; democracy; and so forth, and to a different extent in countries or sectors at different stages of development. Moreover, the recent years have witnessed a new generation of Schumpeterian growth models focusing on firm dynamics and reallocation of resources among incumbents and new entrants.[2] These models are easily estimable using micro firm-level datasets, which also bring the rich set of tools from other empirical fields into macroeconomics and endogenous growth.

In this paper, which aims to be accessible to readers with only basic knowledge in economics and is thus largely self-contained, we shall consider four aspects on which Schumpeterian growth theory delivers distinctive predictions.[3] First, the relationship between growth and industrial organization: faster innovation-led growth is generally associated with higher turnover rates, i.e. higher rates of creation and destruction, of firms and jobs; moreover, competition appears to be positively correlated with growth, and competition policy tends to complement patent policy. Second, the relationship between growth and firm dynamics: small firms exit more frequently than large firms; conditional on survival, small firms grow faster; there is a very strong correlation between firm size and firm age; and finally, firm size distribution is highly skewed. Third, the relationship between growth and development with the notion of appropriate institutions: namely, the idea that different types of policies or institutions appear to be growth-enhancing at different stages of development. Our emphasis will be on the relationship between growth and democracy and on why this relationship appears to be stronger in more frontier economies. Four, the relationship between growth and long-term technological waves: why such waves are associated with an increase in the flow of firm entry and exit; why they may initially generate a productivity slowdown; and why they may increase wage inequality both between and within educational groups. In each case, we show that

[1] The approach was initiated in the fall of 1987 at MIT, where Philippe Aghion was a 1-year assistant professor and Peter Howitt a visiting professor on sabbatical from the University of Western Ontario. During that year they wrote their "model of growth through creative destruction" (see Section 1.2 below), which was published as Aghion and Howitt (1992). Parallel attempts at developing Schumpeterian growth models include Segerstrom et al (1990) and Corriveau (1991).

[2] See Klette and Kortum (2004), Lentz and Mortensen (2008), Akcigit and Kerr (2010), and Acemoglu et al. (2013).

[3] Thus, we are not looking at the aspects or issues that could be addressed by the Schumpeterian model and by other models, including Romer's (1990) product variety model (see Aghion and Howitt, 1998, 2009). Grossman and Helpman (1991) were the first to point out the parallels between the two models, although using a special version of the Schumpeterian model.

Schumpeterian growth theory delivers predictions that distinguish it from other growth models and which can be tested using micro data.

The paper is organized as follows: Section 1.2 lays out the basic Schumpeterian model; Section 1.3 introduces competition and IO into the framework; Section 1.4 analyzes firm dynamics; Section 1.5 looks at the relationship between growth and development and in particular at the role of democracy in the growth process; Section 1.6 discusses technological waves; and Section 1.7 concludes.

A word of caution before we proceed: this paper focuses on the Schumpeterian growth paradigm and some of its applications. It is not a survey of the existing (endogenous) growth literature. There, we refer the reader to growth textbooks (e.g. Acemoglu, 2009; Aghion and Howitt, 1998, 2009; Barro and Sala-i-Martin, 2003; Galor, 2011; Jones and Vollrath, 2013; Weil, 2012).

1.2. SCHUMPETERIAN GROWTH: BASIC MODEL

1.2.1 The Setup

The following model borrows directly from the theoretical IO and patent race literature (see Tirole, 1988). This model is Schumpeterian in that: (i) it is about growth generated by innovations; (ii) innovations result from entrepreneurial investments that are themselves motivated by the prospects of monopoly rents; and (iii) new innovations replace old technologies: in other words, growth involves creative destruction.

Time is continuous and the economy is populated by a continuous mass L of infinitely lived individuals with linear preferences, that discount the future at rate ρ.[4] Each individual is endowed with one unit of labor per unit of time, which he or she can allocate between production and research: in equilibrium, individuals are indifferent between these two activities.

There is a final good, which is also the numeraire. The final good at time t is produced competitively using an intermediate input, namely:

$$Y_t = A_t y_t^{\alpha},$$

where α is between zero and one, y_t is the amount of the intermediate good currently used in the production of the final good, and A_t is the productivity—or quality—of the currently used intermediate input.[5]

The intermediate good y is in turn produced one for one with labor: that is, one unit flow of labor currently used in manufacturing the intermediate input produces one unit of intermediate input of frontier quality. Thus, y_t denotes both the current production of the

[4] The linear preferences (or risk neutrality) assumption implies that the equilibrium interest rate will always be equal to the rate of time preference: $r_t = \rho$ (see Aghion and Howitt, 2009, Chapter 2).

[5] In what follows we will use the words "productivity" and "quality" interchangeably.

intermediate input and the flow amount of labor currently employed in manufacturing the intermediate good.

Growth in this model results from innovations that improve the quality of the intermediate input used in the production of the final good. More formally, if the previous state-of-the-art intermediate good was of quality A, then a new innovation will introduce a new intermediate input of quality γA, where $\gamma > 1$. This immediately implies that growth will involve creative destruction, in the sense that Bertrand competition will allow the new innovator to drive the firm producing the intermediate good of quality A out of the market, since at the same labor cost the innovator produces a better good than that of the incumbent firm.[6]

The innovation technology is directly drawn from the theoretical IO and patent race literatures: namely, if z_t units of labor are currently used in R&D, then a new innovation arrives during the current unit of time at the (memoryless) Poisson rate λz_t.[7] Henceforth, we will drop the time index t, when it causes no confusion.

1.2.2 Solving the Model

1.2.2.1 The Research Arbitrage and Labor Market Clearing Equations

We shall concentrate our attention on balanced growth equilibria where the allocation of labor between production (y) and R&D (z) remains constant over time. The growth process is described by two basic equations.

The first is the labor market clearing equation:

$$L = y + z, \tag{L}$$

reflecting the fact that the total flow of labor supply during any unit of time is fully absorbed between production and R&D activities (i.e. by the demand for manufacturing and R&D labor).

[6] Thus, overall, growth in the Schumpeterian model involves both positive and negative externalities. The positive externality is referred to by Aghion and Howitt (1992) as a "knowledge spillover effect." Namely, any new innovation raises productivity A forever, i.e. the benchmark technology for any subsequent innovation. However, the current (private) innovator captures the rents from his or her innovation only during the time interval until the next innovation occurs. This effect is also featured in Romer (1990), where it is referred to instead as "non-rivalry plus limited excludability." But in addition, in the Schumpeterian model, any new innovation has a negative externality as it destroys the rents of the previous innovator. Following the theoretical IO literature, Aghion and Howitt (1992) refer to this as the "business-stealing effect" of innovation. The welfare analysis in that paper derives sufficient conditions under which the intertemporal spillover effect dominates or is dominated by the business-stealing effect. The equilibrium growth rate under laissez-faire is correspondingly suboptimal or excessive compared to the socially optimal growth rate.

[7] More generally, if z_t units of labor are invested in R&D during the time interval $[t, t + dt]$, the probability of innovation during this time interval is $\lambda z_t dt$.

The second equation reflects individuals' indifference in equilibrium between engaging in R&D or working in the intermediate good sector. We call it the research-arbitrage equation. The remaining part of the analysis consists of spelling out this research-arbitrage equation.

More formally, let w_k denote the current wage rate conditional on there having already been $k \in \mathbb{Z}_{++}$ innovations from time 0 until current time t (since innovation is the only source of change in this model, all other economic variables remain constant during the time interval between two successive innovations). And let V_{k+1} denote the net present value of becoming the next ($(k+1)$th) innovator.

During a small time interval dt, between the kth and $(k+1)$th innovations, an individual faces the following choices: Either she employs her (flow) unit of labor for the current unit of time in manufacturing at the current wage, in which case she gets $w_t dt$. Or she devotes her flow unit of labor to R&D, in which case she will innovate during the current time period with probability λdt and then get V_{k+1}, whereas she gets nothing if she does not innovate.[8] The research-arbitrage equation is then simply expressed as:

$$w_k = \lambda V_{k+1}. \tag{R}$$

The value V_{k+1} is in turn determined by a Bellman equation. We will use Bellman equations repeatedly in this survey; thus, it is worth going slowly here. During a small time interval dt, a firm collects $\pi_{k+1}dt$ profits. At the end of this interval, it is replaced by a new entrant with probability $\lambda z dt$ through creative destruction; otherwise, it preserves the monopoly power and V_{k+1}. Hence, the value function is written as:

$$V_{k+1} = \pi_{k+1}dt + (1 - rdt)\begin{bmatrix} \lambda z dt \times 0 + \\ (1 - \lambda z dt) \times V_{k+1} \end{bmatrix}.$$

Dividing both sides by dt, then taking the limit as $dt \to 0$ and using the fact that the equilibrium interest rate is equal to the time preference, the Bellman equation for V_{k+1} can be rewritten as:

$$\rho V_{k+1} = \pi_{k+1} - \lambda z V_{k+1}.$$

In other words, the annuity value of a new innovation (i.e. its flow value during a unit of time) is equal to the current profit flow π_{k+1} minus the expected capital loss $\lambda z V_{k+1}$ due to creative destruction, i.e. to the possible replacement by a subsequent innovator. If innovating gave the innovator access to a permanent profit flow π_{k+1}, then we know that

[8] Note that we are implicitly assuming that previous innovators are not candidates for being new innovators. This in fact results from a replacement effect pointed out by Arrow (1962). Namely, an outsider goes from zero to V_{k+1} if he or she innovates, whereas the previous innovator would go from V_k to V_{k+1}. Given that the R&D technology is linear, if outsiders are indifferent between innovating and working in manufacturing, then incumbent innovators will strictly prefer to work in manufacturing. Thus, new innovations end up being made by outsiders in equilibrium in this model. This feature will be relaxed in the next section.

the value of the corresponding perpetuity would be π_{k+1}/r.[9] However, there is creative destruction at rate λz. As a result, we have:

$$V_{k+1} = \frac{\pi_{k+1}}{\rho + \lambda z}, \tag{1.1}$$

that is, the value of innovation is equal to the profit flow divided by the risk-adjusted interest rate $\rho + \lambda z$ where the risk is that of being displaced by a new innovator.

1.2.2.2 Equilibrium Profits, Aggregate R&D, and Growth

We solve for equilibrium profits π_{k+1} and the equilibrium R&D rate z by backward induction. That is, first, for a given productivity of the current intermediate input, we solve for the equilibrium profit flow of the current innovator; then we move one step back and determine the equilibrium R&D using Equations (L) and (R).

Equilibrium profits Suppose that k_t innovations have already occurred until time t, so that the current productivity of the state-of-the-art intermediate input is $A_{k_t} = \gamma^{k_t}$. Given that the final good production is competitive, the intermediate good monopolist will sell his or her input at a price equal to its marginal product, namely:

$$p_k(y) = \frac{\partial (A_k y^\alpha)}{\partial y} = A_k \alpha y^{\alpha - 1}. \tag{1.2}$$

This is the inverse demand curve faced by the intermediate good monopolist.

Given that inverse demand curve, the monopolist will choose y to:

$$\pi_k = \max_y \{p_k(y)y - w_k y\}, \quad \text{subject to (1.2)} \tag{1.3}$$

since it costs $w_k y$ units of the numeraire to produce y units of the intermediate good. Given the Cobb-Douglas technology for the production of the final good, the equilibrium price is a constant markup over the marginal cost ($p_k = w_k/\alpha$) and the profit is simply equal to $\frac{1-\alpha}{\alpha}$ times the wage bill, namely:[10]

$$\pi_k = \frac{1-\alpha}{\alpha} w_k y, \tag{1.4}$$

where y solves (1.3).

Equilibrium aggregate R&D Combining (1.1), (1.4), and (R), we can rewrite the research-arbitrage equation as:

$$w_k = \lambda \frac{\frac{1-\alpha}{\alpha} w_{k+1} y}{\rho + \lambda z}. \tag{1.5}$$

[9] Indeed, the value of the perpetuity is:

$$\int_0^\infty \pi_{k+1} e^{-rt} dt = \frac{\pi_{k+1}}{r}.$$

[10] To see that $p_k = w_k/\alpha$, simply combine the first-order condition of (1.3) with expression (1.2).

Using the labor market clearing condition (L) and the fact that on a balanced growth path all aggregate variables (the final output flow, profits, and wages) are multiplied by γ each time a new innovation occurs, we can solve (1.5) for the equilibrium aggregate R&D z as a function of the parameters of the economy:

$$z = \frac{\frac{1-\alpha}{\alpha}\gamma L - \frac{\rho}{\lambda}}{1 + \frac{1-\alpha}{\alpha}\gamma}. \tag{1.6}$$

Clearly, it is sufficient to assume that $\frac{1-\alpha}{\alpha}\gamma L > \frac{\rho}{\lambda}$ to ensure positive R&D in equilibrium. Inspection of (1.6) delivers a number of important comparative statics. In particular, a higher productivity of the R&D technology as measured by λ or a larger size of innovations γ or a larger size of the population L has a positive effect on aggregate R&D. On the other hand, a higher α (which corresponds to the intermediate producer facing a more elastic inverse demand curve and therefore getting lower monopoly rents) or a higher discount rate ρ tends to discourage R&D.

Equilibrium expected growth Once we have determined the equilibrium aggregate R&D, it is easy to compute the expected growth rate. First note that during a small time interval $[t, t + dt]$, there will be a successful innovation with probability $\lambda z dt$. Second, the final output is multiplied by γ each time a new innovation occurs. Therefore, the expected log-output is simply:

$$\mathbb{E}(\ln Y_{t+dt}) = \lambda z dt \ln \gamma Y_t + (1 - \lambda z dt) \ln Y_t.$$

Subtracting $\ln Y_t$ from both sides, dividing through dt, and finally taking the limit leads to the following expected growth:

$$\mathbb{E}(g_t) = \lim_{dt \to 0} \frac{\ln Y_{t+dt} - \ln Y_t}{dt} = \lambda z \ln \gamma,$$

which inherits the comparative static properties of z with respect to the parameters $\lambda, \gamma, \alpha, \rho$, and L.

A distinct prediction of the model is:

Prediction 0: *The turnover rate λz is positively correlated with the growth rate g.*

1.3. GROWTH MEETS IO

Empirical studies (starting with Nickell (1996), Blundell et al. (1995, 1999)) point to a positive correlation between growth and product market competition. Also, the idea that competition—or free entry—should be growth-enhancing is also prevalent among policy advisers. Yet, non-Schumpeterian growth models cannot account for it: AK models assume perfect competition and therefore have nothing to say about the relationship between competition and growth. And in Romer's product variety model, higher competition amounts to a higher degree of substitutability between the horizontally differentiated inputs, which in turn implies lower rents for innovators and therefore lower R&D incentives and thus lower growth.

In contrast, the Schumpeterian growth paradigm can rationalize the positive correlation between competition and growth found in linear regressions. In addition, it can account for several interesting facts about competition and growth that no other growth theory can explain.[11] We shall concentrate on three such facts. First, innovation and productivity growth by incumbent firms appear to be stimulated by competition and entry, particularly in firms near the technology frontier or in firms that compete neck-and-neck with their rivals, less so than in firms below the frontier. Second, competition and productivity growth display an inverted-U relationship: starting from an initially low level of competition, higher competition stimulates innovation and growth; starting from a high initial level of competition, higher competition has a less positive or even a negative effect on innovation and productivity growth. Third, patent protection complements product market competition in encouraging R&D investments and innovation.

Understanding the relationship between competition and growth also helps improve our understanding of the relationship between trade and growth. Indeed, there are several dimensions to that relationship. First is the scale effect, whereby liberalizing trade increases the market for successful innovations and therefore the incentives to innovate; this is naturally captured by any innovation-based model of growth, including the Schumpeterian growth model. But there is also a competition effect of trade openness, which only the Schumpeterian model can capture. This latter effect appears to have been at work in emerging countries that implemented trade liberalization reforms (for example, India in the early 1990s),[12] and it also explains why trade restrictions are more detrimental to growth in more frontier countries (see Section 1.5 below).

1.3.1 From Leapfrogging to Step-By-Step Innovation[13]

1.3.1.1 The Argument

To reconcile theory with the evidence on productivity growth and product market competition, we replace the leapfrogging assumption of the model in the previous section (where incumbents are systematically overtaken by outside researchers) with a less radical *step-by-step* assumption. Namely, a firm that is currently m steps behind the technological leader in the same sector or industry must catch up with the leader before becoming a leader itself. This step-by-step assumption can be rationalized by supposing that an

[11] See Aghion and Griffith (2006) for a first attempt at synthesizing the theoretical and empirical debates on competition and growth.

[12] See, for instance, De Loecker et al. (2012), Goldberg et al. (2010), Sivadasan (2009), and Topalova and Khandelwal (2011).

[13] The following model and analysis are based on Aghion et al. (1997), Aghion et al. (2001), Aghion et al. (2005), and Acemoglu and Akcigit (2012). See also Peretto (1998) for related work.

innovator acquires tacit knowledge that cannot be duplicated by a rival without engaging in its own R&D to catch up. This leads to a richer analysis of the interplay between product market competition, innovation, and growth by allowing firms in some sectors to be *neck-and-neck*. In such sectors, increased product market competition, by making life more difficult for neck-and-neck firms, will encourage them to innovate in order to acquire a lead over their rival in the sector. This we refer to as the escape–competition effect. On the other hand, in sectors that are not neck-and-neck, increased product market competition will have a more ambiguous effect on innovation. In particular, it will discourage innovation by laggard firms when these do not put much weight on the (more remote) prospect of becoming a leader and instead mainly look at the short run extra profit from catching up with the leader. This we call the Schumpeterian effect. Finally, the steady-state fraction of neck-and-neck sectors will itself depend upon the innovation intensities in neck-and-neck versus unleveled sectors. This we refer to as the composition effect.

1.3.1.2 Household

Time is again continuous and a continuous measure L of individuals work in one of two activities: as production workers and as R&D workers. We assume that the representative household consumes C_t, has logarithmic instantaneous utility $U(C_t) = \ln C_t$, and discounts the future at a rate $\rho > 0$. Moreover, the household holds a balanced portfolio of all the firms, $\mathcal{A}_t$. Hence, its budget constraint is simply $C_t + \dot{\mathcal{A}}_t = r_t\mathcal{A}_t + Lw_t$. These assumptions deliver the household's Euler equation as $g_t = r_t - \rho$. All costs in this economy are in terms of labor units. Therefore, the household's consumption is equal to the final good production $C_t = Y_t$, which is also the resource constraint of this economy.

1.3.1.3 A Multi-Sector Production Function

To formalize these various effects, in particular the composition effect, we obviously need a multiplicity of intermediate sectors instead of one, as in the previous section. One simple way to extend the Schumpeterian paradigm to a multiplicity of intermediate sectors is, as in Grossman and Helpman (1991), to assume that the final good is produced using a continuum of intermediate inputs, according to the logarithmic production function:

$$\ln Y_t = \int_0^1 \ln y_{jt}\,dj. \tag{1.7}$$

Next, we introduce competition by assuming that each sector j is duopolistic with respect to production and research activities. We denote the two duopolists in sector j as A_j and B_j and assume, for simplicity, that y_j is the sum of the intermediate goods produced by the two duopolists in sector j:

$$y_j = y_{Aj} + y_{Bj}.$$

The above logarithmic technology implies that in equilibrium the same amount is spent at any time by final good producers on each basket y_j.[14] We normalize the price of the final good to be 1. Thus, a final good producer chooses each y_{Aj} and y_{Bj} to maximize $y_{Aj} + y_{Bj}$ subject to the budget constraint: $p_{Aj}y_{Aj} + p_{Bj}y_{Bj} = Y$. That is, he or she will devote the entire unit expenditure to the least expensive of the two goods.

1.3.1.4 Technology and Innovation

Each firm takes the wage rate as given and produces using labor as the only input according to the following linear production function:

$$y_{it} = A_{it}l_{it}, \quad i \in \{A, B\},$$

where l_{jt} is the labor employed. Let k_i denote the technology level of duopoly firm i in some industry j; that is, $A_i = \gamma^{k_i}$, $i = A, B$, and $\gamma > 1$ is a parameter that measures the size of a leading-edge innovation. Equivalently, it takes γ^{-k_i} units of labor for firm i to produce one unit of output. Thus, the unit cost of production is simply $c_i = w\gamma^{-k_i}$, which is independent of the quantity produced.

An industry j is thus fully characterized by a pair of integers (k_j, m_j) where k_j is the leader's technology and m_j is the technological gap between the leader and the follower.[15]

For expositional simplicity, we assume that knowledge spillovers between the two firms in any intermediate industry are such that neither firm can get more than one technological level ahead of the other, that is:

$$m \leq 1.$$

In other words, if a firm that is already one step ahead innovates, the lagging firm will automatically learn to copy the leader's previous technology and thereby remain only one step behind. Thus, at any point in time, there will be two kinds of intermediate sectors in the economy: (i) leveled or neck-and-neck sectors, where both firms are on a technological par with one another; and (ii) unleveled sectors, where one firm (the leader) lies one step ahead of its competitor (the laggard or follower) in the same industry.[16]

[14] To see this, note that a final good producer will choose the y_j's to maximize $u = \int \ln y_j dj$ subject to the budget constraint $\int p_j y_j dj = E$, where E denotes current expenditures. The first-order condition for this is:

$$\partial u/\partial y_j = 1/y_j = \lambda p_j \quad \text{for all} \quad j,$$

where λ is a Lagrange multiplier. Together with the budget constraint this first-order condition implies:

$$p_j y_j = 1/\lambda = E \quad \text{for all} \quad j.$$

[15] The above logarithmic final good technology, together with the linear production cost structure for intermediate goods, implies that the equilibrium profit flows of the leader and the follower in an industry depend only on the technological gap, m, between the two firms. We will see this below for the case where $m \leq 1$.

[16] Aghion et al. (2001) and Acemoglu and Akcigit (2012) analyze the more general case where there is no limit to how far ahead the leader can get.

To complete the description of the model, we just need to specify the innovation technology. Here we simply assume that by spending the R&D cost $\psi(z) = z^2/2$ in units of labor, a leader (or frontier) firm moves one technological step ahead at the rate z. We call z the innovation rate or R&D intensity of the firm. We assume that a follower firm can move one step ahead with probability h, even if it spends nothing on R&D, by copying the leader's technology. Thus, $z^2/2$ is the R&D cost (in units of labor) of a follower firm moving ahead with probability $z + h$. Let z_0 denote the R&D intensity of each firm in a neck-and-neck industry, and let z_{-1} denote the R&D intensity of a follower firm in an unleveled industry; if z_1 denotes the R&D intensity of the leader in an unleveled industry, note that $z_1 = 0$, since our assumption of automatic catch-up means that a leader cannot gain any further advantage by innovating.

1.3.2 Equilibrium Profits and Competition in Leveled and Unleveled Sectors

We can now determine the equilibrium profits of firms in each type of sector and link them with product market competition. The final good producer in (1.7) generates a unit-elastic demand with respect to each variety:

$$y_j = \frac{Y}{p_j}. \tag{1.8}$$

Consider first an unleveled sector where the leader's unit cost is c. The leader's monopoly profit is:

$$\begin{aligned} p_1 y_1 - c y_1 &= \left(1 - \frac{c}{p_1}\right) Y \\ &= \pi_1 Y, \end{aligned}$$

where the first line uses (1.8) and the second line defines π_1 as the equilibrium profit normalized by the final output Y. Note that the monopoly profit is monotonically increasing in the unit price p_1. However, the monopolist is constrained to setting a price $p_1 \leq \gamma c$, because γc is the rival's unit cost, so at any higher price the rival could profitably undercut his or her price and steal all their business. He or she will therefore choose the maximum possible price $p_1 = \gamma c$, such that the normalized profit in equilibrium is:

$$\pi_1 = 1 - \frac{1}{\gamma}.$$

The laggard in the unleveled sector will be priced out of the market and hence will earn a zero profit:

$$\pi_{-1} = 0.$$

Consider now a leveled (neck-and-neck) sector. If the two firms engaged in open price competition with no collusion, the equilibrium price would fall to the unit cost c

of each firm, resulting in zero profit. At the other extreme, if the two firms colluded so effectively as to maximize their joint profits and shared the proceeds, then they would together act like the leader in an unleveled sector, each setting $p = \gamma c$ (we assume that any third firm could compete using the previous best technology, just like the laggard in an unleveled sector), and each earning a normalized profit equal to $\pi_1/2$.

So in a leveled sector, both firms have an incentive to collude. Accordingly, we model the degree of product market competition inversely by the degree to which the two firms in a neck-and-neck industry are able to collude. (They do not collude when the industry is unleveled because the leader has no interest in sharing their profit.) Specifically, we assume that the normalized profit of a neck-and-neck firm is:

$$\pi_0 = (1 - \Delta)\,\pi_1, \quad 1/2 \leq \Delta \leq 1,$$

and we parameterize product market competition by Δ, that is, one minus the fraction of a leader's profits that the leveled firm can attain through collusion. Note that Δ is also the incremental profit of an innovator in a neck-and-neck industry, normalized by the leader's profit.

We next analyze how the equilibrium research intensities z_0 and z_{-1} of neck-and-neck and backward firms, respectively, and consequently the aggregate innovation rate, vary with our measure of competition Δ.

1.3.3 The Schumpeterian and Escape–Competition Effects

On a balanced growth path, all aggregate variables, including firm values, will grow at the rate g. For tractability, we will normalize all growing variables by the aggregate output Y. Let V_m (resp. V_{-m}) denote the normalized steady-state value of currently being a leader (resp. a follower) in an industry with technological gap m, and let $\omega = w/Y$ denote the normalized steady-state wage rate. We have the following Bellman equations:[17]

$$\rho V_0 = \max_{z_0} \left\{ \pi_0 + \overline{z}_0 (V_{-1} - V_0) + z_0 (V_1 - V_0) - \omega z_0^2/2 \right\}, \tag{1.9}$$

$$\rho V_{-1} = \max_{z_{-1}} \left\{ \pi_{-1} + (z_{-1} + h)(V_0 - V_{-1}) - \omega z_{-1}^2/2 \right\}, \tag{1.10}$$

$$\rho V_1 = \pi_1 + (z_{-1} + h)(V_0 - V_1), \tag{1.11}$$

where $\overline{z}_0$ denotes the R&D intensity of the other firm in a neck-and-neck industry (we focus on a symmetric equilibrium where $\overline{z}_0 = z_0$). Note that we already used $z_1 = 0$ in (1.11).

[17] Note that originally the left-hand side is written as $rV_0 - \dot{V}_0$. Note that on a BGP, $\dot{V}_0 = gV_0$; therefore, we get $(r - g)V_0$. Finally, using the household's Euler equation, $r - g = \rho$, leads to the Bellman equations in the text.

In words, the growth-adjusted annuity value ρV_0 of currently being neck-and-neck is equal to the corresponding profit flow π_0 plus the expected capital gain $z_0(V_1 - V_0)$ of acquiring a lead over the rival plus the expected capital loss $\overline{z}_0(V_{-1} - V_0)$, if the rival innovates and thereby becomes the leader, minus the R&D cost $\omega z_0^2/2$. Similarly, the annuity value ρV_1 of being a technological leader in an unleveled industry is equal to the current profit flow π_1 plus the expected capital loss $z_{-1}(V_0 - V_1)$ if the leader is being caught up by the laggard (recall that a leader does not invest in R&D in equilibrium). Finally, the annuity value ρV_{-1} of currently being a laggard in an unleveled industry is equal to the corresponding profit flow π_{-1} plus the expected capital gain $(z_{-1} + h)$ $(V_0 - V_{-1})$ of catching up with the leader, minus the R&D cost $\omega z_{-1}^2/2$.

Using the fact that z_0 maximizes (1.9) and z_{-1} maximizes (1.10), we have the first-order conditions:

$$\omega z_0 = V_1 - V_0, \tag{1.12}$$

$$\omega z_{-1} = V_0 - V_{-1}. \tag{1.13}$$

In Aghion et al. (1997) the model is closed by a labor market clearing equation that determines ω as a function of the aggregate demand for R&D plus the aggregate demand for manufacturing labor. Here, for simplicity we shall ignore that equation and take the wage rate ω as given, normalizing it at $\omega = 1$.

Then, using (1.12) and (1.13) to eliminate the V's from the system of Equations (1.9) –(1.11), we end up with a system of two equations in the two unknowns z_0 and z_{-1}:

$$z_0^2/2 + (\rho + h)z_0 - (\pi_1 - \pi_0) = 0, \tag{1.14}$$

$$z_{-1}^2/2 + (\rho + z_0 + h)z_{-1} - (\pi_0 - \pi_{-1}) - z_0^2/2 = 0. \tag{1.15}$$

These equations solve recursively for unique positive values of z_0 and z_{-1}, and we are mainly interested in how equilibrium R&D intensities are affected by an increase in product market competition Δ. It is straightforward to see from Equation (1.14) and the fact that:

$$\pi_1 - \pi_0 = \Delta \pi_1,$$

that an increase in Δ will increase the innovation intensity $z_0(\Delta)$ of a neck-and-neck firm. This is the escape–competition effect.

Then, plugging $z_0(\Delta)$ into (1.15), we can look at the effect of an increase in competition Δ on the innovation intensity z_{-1} of a laggard. This effect is ambiguous in general: in particular, for very high ρ, the effect is negative, since then z_{-1} varies like:

$$\pi_0 - \pi_{-1} = (1 - \Delta)\pi_1.$$

In this case, the laggard is very impatient and thus looks at its short-term net profit flow if it catches up with the leader, which in turn decreases when competition increases. This is the *Schumpeterian effect*. However, for low values of ρ, this effect is counteracted by an anticipated escape–competition effect.

Thus, the effect of competition on innovation depends on what situation a sector is in. In unleveled sectors, the Schumpeterian effect is at work even if it does not always dominate. But in leveled (neck-and-neck) sectors, the escape–competition effect is the only effect at work; that is, more competition induces neck-and-neck firms to innovate in order to escape from a situation in which competition constrains profits.

On average, an increase in product market competition will have an ambiguous effect on growth. It induces faster productivity growth in currently neck-an-neck sectors and slower growth in currently unleveled sectors. The overall effect on growth will thus depend on the (steady-state) fraction of leveled versus unleveled sectors. But this steady-state fraction is itself endogenous, since it depends on equilibrium R&D intensities in both types of sectors. We proceed to show under which condition this overall effect is an inverted U and, at the same time, derive additional predictions for further empirical testing.

1.3.3.1 Composition Effect and the Inverted U

In a steady state, the fraction of sectors μ_1 that are unleveled is constant, as is the fraction $\mu_0 = 1 - \mu_1$ of sectors that are leveled. The fraction of unleveled sectors that become leveled each period will be $z_{-1} + h$, so the sectors moving from unleveled to leveled represent the fraction $(z_{-1} + h)\,\mu_1$ of all sectors. Likewise, the fraction of all sectors moving in the opposite direction is $2z_0\mu_0$, since each of the two neck-and-neck firms innovates with probability z_0. In the steady state, the fraction of firms moving in one direction must equal the fraction moving in the other direction:

$$(z_{-1} + h)\mu_1 = 2z_0\,(1 - \mu_1)\,,$$

which can be solved for the steady-state fraction of unleveled sectors:

$$\mu_1 = \frac{2z_0}{z_{-1} + h + 2z_0}. \tag{1.16}$$

This implies that the aggregate flow of innovations in all sectors is[18]:

$$x = \frac{4\,(z_{-1} + h)\,z_0}{z_{-1} + h + 2z_0}.$$

One can show that for ρ large but h not too large, aggregate innovation x follows an inverted-U pattern: it increases with competition Δ for small enough values of Δ and

[18] x is the sum of the two flows: $(z_{-1} + h)\,\mu_1 + 2z_0\,(1 - \mu_1)$. But since the two flows are equal, $x = 2\,(z_{-1} + h)\,\mu_1$. Substituting for μ_1 using (1.16) yields $x = \frac{4(z_{-1}+h)z_0}{z_{-1}+h+2z_0}$.

decreases for large enough Δ. The inverted-U shape results from the composition effect whereby a change in competition changes the steady-state fraction of sectors that are in the leveled state, where the escape–competition effect dominates, versus the unleveled state, where the Schumpeterian effect dominates. At one extreme, when there is not much product market competition, there is not much incentive for neck-and-neck firms to innovate, and therefore, the overall innovation rate will be highest when the sector is unleveled. Thus, the industry will be quick to leave the unleveled state (which it does as soon as the laggard innovates) and slow to leave the leveled state (which will not happen until one of the neck-and-neck firms innovates). As a result, the industry will spend most of the time in the leveled state, where the escape–competition effect dominates (z_0 is increasing in Δ). In other words, if the degree of competition is very low to begin with, an increase in competition should result in a faster average innovation rate. At the other extreme, when competition is initially very high, there is little incentive for the laggard in an unleveled state to innovate. Thus, the industry will be slow to leave the unleveled state. Meanwhile, the large incremental profit $\pi_1 - \pi_0$ gives firms in the leveled state a relatively large incentive to innovate, so that the industry will be relatively quick to leave the leveled state. As a result, the industry will spend most of the time in the unleveled state where the Schumpeterian effect is the dominant effect. In other words, if the degree of competition is very high to begin with, an increase in competition should result in a slower average innovation rate.

Finally, using the fact that the log of an industry's output rises by the amount $\ln\gamma$ each time the industry completes two cycles from neck-and-neck ($m = 0$) to unleveled ($m = 1$) and then back to neck-and-neck, the average growth rate of final output g is simply equal to the frequency of completed cycles times $\ln\gamma$. But the frequency of completed cycles is itself equal to the fraction of time μ_1 spent in the unleveled state times the frequency $(z_{-1} + h)$ of innovation when in that state. Hence, overall, we have:

$$g = \mu_1 (z_{-1} + h) \ln\gamma = \frac{x}{2} \ln\gamma.$$

Thus, productivity growth follows the same pattern as aggregate innovation with regard to product market competition.

1.3.4 Predictions

The main testable predictions are:

Prediction 1: *The relationship between competition and innovation follows an inverted-U pattern and the average technological gap within a sector (μ_1 in the above model) increases with competition.*

This prediction is tested by Aghion et al. (2005) (hereafter ABBGH) using panel data on UK firms spanning 17 two-digit SIC industries between 1973 and 1994. The chosen measure of product market competition is equal to 1 minus the Lerner index. The Lerner index, or price–cost margin, is itself defined by operating profits net of

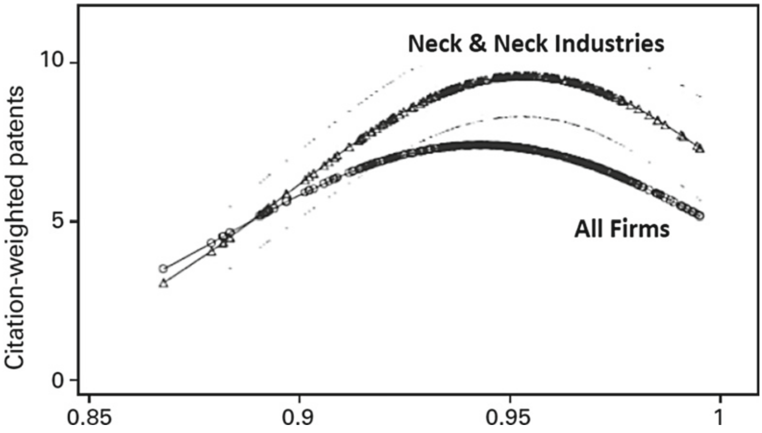


Figure 1.1 Competition and innovation (regression lines).

depreciation, provisions and financial cost of capital, divided by sales, averaged across firms within an industry-year. Figure 1.1 shows the inverted-U pattern, and it also shows that if we restrict attention to industries above the median degree of neck-and-neckness, the upward-sloping part of the inverted U is steeper than if we consider the whole sample of industries. ABBGH also show that the average technological gap across firms within an industry increases with the degree of competition the industry is subject to.

Prediction 2: *More intense competition enhances innovation in "frontier" firms but may discourage it in "non-frontier" firms.*

This prediction is tested by Aghion et al. (2009) (hereafter ABGHP). ABGHP use a panel of more than 5000 incumbent lines of businesses in UK firms in 180 four-digit SIC industries over the time period 1987–1993.

Taking the measure of technologically advanced entry of new foreign firms which ABGHP construct from administrative plant-level data as the proxy of competition, Figure 1.2 (taken from ABGHP, 2009) illustrates the following two results. First, the upper line, depicting how productivity growth responds to entry in incumbents that are more-than-median close to the frontier, is upward sloping, and this reflects the escape–competition effect at work in neck-and-neck sectors. Second, the lower line, depicting how productivity growth responds to entry in incumbents that are less-than-median close to the frontier, is downward sloping, which reflects the Schumpeterian effect of competition on innovation in laggards. In the main empirical analysis, ABGHP also control for the influence of trade and average profitability-related competition measures, and address the issue that entry, as well as the other explanatory variables, can be endogenous to incumbent productivity growth, as well as incumbent innovation. To tackle entry endogeneity, in particular, instruments are derived from a broad set of UK and EU-level policy reforms.

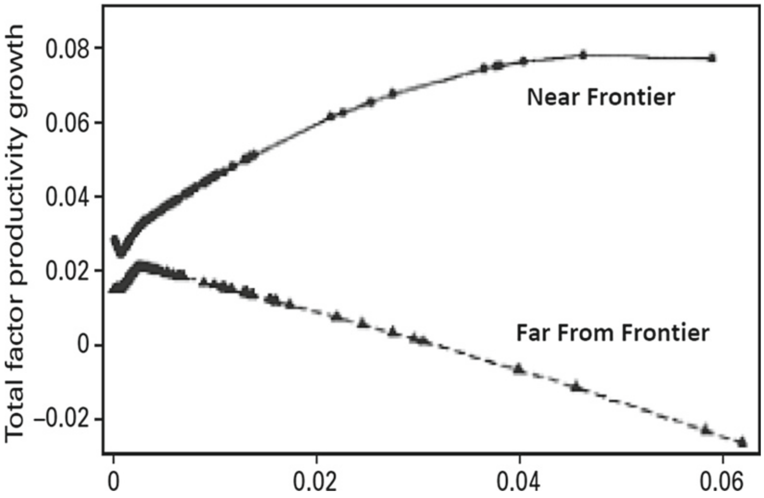


Figure 1.2 Entry and growth (regression lines).

Prediction 3: *There is complementarity between patent protection and product market competition in fostering innovation.*

In the above model, competition reduces the profit flow π_0 of non-innovating neck-and-neck firms, whereas patent protection is likely to enhance the profit flow π_1 of an innovating neck-and-neck firm. Both contribute to raising the net profit gain $(\pi_1 - \pi_0)$ of an innovating neck-and-neck firm; in other words, both types of policies tend to enhance the escape–competition effect. That competition and patent protection should be complementary in enhancing growth rather than mutually exclusive is at odds with Romer's (1990) product variety model, where competition is always detrimental to innovation and growth (as we discussed above) for exactly the same reason that intellectual property rights (IPRs) in the form of patent protection are good for innovation. Namely, competition reduces post-innovation rents, whereas patent protection increases these rents.[19] Empirical evidence in line with Prediction 3 has recently been provided. Qian (2007) uses the spreading of national pharmaceutical patent laws during the 1980s and 1990s to investigate the effects of patent protection on innovation. She reports that introducing national patent laws stimulates pharmaceutical innovation not on average across all countries, but, among others, in countries with high values of a country-level index of economic

[19] Similarly, in Boldrin and Levine (2008), patenting is detrimental to competition and thereby to innovation for the same reason that competition is good for innovation. To provide support to their analysis the two authors build a growth model in which innovation and growth can occur under perfect competition. The model is then used to argue that monopoly rents and therefore patents are not needed for innovation and growth. On the contrary, patents are detrimental to innovation because they reduce competition.

freedom. The index is the Fraser Institute index, which aggregates proxies of freedom to trade, in addition to measures of access to money, legal structure, and property rights.

Aghion et al. (2013) (hereafter AHP) set out to study whether patent protection can foster innovation when being complemented by product market competition, using country-industry panel data for many industries in OECD countries since the 1980s. AHP find that the implementation of a competition-increasing product market reform, the large-scale European Single Market Program, has increased innovation in industries of countries with strong IPRs since the pre-sample period, but not so in those with weaker IPRs. Moreover, the positive response of innovation to the product market reform in strong IPR countries is more pronounced among firms in industries that rely more on patenting than in other industries. Overall, these empirical results are consistent with a complementarity between IPRs and competition.

1.4. SCHUMPETERIAN GROWTH AND FIRM DYNAMICS

One of the main applications of the Schumpeterian theory has been the study of firm dynamics. The empirical literature has documented various stylized facts using micro firm-level data. Some of these facts are: (i) the firm size distribution is highly skewed; (ii) firm size and firm age are highly correlated; (iii) small firms exit more frequently, but the ones that survive tend to grow faster than the average growth rate; (iv) a large fraction of R&D in the US is done by incumbents; and (v) reallocation of inputs between entrants and incumbents is an important source of productivity growth.

These are some of the well-known empirical facts that non-Schumpeterian growth models cannot account for. In particular, the first four facts listed require a new firm to enter, expand, then shrink over time, and eventually be replaced by new entrants. These and the last fact on the importance of reallocation are all embodied in the Schumpeterian idea of creative destruction.

We will now consider a setup that closely follows the highly influential work by Klette and Kortum (2004). This model will add two elements to the baseline model of Section 1.2: First, innovations will come from both entrants and incumbents. Second, firms will be defined as a collection of production units where successful innovations by incumbents will allow them to expand in product space. Creative destruction will be the central force that drives innovation, invariant firm size distribution, and aggregate productivity growth on a balanced growth path.

1.4.1 The Setup

Time is again continuous and a continuous measure L of individuals work in one of three activities: (i) as production workers, l; (ii) as R&D scientists in incumbent firms, s_i; and (iii) as R&D scientists in potential entrants, s_e. The utility function is logarithmic; therefore, the household's Euler equation is $g_t = r_t - \rho$. The final good is produced

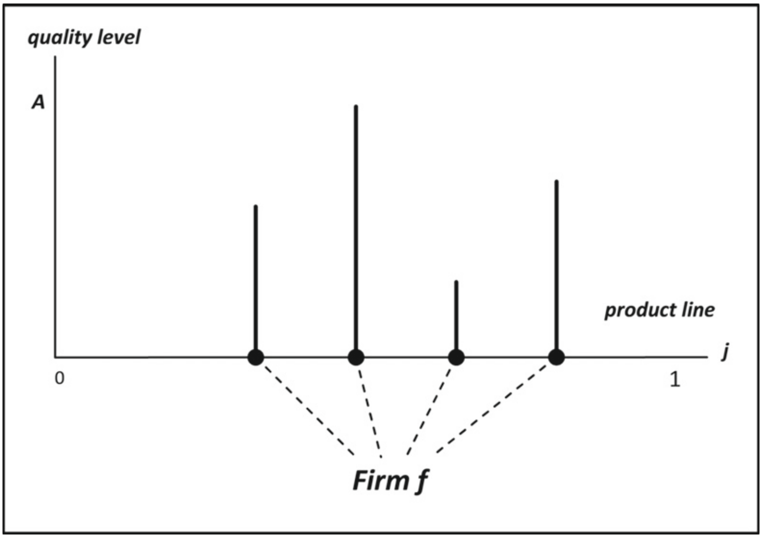


Figure 1.3 Example of a firm.

competitively using a combination of intermediate goods according to the following production function:

$$\ln Y_t = \int_0^1 \ln y_{jt} dj, \tag{1.17}$$

where y_j is the quantity produced of intermediate j. Intermediates are produced monopolistically by the innovator who innovated last within that product line j, according to the following linear technology:

$$y_{jt} = A_{jt} l_{jt},$$

where A_{jt} is the product-line-specific labor productivity and l_{jt} is the labor employed for production. This implies that the marginal cost of production in j is simply w_t/A_{jt} where w_t is the wage rate in the economy at time t.

A firm in this model is defined as a collection of n production units (product lines) as illustrated in Figure 1.3. Firms expand in product space through successful innovations. To innovate, firms combine their existing knowledge stock that they accumulated over time (n) with scientists (S_i) according to the following Cobb-Douglas production function:

$$Z_i = \left(\frac{S_i}{\zeta}\right)^{\frac{1}{\eta}} n^{1-\frac{1}{\eta}}, \tag{1.18}$$

where Z_i is the Poisson innovation flow rate, $\frac{1}{\eta}$ is the elasticity of innovation with respect to scientists, and ζ is a scale parameter. Note that this production function generates the following R&D cost of innovation:

$$C(z_i, n) = \zeta wnz_i^{\eta},$$

where $z_i \equiv Z_i/n$ is simply defined as the innovation intensity of the firm. When a firm is successful in its current R&D investment, it innovates over a random product line $j' \in [0, 1]$. Then, the productivity in line j' increases from $A_{j'}$ to $\gamma A_{j'}$. The firm becomes the new monopoly producer in line j' and thereby increases the number of its production lines to $n + 1$. At the same time, each of its n current production lines is subject to the creative destruction x by new entrants and other incumbents. Therefore, during a small time interval dt, the number of production units of a firm increases to $n+1$ with probability $Z_i dt$ and decreases to $n-1$ with probability $nxdt$. A firm that loses all of its product lines exits the economy.

1.4.2 Solving the Model

As before, our focus is on a balanced growth path, where all aggregate variables grow at the same rate g (to be determined). We will now proceed in two steps. First, we will solve for the static production decision and then turn to the dynamic innovation decision of firms, which will determine the equilibrium rate of productivity growth, as well as various firm moments along with the invariant firm size distribution.

1.4.2.1 Static Production Decision

As in Section 1.3, the final good producer spends the same amount Y_t on each variety j. As a result, the final good production function in (1.17) generates a unit-elastic demand with respect to each variety: $y_{jt} = Y_t/p_{jt}$. Combined with the fact that firms in a single-product line compete à la Bertrand, this implies that a monopolist with marginal cost w_t/A_{jt} will follow limit pricing by setting its price equal to the marginal cost of the previous innovator $p_{jt} = \gamma w_t/A_{jt}$. The resulting equilibrium quantity and profit in product line j are:

$$y_{jt} = \frac{A_{jt} Y_t}{\gamma w_t} \quad \text{and} \quad \pi_{jt} = \pi Y_t, \tag{1.19}$$

where $\pi \equiv \frac{\gamma-1}{\gamma}$. Note that profits are constant across product lines, which will significantly simplify the aggregation up to the firm level. Note also that the demand for production workers in each line is simply $Y_t/(\gamma w_t)$.

1.4.2.2 Dynamic Innovation Decision

Next we turn to the innovation decision of the firms. The stock-market value of an n-product firm $V_t(n)$ at date t satisfies the Bellman equation:

$$rV_t(n) - \dot{V}_t(n) = \max_{z_i \geq 0} \left\{ \begin{array}{c} n\pi_t - w_t \zeta n z_i^{\eta} \\ +nz_i \left[V_t(n+1) - V_t(n) \right] \\ +nx \left[V_t(n-1) - V_t(n) \right] \end{array} \right\}. \tag{1.20}$$

The intuition behind this expression is as follows. The firm collects a total of $n\pi_t$ profits from n product lines and invests in total $w_t\zeta nz_i^\eta$ in R&D. As a result, it innovates at the flow rate $Z_i \equiv nz_i$, in which case it gains $V_t(n+1) - V_t(n)$. In addition, the firm loses each of its product lines through creative destruction at the rate x, which means that a production line will be lost overall at a rate nx, leading to a loss of $V_t(n) - V_t(n-1)$. It is a straightforward exercise to show that the value function in (1.20) is linear in the number of product lines n and proportional to aggregate output Y_t, with the form:

$$V_t(n) = nvY_t.$$

In this expression, $v = V_t(n)/nY_t$ is simply the average normalized value of a production unit that is endogenously determined as:

$$v = \frac{\pi - \zeta\omega z_i^\eta}{\rho + x - z_i}. \tag{1.21}$$

Note that this expression uses the Euler equation $\rho = r - g$ and that labor share is defined as $\omega \equiv w_t/Y_t$, which is constant on a balanced growth path. In the absence of incumbent innovation, i.e. $z_i = 0$, this value is equivalent to the baseline model (1.1). The fact that incumbents can innovate modifies the baseline value in two opposite directions: First, the cost of R&D investment is subtracted from the gross profit, which lowers the net instantaneous return $\pi - \zeta\omega z_i^\eta$. However, each product line comes with an R&D option value, that is, having one more production unit increases the firm's R&D capacity as in (1.18) and therefore the firm's value.

The equilibrium innovation decision of an incumbent is simply found through the first-order condition of (1.20):

$$z_i = \left(\frac{v}{\eta\zeta\omega}\right)^{\frac{1}{\eta-1}}. \tag{1.22}$$

As expected, innovation intensity is increasing in the value of innovation v and decreasing in the labor cost ω.

1.4.2.3 Free Entry

We consider a mass of entrants that produce one unit of innovation by hiring ψ number of scientists. When a new entrant is successful, it innovates over a random product line by improving its productivity by $\gamma > 1$. It then starts out as a single-product firm. Let us denote the entry rate by z_e. The free-entry condition equates the value of a new entry $V_t(1)$ to the cost of innovation ψw_t, such that:

$$v = \omega\psi. \tag{1.23}$$

Recall that the rate of creative destruction is simply the entry rate plus an incumbent's innovation intensity, i.e. $x = z_i + z_e$. Using this fact, together with (1.21)–(1.23), delivers the equilibrium entry rate and incumbent innovation intensity:

$$z_e = \frac{\pi}{\omega\psi} - \frac{1}{\eta}\left(\frac{\psi}{\eta\zeta}\right)^{\frac{1}{\eta-1}} - \rho \quad \text{and} \quad z_i = \left(\frac{\psi}{\eta\zeta}\right)^{\frac{1}{\eta-1}}.$$

1.4.2.4 Labor Market Clearing

Now we are ready to close the model by imposing the labor market clearing condition. The equilibrium labor share ω equates the supply of labor L to the sum of aggregate labor demand coming from (i) production, $(\gamma\omega)^{-1}$, (ii) incumbent R&D, $\zeta\left(\frac{\psi}{\eta\zeta}\right)^{\frac{\eta}{\eta-1}}$, and (iii) outside entrants, $\frac{\pi}{\omega} - \zeta\left(\frac{\psi}{\eta\zeta}\right)^{\frac{\eta}{\eta-1}} - \psi\rho$. The resulting labor share is:

$$\omega = \frac{w_t}{Y_t} = \frac{1}{L + \rho\psi}.$$

1.4.3 Equilibrium Growth Rate

In this model, innovation takes place by both incumbents and entrants at the total rate of $x = z_i + z_e$. Hence, the equilibrium growth rate is:

$$\begin{aligned} g &= x \ln \gamma \\ &= \left[\left(\frac{\gamma-1}{\gamma}\right)\frac{L}{\psi} + \left(\frac{\eta-1}{\eta}\right)\left(\frac{\psi}{\eta\zeta}\right)^{\frac{1}{\eta-1}} - \frac{\rho}{\gamma}\right]\ln\gamma. \end{aligned}$$

In addition to the standard effects, such as the growth rate increasing in the size of innovation and decreasing in the discount rate, this model generates an interesting non-linear relationship between entry cost ψ and growth. An increase in the entry cost reduces the entry rate and therefore has a negative effect on equilibrium growth. However, this effect also frees up those scientists that used to be employed by outside entrants and reallocates them to incumbents, hence increasing innovation by incumbents and growth. This is an interesting trade-off for industrial policy. In a recent work, Acemoglu et al. (2013) analyze the effects of various industrial policies on equilibrium productivity growth, including entry subsidy and incumbent R&D subsidy, in an enriched version of the above framework.

1.4.4 Predictions

Now we go back to the initial list of predictions and discuss how they are captured by the above model.

Prediction 1: *The size distribution of firms is highly skewed.*

In this model, firm size is summarized by the number of product lines of a firm. Let us denote by μ_n the fraction of firms that have n products. The invariant distribution μ_n is found by equating the inflows into state n to the outflows from it:

$$\begin{aligned} \mu_1 x &= z_e, \\ (z_i + x)\,\mu_1 &= \mu_2 2x + z_e, \\ (z_i + x)\,n\mu_n &= \mu_{n+1}\,(n+1)\,x + \mu_{n-1}\,(n-1)\,z_i \quad \text{for } n \geq 2. \end{aligned}$$

The first line equates exits to entry. The left-hand side of the second line consists of outflows from being a one-product firm that happen when a one-product firm innovates itself and becomes a two-product firm or is replaced by another firm at the rate x. The right-hand side is the sum of the inflows coming from two-product firms or from outsiders. The third line generalizes the second line to n-product firms. The resulting firm size distribution is geometric as illustrated in Figure 1.4 and has the following exact form:

$$\mu_n\,(z_e/z_i) = \frac{z_e/z_i}{(1 + z_e/z_i)^n\, n},$$

and highly skewed as shown in a vast empirical literature (Simon and Bonini, 1958; Ijiri and Simon, 1977; Schmalensee, 1989; Stanley et al. 1995; Axtell, 2001; Rossi-Hansberg and Wright, 2007). Several alternative Schumpeterian models have been proposed after (Klette and Kortum, 2004) that feature invariant firm size distributions with a Pareto tail. (See Acemoglu and Cao (2011) for an example and a discussion of the literature.)

Prediction 2: *Firm size and firm age are positively correlated.*

In the current model, firms are born with a size of 1. Subsequent successes are required for firms to grow in size, which naturally produces a positive correlation between size and age. This regularity has been documented extensively in the literature. (For recent discussions and additional references, see Haltiwanger et al. (2010) and Akcigit and Kerr (2010)).

Prediction 3: *Small firms exit more frequently. The ones that survive tend to grow faster than average.*

In the above model, firm exit happens through the loss of product lines. Conditional on not producing a new innovation, a firm's probability of losing all of its product lines and exiting within a period is $(x\Delta t)^n$, which decreases in n. Clearly it becomes much more difficult for a firm to exit when it expands in product space.

The facts that small firms exit more frequently and grow faster conditional on survival have been widely documented in the literature (for early work, see Birch (1981, 1987) and Davis et al. (1996). For more recent work, see Haltiwanger et al. (2010), Akcigit and Kerr (2010), and Neumark et al. (2008)).

Prediction 4: *A large fraction of R&D is done by incumbents.*

There is an extensive literature that studies R&D investment and the patenting behavior of existing firms in the US (see, for instance, among many others, Acs and Audretsch

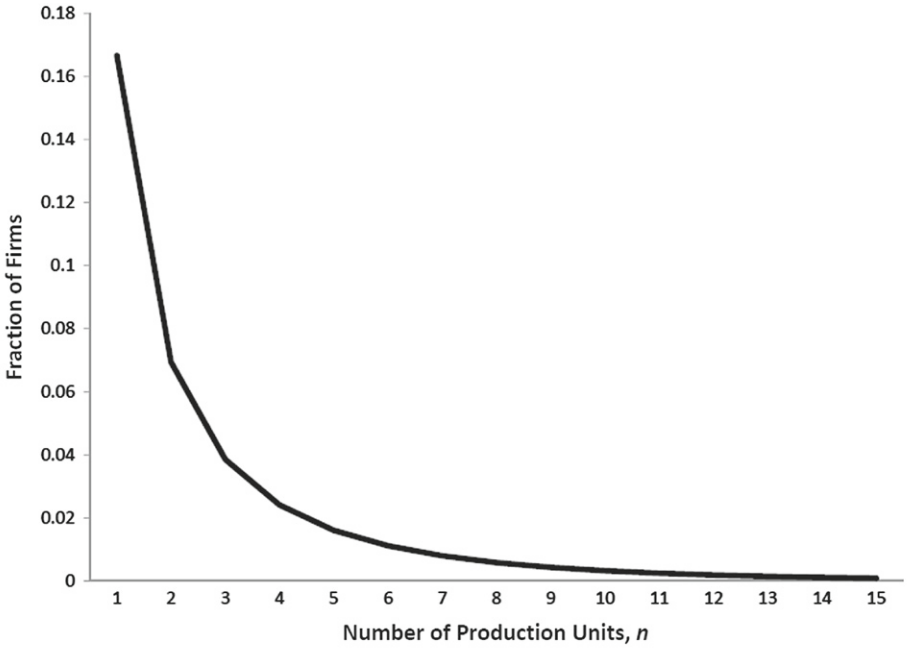


Figure 1.4 Firm size distribution.

(1988, 1991), Griliches (1990), Hall et al. (2001), Cohen (1995), and Cohen and Klepper (1996)). In particular, Freeman (1982), Pennings and Buitendam (1987), Tushman and Anderson (1986), Scherer (1984), and Akcigit and Kerr (2010) show that large incumbents focus on improving the existing technologies, whereas small new entrants focus on innovating with radical new products or technologies. Similarly, Akcigit et al. (2012) provide empirical evidence on French firms showing that large incumbents with a broad technological spectrum account for most of the private basic research investment.

On the theory side, Akcigit and Kerr (2010), Acemoglu and Cao (2011), and Acemoglu et al. (2012) have also provided alternative Schumpeterian models that capture this fact.

Prediction 5: *Both entrants and incumbents innovate. Moreover, the reallocation of resources among incumbents, as well as from incumbents to new entrants, is the major source of productivity growth.*

A central feature of this model is that both incumbents and entrants innovate and contribute to productivity growth. New entrants account for:

$$\frac{z_e}{z_e + z_i} = 1 - \left[\left(\left(\frac{\gamma - 1}{\gamma} \right) \frac{L}{\psi} - \frac{\rho}{\gamma} \right) \left(\frac{\eta \zeta}{\psi} \right)^{\frac{1}{\eta - 1}} + \frac{\eta - 1}{\eta} \right]^{-1},$$

percent of innovations in any given period. Bartelsman and Doms (2000) and Foster et al. (2001) have shown that 25% of productivity growth in the US is accounted for by new entry and the remaining 75% by continuing plants. Moreover, Foster et al. (2001, 2006) have shown that reallocation of resources through entry and exit accounts for around 50% of manufacturing and 90% of US retail productivity growth. In a recently growing cross-country literature, Hsieh and Klenow (2009, 2012), Bartelsman et al. (2009), and Syverson (2011) describe how variations in reallocation across countries explain differences in productivity levels. Lentz and Mortensen (2008) and Acemoglu et al. (2013) estimate variants of the baseline model in Klette and Kortum (2004) to quantify the importance of reallocation and study the impacts of industrial policy on reallocation and productivity growth.

1.5. GROWTH MEETS DEVELOPMENT

In this section, we argue that Schumpeterian growth theory helps bridge the gap between growth and development economics, by offering a simple framework to capture the idea that growth-enhancing policies or institutions may vary with a country's level of technological development. In particular, we will look at the role of democracy in the growth process, arguing that democracy matters for growth to a larger extent in more advanced economies.

1.5.1 Innovation Versus Imitation and the Notion of Appropriate Institutions

Innovations in one sector or one country often build on knowledge that was created by innovations in another sector or country. The process of diffusion, or technology spillover, is an important factor behind cross-country convergence. Howitt (2000) showed how this can lead to cross-country conditional convergence of growth rates in Schumpeterian growth models. Specifically, a country that starts far behind the world technology frontier can grow faster than one close to the frontier because the former country will make a larger technological advance every time one of its sectors catches up to the global frontier. In Gerschenkron's (1962) terms, countries far from the frontier enjoy an "advantage of backwardness." This advantage implies that, in the long run, a country with a low rate of innovation will fall behind the frontier but will grow at the same rate as the frontier; as they fall further behind, the advantage of backwardness eventually stabilizes the gap that separates them from the frontier.

These same considerations imply that policies and institutions that are appropriate for countries close to the global technology frontier are often different from those that are appropriate for non-frontier countries, because those policies and institutions that help a country to copy, adapt, and implement leading-edge technologies are not necessarily the same as those that help it to make leading-edge innovations. The idea of appropriate institutions was developed more systematically by Acemoglu et al. (2006), henceforth

AAZ, and it underlies more recent work, in particular, Acemoglu and Robinson's best-selling book *Why Nations Fail* (Acemoglu and Robinson (2012)), in which the authors rely on a rich set of country studies to argue that sustained growth requires creative destruction and therefore is not sustainable in countries with extractive institutions.

A particularly direct and simpler way to formalize the idea of appropriate growth policy is to move for a moment from continuous to discrete time. Following AAZ and more remotely (Nelson and Phelps, 1966), let A_t denote the current average productivity in the domestic country, and $\overline{A}_t$ denote the current (world) frontier productivity. Then, think of innovation as multiplying productivity by factor γ, and of imitation as catching up with the frontier technology.

Then, if the fraction μ_n of sectors innovates and the fraction μ_m imitates, we have:

$$A_{t+1} - A_t = \mu_n (\gamma - 1) A_t + \mu_m \left(\overline{A}_t - A_t\right).$$

This in turn implies that productivity growth hinges upon the country's degree of "frontierness," i.e. its "proximity" $a_t = A_t / \overline{A}_t$ to the world frontier, namely:

$$g_t = \frac{A_{t+1} - A_t}{A_t} = \mu_n (\gamma - 1) + \mu_m \left(a_t^{-1} - 1\right).$$

In particular:

Prediction 1: *The closer to the frontier an economy is, that is, the closer to one the proximity variable a_t is, the more is growth driven by "innovation-enhancing" rather than "imitation-enhancing" policies or institutions.*

1.5.2 Further Evidence on Appropriate Growth Policies and Institutions

In Section 1.3 we mentioned some recent evidence for the prediction that competition and free-entry should be more growth-enhancing. Using a cross-country panel of more than 100 countries over the 1960–2000 period, AAZ regress the average growth rate on a country's distance to the US frontier (measured by the ratio of GDP per capita in that country to per capita GDP in the US) at the beginning of the period. Then, they split the sample of countries into two groups, corresponding respectively to countries that are more open than the median and to countries that are less open than the median. The prediction is:

Prediction 2: *Average growth should decrease more rapidly as a country approaches the world frontier when openness is low.*

To measure openness one can use imports plus exports divided by aggregate GDP. But this measure suffers from obvious endogeneity problems; in particular, exports and imports are likely to be influenced by domestic growth. To deal with the endogeneity problem, Frankel and Romer (1999) construct a more exogenous measure of openness that relies on exogenous characteristics such as land area, common borders, geographical distance, population, etc. and it is this measure that AAZ use to measure openness in the following figures.

Figure 1.5A and B shows the cross-sectional regression. Here, average growth over the whole 1960–2000 period is regressed over the country's distance to the world technology frontier in 1965, respectively for less open and more open countries. A country's distance

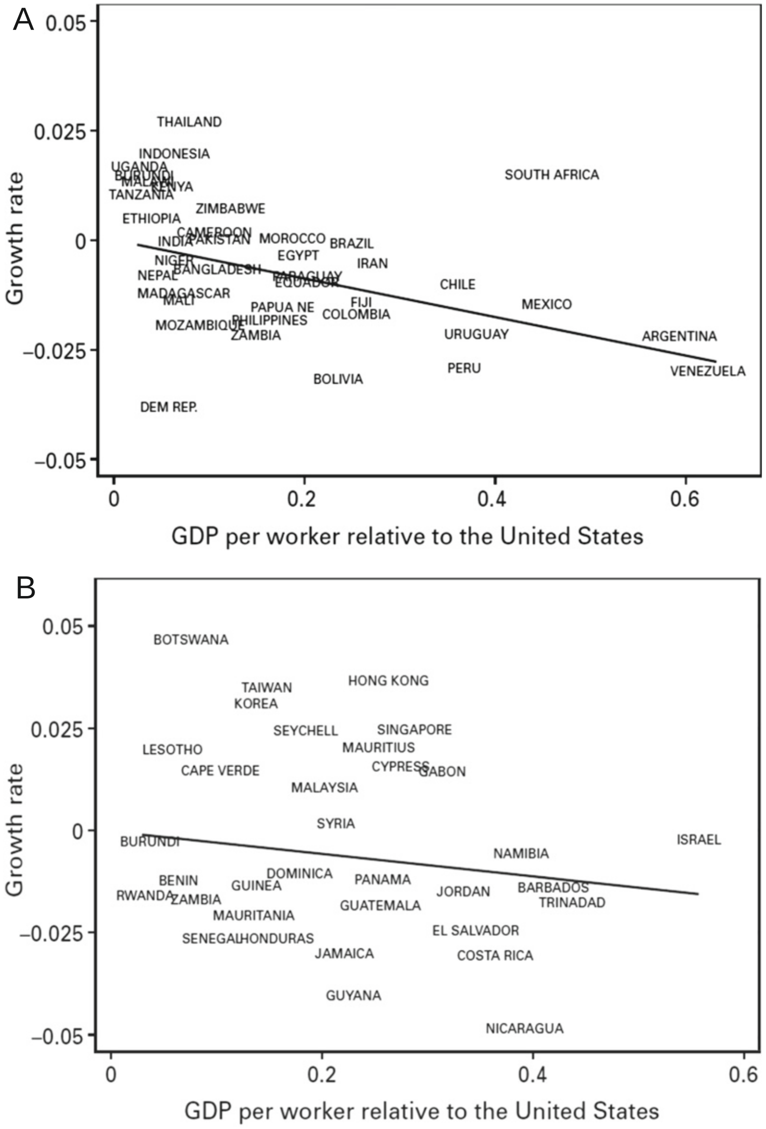


Figure 1.5 Growth, openness and distance to frontier. A: less open countries (cross-section) B: more open countries (cross-section) C: less open countries (Panel) D: more open countries (panel).

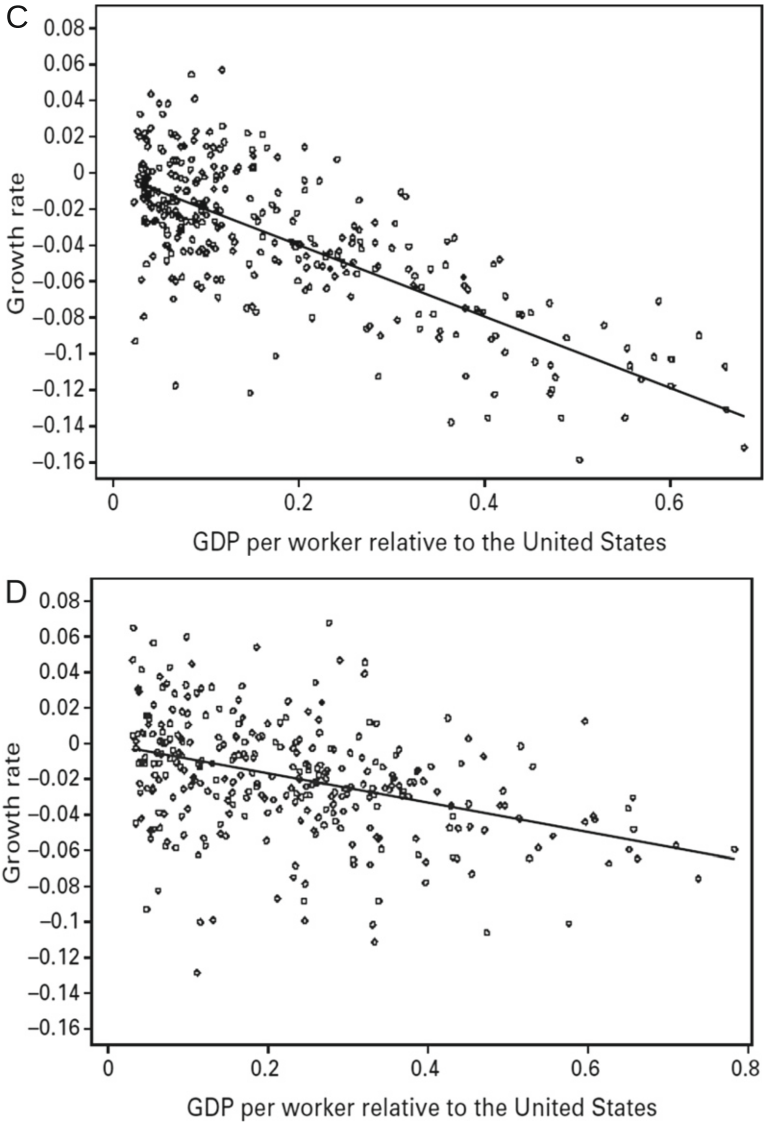


Figure 1.5 (*Continued*).

to the frontier is measured by the ratio between the log of this country's level of per capita GDP and the maximum of the logs of per capita GDP across all countries (which corresponds to the log of per capita GDP in the US).[20]

[20] That the regression lines should all be downward sloping reflects the fact that countries farther below the world technology frontier achieve bigger technological leaps whenever they successfully catch up with

Figure 1.5C and D shows the results of panel regressions where AAZ decompose the period 1960–2000 in 5-year subperiods and then for each subperiod AAZ regress average growth over the period on distance to the frontier at the beginning of the subperiod, respectively for less open and more open countries. These latter regressions control for country fixed effects. In both cross-sectional and panel regressions we see that while a low degree of openness does not appear to be detrimental to growth in countries far below the world frontier, it becomes increasingly detrimental to growth as the country approaches the frontier.

AAZ repeat the same exercise using entry costs faced by new firms instead of openness. The prediction is:

Prediction 3: *High entry barriers become increasingly detrimental to growth as the country approaches the frontier.*

Entry costs in turn are measured by the number of days to create a new firm in the various countries (see Djankov et al. 2002). Here, the country sample is split between countries with high barriers relative to the median and countries with low barriers relative to the median. Figure 1.6A and B shows the cross-sectional regressions, respectively, for high and low barrier countries, whereas Figure 1.6C and D shows the panel regressions for the same two subgroups of countries. Both types of regressions show that while high entry barriers do not appear to be detrimental to growth in countries far below the world frontier, they indeed become increasingly detrimental to growth as the country approaches the frontier.

These two empirical exercises point to the importance of interacting institutions or policies with technological variables in growth regressions: openness is particularly growth-enhancing in countries that are closer to the technological frontier; entry is more growth-enhancing in countries or sectors that are closer to the technological frontier; below we will see that higher (in particular, graduate) education tends to be more growth-enhancing in countries or in US states that are closer to the technological frontier, whereas primary-secondary (possibly undergraduate) education tends to be more growth enhancing in countries or in US states that are farther below the frontier.

A third piece of evidence is provided by Aghion et al. (2009), who use cross-US-states panel data to look at how spending on various levels of education matter differently for growth across US states with different levels of frontierness as measured by their average productivity compared to frontier-state (Californian) productivity. The gray bars in Figure 1.7 do not factor in the mobility of workers across US states, whereas the solid black bars do. The more frontier a country or region is, the more its growth relies on frontier innovation and therefore our prediction is:

the frontier (this is the "advantage of backwardness" we mentioned above). More formally, for given μ_n and $\mu_m, g_t = \mu_n(\gamma - 1) + \mu_m\left(a_t^{-1} - 1\right)$ is decreasing in a_t.

Prediction 4: *The more frontier an economy is, the more growth in this economy relies on research education.*

As shown in the figure below, research-type education is always more growth-enhancing in states that are more frontier, whereas a bigger emphasis on 2-year colleges is more growth-enhancing in US states that are farther below the productivity frontier. This is not surprising:Vandenbussche et al. (2006) obtain similar conclusions using

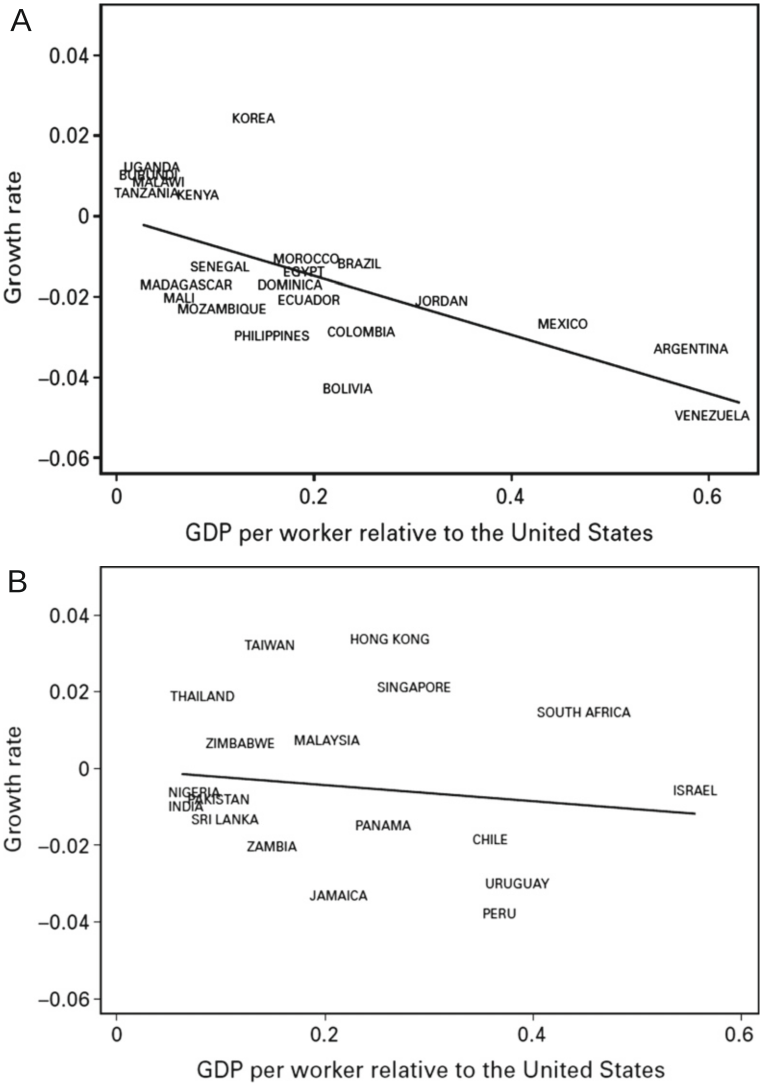


Figure 1.6 Growth, entry and distance to frontier. A: high barrier countries (cross-section) B: low barrier countries (cross-section) C: high barrier countries (panel) D: low barrier countries (panel).

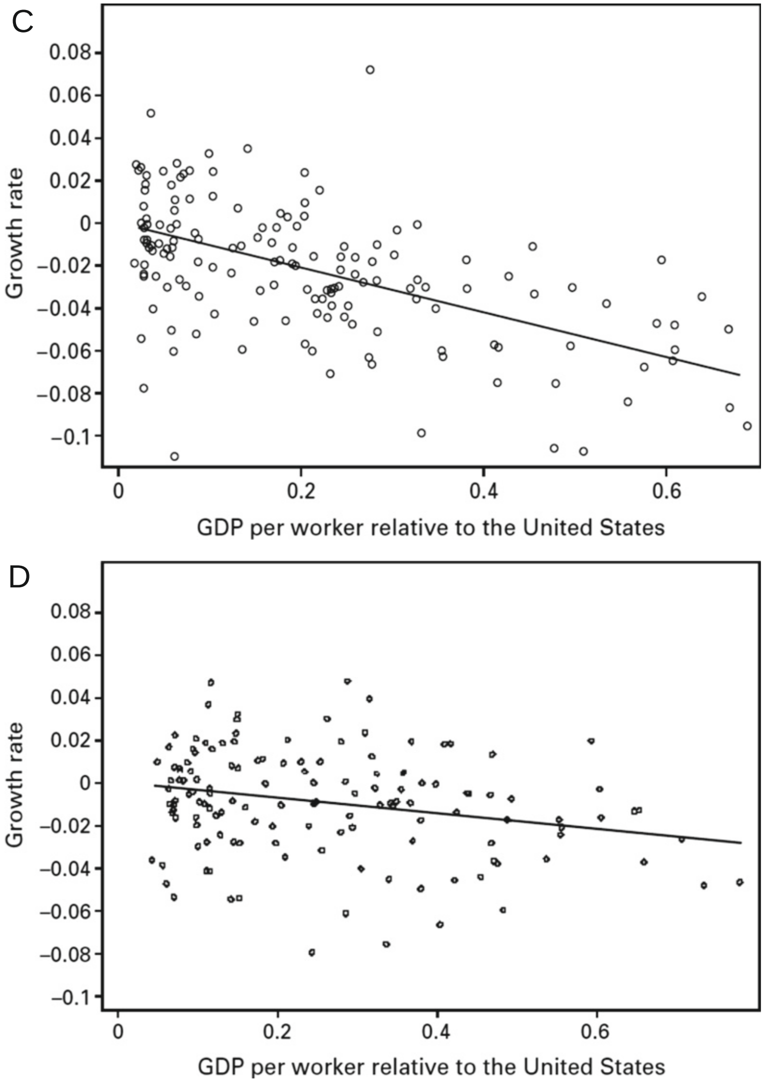


Figure 1.6 (*Continued*).

cross-country panel data, namely, that tertiary education is more positively correlated with productivity growth in countries that are closer to the world technology frontier.

1.5.3 Political Economy of Creative Destruction

Does democracy enhance or hamper economic growth? One may think of various channels whereby democracy should affect per capita GDP growth. A first channel is that democracy pushes for more redistribution from rich to poor, and that redistribution in

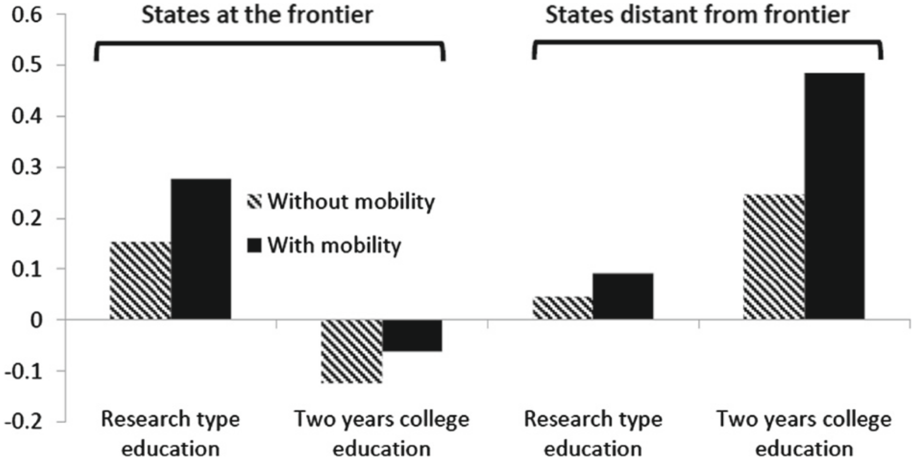


Figure 1.7 Growth, education, and distance to frontier.

turn affects growth. Thus, Persson and Tabellini (1994) and Alesina and Rodrik (1994) analyze the relationship between inequality, democratic voting, and growth. They develop models in which redistribution from rich to poor is detrimental to growth as it discourages capital accumulation. More inequality is then also detrimental to growth because it results in the median voter becoming poorer and therefore demanding more redistribution. A second channel, which we explore in this section, is Schumpeterian: namely, democracy reduces the scope for expropriating successful innovators or for incumbents to prevent new entry by using political pressure or bribes. In other words, democracy facilitates creative destruction and thereby encourages innovation.[21] To the extent that innovation matters more for growth in more frontier economies, the prediction is:

Prediction 5: *The correlation between democracy and innovation/growth is more positive and significant in more frontier economies.*

The relationship between democracy, "frontierness" and growth, thus provides yet another illustration of our notion of appropriate institutions. In the next subsection we develop a simple Schumpeterian model that generates this prediction.

1.5.3.1 The Formal Argument

Consider the following Schumpeterian model in discrete time. All agents and firms live for one period. In each period t a final good (henceforth the numeraire) is produced in each state by a competitive sector using a continuum one of intermediate inputs,

[21] Acemoglu and Robinson (2006) formalize another reason, also Schumpeterian, as to why democracy matters for innovation: namely, new innovations not only destroy the economic rents of incumbent producers, they also threaten the power of incumbent political leaders.

according to the technology:

$$\ln Y_t = \int_0^1 \ln y_{jt}\, dj, \tag{1.24}$$

where the intermediate products are produced again by labor according to:

$$y_{jt} = A_{jt} l_{jt}. \tag{1.25}$$

There is a competitive fringe of firms in each sector that are capable of producing a product with technology level A_{jt}/γ. So, as before, each incumbent's profit flow is:

$$\pi_{jt} = \pi Y_t,$$

where $\pi \equiv \frac{\gamma - 1}{\gamma}$. Note that as in (1.19), each incumbent will produce using the same amount of labor:

$$l_{jt} = \frac{Y_t}{\gamma w_t} \equiv l, \tag{1.26}$$

where l is the economy's total use of manufacturing labor. We assume that there is measure one unit of labor that is used only for production. Therefore $l = 1$ implies:

$$w_t = \frac{Y_t}{\gamma}.$$

Finally, (1.24)–(1.26) deliver the final output as a function of the aggregate productivity A_t in this economy:

$$Y_t = A_t,$$

where $\ln A_t \equiv \int_0^1 \ln A_{jt}\, dj$ is the end-of-period-t aggregate productivity index.

Technology and entry Let $\overline{A}_t$ denote the new world productivity frontier at date t and assume that:

$$\overline{A}_t = \gamma \overline{A}_{t-1},$$

with $\gamma > 1$ exogenously given. We shall again emphasize the distinction already made in the previous section between sectors in which the incumbent producer is neck-and-neck with the frontier and those in which the incumbent firm is below the frontier; at the beginning of date t, a sector j can either be at the current frontier, with productivity level $A^b_{jt} = \bar{A}_{t-1}$ (advanced sector) or one step below the frontier, with productivity level $A^b_{jt} = \bar{A}_{t-2}$ (backward sector). Thus, imitation—or knowledge spillovers—in this model means that whenever the frontier moves up one step from $\bar{A}_{t-1}$ to $\overline{A}_t$, then backward sectors also automatically move up one step from $\overline{A}_{t-2}$ to $\overline{A}_{t-1}$.

In each intermediate sector j, only one incumbent firm I_j, and one potential entrant E_j, are active in each period. In this model, innovation in a sector is made only by

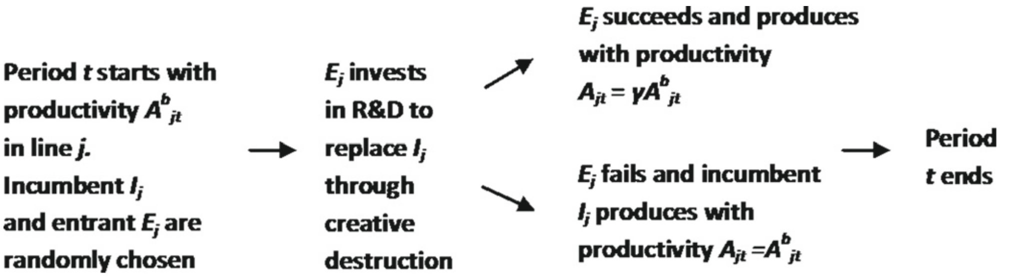

Period t starts with productivity A^b_{jt} in line j. Incumbent I_j and entrant E_j are randomly chosen
E_j invests in R&D to replace I_j through creative destruction
E_j succeeds and produces with productivity $A_{jt} = \gamma A^b_{jt}$
E_j fails and incumbent I_j produces with productivity $A_{jt} = A^b_{jt}$
Period t ends

Figure 1.8 Timing of events.

a potential entrant E_j since innovation does not change the incumbent's profit rate. Before production takes place, potential entrant E_j invests in R&D in order to replace the incumbent I_j. If successful, it increases the current productivity of sector j to $A_{jt} = \gamma A^b_{jt}$ and becomes the new monopolist and produces. Otherwise, the current incumbent preserves its monopoly right and produces with the beginning-of-period productivity $A_{jt} = A^b_{jt}$ and the period ends. The timing of events is described in Figure 1.8.

Finally, the innovation technology is as follows: if a potential entrant E_j spends $A_t \lambda z^2_{jt}/2$ on R&D in terms of the final good, then she innovates with probability z_{jt}.

Democracy Entry into a sector is subject to the democratic environment in the domestic country. Similar to Acemoglu and Robinson (2006), we model democracy as freedom to enter. More specifically, in a country with democracy level $\beta \in [0, 1]$, a successful innovation leads to successful entry only with probability β, and it is blocked with probability $(1 - \beta)$. As a result, the probability of an unblocked entry is βz_j. An unblocked entrant raises productivity from A^b_{jt} to γA^b_{jt} and becomes the new monopoly producer.

Equilibrium innovation investments We can now analyze the innovation decision of the potential entrant E_j:

$$\max_{z_{jt}} \left\{ z_{jt} \beta \pi Y_t - A_t \lambda \frac{z^2_{jt}}{2} \right\}.$$

In equilibrium we get:

$$z_{jt} = \bar{z} = \frac{\beta \pi}{\lambda},$$

where we used the fact that $Y_t = A_t$. Thus, the aggregate equilibrium innovation effort is increasing in profit π and decreasing in R&D cost λ. Most important for us in this section, the innovation rate is increasing in the democracy level β:

$$\frac{\partial \bar{z}}{\partial \beta} > 0.$$

Growth Now we can turn to the equilibrium growth rate of average productivity. We will denote the fraction of advanced sectors by μ, which will also be the index for the frontierness of the domestic country. The average productivity of a country at the beginning of date t is:

$$A_{t-1} \equiv \int_0^1 A_{jt}\,dj = \mu \bar{A}_{t-1} + (1-\mu)\,\bar{A}_{t-2}.$$

Average productivity at the end of the same period is:[22]

$$A_t = \mu\left[\beta\bar{z}\gamma\,\bar{A}_{t-1} + (1-\beta\bar{z})\,\bar{A}_{t-1}\right] + (1-\mu)\,\bar{A}_{t-1}.$$

Then the growth rate of average productivity is simply equal to:

$$g_t = \frac{A_t - A_{t-1}}{A_{t-1}} = \gamma\frac{\mu\beta\bar{z}\,(\gamma-1)+1}{\mu\,(\gamma-1)+1} - 1 > 0.$$

As is clear from the above expression, democracy is always growth-enhancing:

$$\frac{\partial g_t}{\partial \beta} = \left(\bar{z} + \frac{\partial \bar{z}}{\partial \beta}\beta\right)\frac{\gamma\mu\,(\gamma-1)}{\mu\,(\gamma-1)+1} > 0.$$

Moreover, democracy is more growth enhancing the closer the domestic country is to the world technology frontier:

$$\frac{\partial^2 g_t}{\partial \beta\,\partial \mu} = \left(\bar{z} + \frac{\partial \bar{z}}{\partial \beta}\beta\right)\frac{(\gamma-1)\,\gamma}{[\mu\,(\gamma-1)+1]^2} > 0.$$

This result is quite intuitive. Democratization allows for more turnover which in turn encourages outsiders to innovate and replace the incumbents. Since frontier countries rely more on innovation and benefit less from imitation or spillover, the result follows.

1.5.3.2 Evidence

A first piece of evidence supporting Prediction 5 is provided by Aghion et al. (2007), henceforth AAT. The paper uses employment and productivity data at the industry level across countries and over time. Their sample includes 28 manufacturing sectors for 180 countries over the period 1964–2003. Democracy is measured using the Polity 4 indicator, which itself is constructed from combining constraints on the executive; the openness and competitiveness of executive recruitment; and the competitiveness of political participation. Frontierness is measured by the log of the value added of a sector divided by the maximum of the log of the same variable in the same sectors across all countries or

[22] Here we make use of the assumption that backward sectors are automatically upgraded as the technology frontier moves up.

by ratio of the log of GDP per worker in the sector over the maximum of the log of per capita GDP in similar sectors across all countries. AAT take one minus these ratios as proxies for a sector's distance to the technological frontier. AAT focus on 5-year and 10-year growth rates. They compute rates over non-overlapping periods and in particular 5-year growth rates are computed over the periods 1975, 1980, 1985, 1990, 1995, and 2000. For the 10-year growth rates they use alternatively the years 1975, 1985, 1995, and the years 1980, 1990, and 2000.

AAT regress the growth of either value added or employment in an industrial sector on democracy (and other measures of civil rights), the country's or industry's frontierness, and the interaction term between the latter two. AAT also add time, country, and industry fixed effects.

The result is that the interaction coefficient between frontierness and democracy is positive and significant, meaning that the more frontier the industry is, the more growth-enhancing is democracy in the country for that sector. Figure 1.9 below provides an illustration of the results. It plots the rate of value-added growth against a measure of the country's proximity to the technological frontier (namely, the ratio of the country's labor productivity to the frontier labor productivity). The dotted line shows the linear regression of industry growth on democracy for countries that are less democratic than the median country (on the democracy scale), whereas the solid line shows the corresponding relationship for countries that are more democratic than the median country. We see that growth is higher in more democratic countries when they are close to the technological frontier, but not when they are far below the frontier.

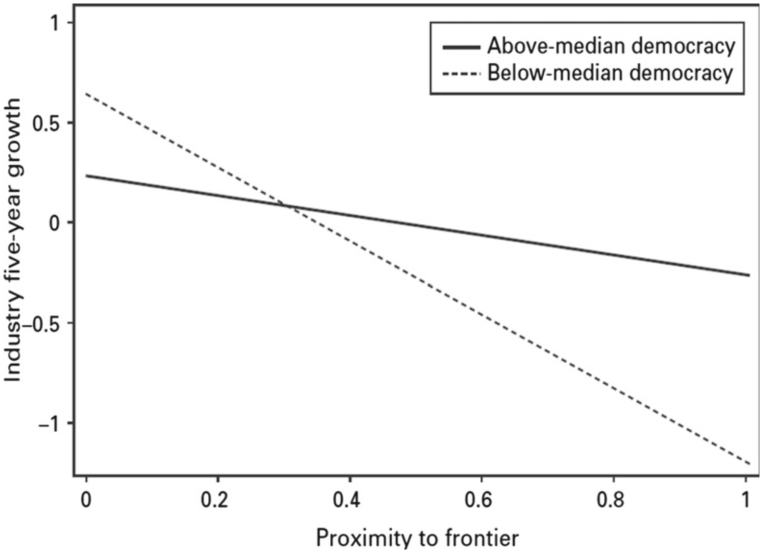


Figure 1.9 Growth, democracy, and distance to frontier (regression lines).

1.6. SCHUMPETERIAN WAVES

What causes long-term accelerations and slowdowns in economic growth and underlies the long swings sometimes referred to as Kondratieff cycles? In particular, what caused American growth in GDP and productivity to accelerate starting in the mid-1990s? The most popular explanation relies on the notion of general-purpose technologies (GPTs).

Bresnahan and Trajtenberg (1995) define a GPT as a technological innovation that affects production and/or innovation in many sectors of an economy. Well-known examples in economic history include the steam engine, electricity, the laser, turbo reactors, and more recently the information technology (IT) revolution. Three fundamental features characterize most GPTs. First, their pervasiveness: GPTs are used in most sectors of an economy and thereby generate palpable macroeconomic effects. Second, their scope for improvement: GPTs tend to underperform upon being introduced; only later do they fully deliver their potential productivity growth. Third, innovation spanning: GPTs make it easier to invent new products and processes—that is, to generate new secondary innovations—of higher quality.

Although each GPT raises output and productivity in the long run, it can also cause cyclical fluctuations while the economy adjusts to it. As David (1990) and Lipsey and Bekar (1995) have argued, GPTs like the steam engine, the electric dynamo, the laser, and the computer require costly restructuring and adjustment to take place, and there is no reason to expect this process to proceed smoothly over time. Thus, contrary to the predictions of real-business-cycle theory, the initial effect of a positive technology shock may not be to raise output, productivity, and employment, but to reduce them.[23]

Note that GPTs are Schumpeterian in nature, as they typically lead to older technologies in all sectors of the economy being abandoned as they diffuse to these sectors. Thus, it is no surprise that Helpman and Trajtenberg (1998) used the Schumpeterian apparatus to develop their model of GPT and growth. The basic idea of this model is that GPTs do not come ready to use off the shelf. Instead, each GPT requires an entirely new set of intermediate goods before it can be implemented. The discovery and development of these intermediate goods is a costly activity, and the economy must wait until some critical mass of intermediate components has been accumulated before it is profitable for firms to switch from the previous GPT. During the period between the discovery of a new GPT and its ultimate implementation, national income will fall as resources are taken out of production and put into R&D activities aimed at the discovery of new intermediate input components.

[23] For instance, Greenwood and Yorukoglu (1974) and Hornstein and Krusell (1996) have studied the productivity slowdown during the late 1970s and early 1980s caused by the IT revolution.

1.6.1 Back to the Basic Schumpeterian Model

As a useful first step toward a growth model with GPT, let us go back to the basic Schumpeterian model laid out in Section 1.2, but present it somewhat differently. Recall that the representative household has linear utility and the final good is produced with a single intermediate product according to:

$$Y_t = A_t y^{\alpha},$$

where y is the flow of intermediate input and A is the productivity parameter measuring the quality of intermediate input y.

Each innovation results in an intermediate good of higher quality. Specifically, a new innovation multiplies the productivity parameter A_k by $\gamma > 1$, so that:

$$A_{k+1} = \gamma A_k.$$

Innovations in turn arrive discretely with Poisson rate λz, where z is the current flow of research.

In the steady state the allocation of labor between research and manufacturing remains constant over time, and is determined by the research-arbitrage equation:

$$\omega_k = \lambda \gamma v_k, \tag{1.27}$$

where the LHS of (1.27) is the productivity-adjusted wage $\omega_k \equiv w_k/A_k$, which a worker earns by working in the manufacturing sector; $v_k \equiv V_k/A_k$ is the productivity-adjusted value and $\lambda \gamma v_k$ is the expected reward from investing one unit flow of labor in research.[24] The productivity-adjusted value v_k of an innovation is in turn determined by the Bellman equation:

$$\rho v_k = \widetilde{\pi}(\omega_k) - \lambda z v_k, \tag{1.28}$$

where $\pi\,(\omega_k) = A_k\,[1-\alpha]\,\alpha^{\frac{1+\alpha}{1-\alpha}}\omega_k^{\frac{\alpha}{\alpha-1}}$ is the equilibrium profit and $\widetilde{\pi}(\omega_k) \equiv \pi\,(\omega_k)\,/A_k$ denotes the productivity-adjusted flow of monopoly profits accruing to a successful innovator and we used the fact that $r_t = \rho$. In (1.28) the term $(-\lambda z v)$ corresponds to the capital loss involved in being replaced by a subsequent innovator. In the steady state, the productivity-adjusted variables ω_k and v_k remain constant; therefore, the subscript k will henceforth be dropped.

The above arbitrage equation, which can now be re-expressed as:

$$\omega = \lambda \gamma \frac{\widetilde{\pi}(\omega)}{\rho + \lambda z},$$

[24] Equation (1.27) is just a rewrite of Equation (R) in Section 1.2. Recall that the latter is expressed as:

$$w_k = \lambda V_{k+1};$$

using the fact that $V_{k+1} = \gamma V_k$, this immediately leads to Equation (1.27).

the labor-market clearing condition:

$$y(\omega) + z = L,$$

where $y(\omega)$ is the manufacturing demand for labor, jointly determine the steady-state amount of research z as a function of the parameters $\lambda, \gamma, L, \rho, \alpha$.

In a steady state the flow of the final good produced between the kth and $(k+1)$th innovation is:

$$Y_k = A_k \left[L - z\right]^{\alpha}.$$

Thus, the log of final output increases by $\ln \gamma$ each time a new innovation occurs. Then the average growth rate of the economy is equal to the size of each step $\ln \gamma$ times the average number of innovations per unit of time, λz: i.e.:

$$\mathbb{E}\left(g\right) = \lambda z \ln \gamma.$$

Note that this is a one-sector economy where each innovation corresponds by definition to a major technological change (i.e. to the arrival of a new GPT), and thus where growth is uneven with the time path of output being a random step function. But although it is uneven, the time path of aggregate output does not involve any slump. Accounting for the existence of slumps requires an extension of the basic Schumpeterian model, which brings us to the GPT growth model.

1.6.2 A Model of Growth with GPTs

As before, there are L workers who can engage either in the production of existing intermediate goods or in research aimed at discovering new intermediate goods. Again, each intermediate good is linked to a particular GPT. We follow Helpman and Trajtenberg (1998) in supposing that before any of the intermediate goods associated with a GPT can be used profitably in the final-goods sector, some minimal number of them must be available. We lose nothing essential by supposing that this minimal number is one. Once the good has been invented, its discoverer profits from a patent on its exclusive use in production, exactly as in the basic Schumpeterian model reviewed earlier.

Thus, the difference between this model and our basic model is that now the discovery of a new generation of intermediate goods comes in two stages. First, a new GPT must come, and then the intermediate good must be invented that implements that GPT. Neither can come before the other. You need to see the GPT before knowing what sort of good will implement it, and people need to see the previous GPT in action before anyone can think of a new one. For simplicity we assume that no one directs R&D toward the discovery of a new GPT. Instead, the discovery arrives as a serendipitous by-product of learning-by-doing with the previous one.

The economy will pass through a sequence of cycles, each having two phases, as indicated in Figure 1.10. GPT$_i$ arrives at time t_i. At that time, the economy enters phase

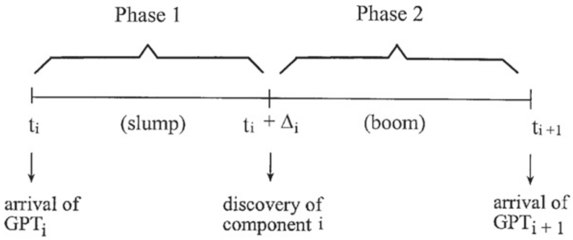


Figure 1.10 Phases of GPT cycles.

1 of the *i*th cycle. During phase 1, the amount z of labor is devoted to research. Phase 2 begins at time $t_i + \Delta_i$ when this research discovers an intermediate good to implement GPT$_i$. During phase 2, all labor is allocated to manufacturing until GPT$_{i+1}$ arrives, at which time the next cycle begins. Over the cycle, output is equal to $A_{i-1}F(L-z)$ during phase 1 and to $A_iF(L)$ during phase 2. Thus, the drawing of labor out of manufacturing and into research causes output to fall each time a GPT is discovered, by an amount equal to $A_{i-1}[F(L) - F(L-z)]$.

A steady-state equilibrium is one in which people choose to do the same amount of research each time the economy is in phase 1; that is, z is constant from one GPT to the next. As before, we can solve for the equilibrium value of z using a research-arbitrage equation and a labor-market-equilibrium condition. Let ω_j be the (productivity-adjusted) wage, and v_j the expected (productivity-adjusted) present value of the incumbent (intermediate good) monopolist when the economy is in phase $j \in \{1, 2\}$. In a steady state these productivity-adjusted variables will all be independent of which GPT is currently in use.

Because research is conducted in phase 1 but pays off when the economy enters into phase 2 with a productivity parameter raised by the factor γ, the following research-arbitrage condition must hold in order for there to be a positive level of research in the economy:

$$\omega_1 = \lambda\gamma v_2.$$

Suppose that once we are in phase 2, the new GPT is delivered by a Poisson process with constant arrival rate μ. Then the value v_2 is determined by the Bellman equation:

$$\rho v_2 = \widetilde{\pi}(\omega_2) + \mu\left[v_1 - v_2\right].$$

By analogous reasoning, we have:

$$\rho v_1 = \widetilde{\pi}(\omega_1) - \lambda z v_1.$$

Combining the above three equations yields the research-arbitrage equation:

$$\omega_1 = \frac{\lambda\gamma}{\rho+\mu}\left[\widetilde{\pi}(\omega_2) + \frac{\mu\widetilde{\pi}(\omega_1)}{\rho+\lambda z}\right]. \tag{1.29}$$

Because no one does research in phase 2, we know that the value of ω_2 is determined independently of research, by the market clearing condition:

$$L = \gamma(\omega_2).$$

Thus, we can take this value as given and regard the preceding research-arbitrage condition (1.29) as determining ω_1 as a function of z. The value of z is then determined, as in the previous subsection, by the labor-market equation:

$$L - z = \gamma(\omega_1).$$

The average growth rate will be the frequency of innovation times the size $\ln \gamma$, for exactly the same reason as in the basic model. The frequency, however, is determined a little differently than before because the economy must pass through two phases. An innovation is implemented each time a full cycle is completed. The frequency with which this implementation occurs is the inverse of the expected length of a complete cycle. This in turn is just the expected length of phase 1 plus the expected length of phase 2: $1/\lambda z + 1/\mu = [\mu + \lambda z] / \mu\lambda z$. Thus, the growth rate will be:

$$g = \ln \gamma \frac{\mu\lambda z}{\mu + \lambda z}$$

which is positively affected by anything that raises research. Note also that growth tapers off in the absence of the arrival of new GPTs, i.e. if $\mu = 0$. This leads Gordon (2012) to predict a durable slowdown of growth in the US and other developed economies as the ITC revolution is running out of steam.

The size of the slump $\ln(F(L)) - \ln(F(L - z))$ that occurs when each GPT arrives is also an increasing function of z, and hence it will tend to be positively correlated with the average growth rate.

One further property of this cycle worth mentioning is that the wage rate will rise when the economy goes into a slump. That is, because there is no research in phase 2, the normalized wage must be low enough to provide employment for all L workers in the manufacturing sector; whereas, with the arrival of the new GPT, the wage must rise to induce manufacturers to release workers into research. This brings us directly to the next subsection on wage inequality.

1.6.3 GPT and Wage Inequality

In this subsection we show how the model of the previous section can account for the rise in the skill premium during the IT revolution. We modify that model by assuming that there are two types of labor. Educated labor can work in both research and manufacturing, whereas uneducated labor can only work in manufacturing. Let L^s and L^u denote the supply of educated (skilled) and uneducated (unskilled) labor, let ω_1^s and ω_1^u denote their respective productivity-adjusted wages in phase 1 of the cycle (when research activities on complementary inputs actually take place), and let ω_2 denote the productivity-adjusted wage of labor in phase 2 (when new GPTs have not yet appeared and therefore labor is entirely allocated to manufacturing).

If in equilibrium the labor market is segmented in phase 1, with all skilled labor being employed in research while unskilled workers are employed in manufacturing, we have the labor-market-clearing conditions:

$$L^s = z, \quad L^u = \gamma(\omega_1^u), \quad \text{and} \quad L^s + L^u = \gamma(\omega_2),$$

and the research-arbitrage condition:

$$\omega_1^s = \lambda\gamma v_2, \tag{1.30}$$

where v_2 is the productivity-adjusted value of an intermediate producer in stage 2. This value is itself determined as before by the two Bellman equations:

$$\rho v_2 = \widetilde{\pi}(\omega_2) + \mu\left[v_1 - v_2\right],$$

and:

$$\rho v_1 = \widetilde{\pi}(\omega_1^u) - \lambda z v_1.$$

Thus, the above research-arbitrage Equation (1.30) expresses the wage of skilled labor as being equal to the expected value of investing (skilled) labor in R&D for discovering complementary inputs to the new GPT.

The labor market will be truly segmented in phase 1, if and only if, ω_1^s defined by research-arbitrage condition (1.30) satisfies:

$$\omega_1^s > \omega_1^u,$$

which in turn requires that L^s not be too large. Otherwise the labor market remains unsegmented, with $z < L^s$ and:

$$\omega_1^s = \omega_1^u,$$

in equilibrium. In the former case, the arrival of a new GPT raises the skill premium (from 0 to $\omega_1^s/\omega_1^u - 1$) at the same time as it produces a productivity slowdown because labor is driven out of production.

1.6.4 Predictions

The above GPT model delivers the following predictions.[25]

[25] While Jovanovic and Rousseau (2005) provide evidence for the first three predictions, we refer the reader to Acemoglu (2002, 2009), Aghion et al. (1999), and Aghion and Howitt (2009) for evidence on growth and wage inequality. In particular, Aghion and Howitt contrast the GPT explanation with alternative explanations based on trade, deunionization, or directed technical change considerations.

Prediction 1: *The diffusion of a new GPT is associated with an increase in the flow of firm entry and exit.*

This results from the fact that the GPT is Schumpeterian in nature; thus it generates quality-improving innovations, and therefore creative destruction, in any sector of the economy where it diffuses.

Prediction 2: *The arrival of a new GPT generates a slowdown in productivity growth; this slowdown is mirrored by a decline in stock-market prices.*

The diffusion of a new GPT requires complementary inputs and learning, which may draw resources from normal production activities and may contribute to future productivity in a way that cannot be captured easily by current statistical indicators. Another reason why the diffusion of a new GPT should reduce growth in the shortrun is by inducing the obsolescence of existing capital in the sectors it diffuses to (see Aghion and Howitt, 1998, 2009).

Prediction 3: *The diffusion of a new GPT generates an increase in wage inequality both between and within educational groups.*

An increase in the skill premium occurs as more skilled labor is required to diffuse a new GPT to the economy, as we saw above. The other and perhaps most intriguing feature of the upsurge in wage inequality is that it took place to a large extent within control groups, no matter how narrowly those groups are identified (e.g. in terms of experience, education, gender, industry, occupation). One explanation is that skill-biased technical change enhanced not only the demand for observed skills as described earlier but also the demand for unobserved skills or abilities. Although theoretically appealing, this explanation is at odds with econometric work (Blundell and Preston, 1999) showing that the within-group component of wage inequality in the United States and the United Kingdom is mainly transitory, whereas the between-group component accounts for most of the observed increase in the variance of permanent income. The explanation based on unobserved innate abilities also fails to explain why the rise in within-group inequality has been accompanied by a corresponding rise in individual wage instability (see Gottschalk and Moffitt, 1994). Using a GPT approach, Aghion et al. (2002) argue that the diffusion of a new technological paradigm can affect the evolution of within-group wage inequality in a way that is consistent with these facts. The diffusion of a new GPT raises within-group wage inequality primarily because the rise in the speed of embodied technical progress associated with the diffusion of the new GPT increases the market premium to those workers who adapt quickly to the leading-edge technology and are therefore able to survive the process of creative destruction at work as the GPT diffuses to the various sectors of the economy.[26]

[26] In terms of the preceding model, let us again assume that all workers have the same level of education but that once a new GPT has been discovered, only a fraction α of the total labor force can adapt quickly enough to the new technology so that they can work on looking for a new component that complements the GPT. The other workers, who did not successfully adapt have no alternative but to work in

1.7. CONCLUSION

In this paper, we argued that Schumpeterian growth theory—where current innovators exert positive knowledge spillovers on subsequent innovators as in other innovation-based models, but where current innovators also drive out previous technologies—generates predictions and explains facts about the growth process that could not be accounted for by other theories.

In particular, we saw how Schumpeterian growth theory manages to put IO into growth and to link growth with firm dynamics, thereby generating predictions on the dynamic patterns of markets and firms (entry, exit, reallocation, etc.) and on how these patterns shape the overall growth process. These predictions and the underlying models can be confronted with micro data and this confrontation in turn helps refine the models. This back-and-forth communication between theory and data has been key to the development of the Schumpeterian growth theory over the past 25 years.[27]

Also, we argued that Schumpeterian growth theory helps us reconcile growth with development, in particular, by bringing out the notion of appropriate growth institutions and policies, i.e. the idea that what drives growth in a sector (or country) far below the world technology frontier is not necessarily what drives growth in a sector or country at the technological frontier where creative destruction plays a more important role. In particular, we pointed to democracy being more growth enhancing in more frontier economies. The combination of the creative destruction and appropriate growth institutions ideas also underlies the view[28] that "extractive economies," where creative destruction is deterred by political elites, are more likely to fall into low-growth traps.

manufacturing. Let ω_1^{adapt} denote the productivity-adjusted wage rate of adaptable workers in phase 1 of the cycle, and let ω_1 denote the wage of non-adaptable workers. Labor market clearing implies: $\alpha L = z$; $[1-\alpha]\, L = \gamma(\omega_1)$; $L = \gamma(\omega_2)$, whereas research arbitrage for adaptable workers in phase 1 implies $\omega_1^{adapt} = \lambda \gamma v_2$. When α is sufficiently small, the model generates a positive adaptability premium: $\omega_1^{adapt} > \omega_1$.

[27] For example, when analyzing the relationship between growth and firm dynamics, this back-and-forth process amounts to what one might call a layered approach. Here, we refer the reader to Daron Acemoglu's panel discussion at the Nobel Symposium on Growth and Development (September 2012). The idea here is that of a step-by-step estimation method, where at each step a small subset of parameters are being identified in their neighborhood. Thanks to the rich set of available micro data, one can first identify a parameter and its partial equilibrium effect as well as some of its industry equilibrium effects. Next, one can test the predictions of the model using moments in the data that were not directly targeted in the original estimation. Then one can check that the model also satisfies various out-of-sample properties and reach a macro-aggregation by building on detailed micro moments. Schumpeterian models are well suited for this type of approach as they are able to generate realistic firm dynamics with tractable aggregations.

[28] See Acemoglu and Robinson (2012).

Beyond enhancing our understanding of the growth process, Schumpeterian growth theory is useful in at least two respects. First, as a tool for the design of growth policy: departing from the Washington consensus view whereby the same policies should be recommended everywhere, the theory points to appropriate growth policies, i.e. policies that match the particular context of a country or region. Thus, we saw that more intense competition (lower entry barriers), a higher degree of trade openness, and more emphasis on research education are all more growth-enhancing in more frontier countries.[29]

The Schumpeterian growth paradigm also helps us assess the relative magnitude of the counteracting partial equilibrium effects pointed out by the theoretical IO literature. For example, there is a whole literature on competition, investments, and incentives[30] that points to counteracting partial equilibrium effects without saying much about when one particular effect should be expected to prevail. In contrast, Section 1.3 illustrated how aggregation and the resulting composition effect could help determine under which circumstances the escape–competition effect would dominate the counteracting Schumpeterian effect. Similarly, Section 1.4 showed the importance of reallocation for growth; thus, policies supporting entry or incumbent R&D could contribute positively to economic growth in partial equilibrium, yet in general equilibrium Section 1.4 showed that this is done at the expense of reduced innovation by the rest of the economy.

Where do we see the Schumpeterian growth agenda being pushed over the next years? A first direction is to look more closely at how growth and innovation are affected by the organization of firms and research. Thus, over the past 5 years Nick Bloom and John Van Reenen have popularized fascinating new datasets that allow us to look at how various types of organizations (e.g. more or less decentralized firms) are more or less conducive to innovation. But firms' size and organization are in turn endogenous, and in particular, they depend on factors such as the relative supply of skilled labor or the nature of domestic institutions. Future studies should try to model and then test the relationship from skill endowment and the institutional environment to firm organization and then from firm organization to innovation and growth.

A second and related avenue for future research is to look at growth, firm dynamics, and reallocation in developing economies. Recent empirical evidence (see Hsieh and Klenow, 2009, 2012) has shown that the misallocation of resources is a major source of the productivity gap across countries. What are the causes of misallocation, and why do these countries lack creative destruction that would eliminate the inefficient firms? Schumpeterian theory with firm dynamics could be an invaluable source to shed light on these important issues that lie at the core of the development puzzle.

A third avenue is to look at the role of finance in the growth process. In Section 1.5 we pointed to equity finance being more growth-enhancing in more frontier economies.

[29] Parallel studies point to labor market liberalization and stock-market finance being more growth-enhancing in more advanced countries or regions.

[30] See the recent analytical surveys by Gilbert (2006), Vives (2008), and Schmutzler (2010).

More generally, we still need to better understand how different types of financial instruments map with different sources of growth and different types of innovation activities. Also, we need to better understand why we observe a surge of finance during the acceleration phase in the diffusion of new technological waves, as mentioned in Section 1.6, and how financial sectors evolve when the waves taper off. These and many other microeconomic aspects of innovation and growth await further research.

ACKNOWLEDGMENT

This survey builds on a presentation at the Nobel Symposium on Growth and Development (September 2012) and was subsequently presented as the Schumpeter Lecture at the Swedish Entrepreneurship Forum (January 2013). We thank Pontus Braunerhjelm, Mathias Dewatripont, Sam Kortum, Michael Spence, John Van Reenen, David Warsh, and Fabrizio Zilibotti for helpful comments and encouragements, and Sina Ates, Salome Baslandze, Jonathan Roth, and Felipe Saffie for outstanding editing work. This paper is produced as part of the project Growth and Sustainability Policies for Europe (GRASP), a collaborative project funded by the European Commission's Seventh Research Framework Programme, Contract No. 244725.

REFERENCES

Acemoglu, D., 2002. Technical change, inequality, and the labor market. Journal of Economic Literature 40, 7–72.

Acemoglu, D., 2009. Introduction to Modern Economic Growth. Princeton University Press.

Acemoglu, D., Aghion, P., Zilibotti, F., 2006. Distance to Frontier, Selection, and economic growth. Journal of the European Economic Association, 37–74.

Acemoglu, D., Akcigit, U., 2012. Intellectual property rights policy, competition and innovation. Journal of the European Economic Association 10, 1–42.

Acemoglu, D., Akcigit, U., Bloom, N., Kerr, W., 2013. Innovation, Reallocation and Growth. NBER Working Paper 18993.

Acemoglu, D., Akcigit, U., Hanley, D., Kerr, W., 2012. Transition to Clean Technology. University of Pennsylvania, mimeo.

Acemoglu, D., Cao, D., 2011. Innovation by Entrants and Incumbents. MIT, mimeo.

Acemoglu, D., Robinson, J., 2006. Economic backwardness in political perspective. American Political Science Review 100, 115–131.

Acemoglu, D., Robinson, J., 2012. Why Nations Fail, Crown Business.

Acs, Z., Audretsch, D., 1988. Innovation in large and small firms: an empirical analysis. American Economic Review 78, 678–690.

Acs, Z., Audretsch, D., 1991. Innovation and size at the firm level. Southern Economic Journal 57, 739–744.

Aghion, P., Alesina, A., Trebbi, F., 2007. Democracy, technology and growth. In: Helpman, E. (Eds.), Institutions and Economic Performance, Cambridge University Press.

Aghion, P., Bloom, N., Blundell, R., Griffith, R., Howitt, P., 2005. Competition and innovation: an inverted-U relationship. Quarterly Journal of Economics 120, 701–728.

Aghion, P., Blundell, R., Griffith, R., Howitt, P., Prantl, S., 2009. The effects of entry on incumbent innovation and productivity. Review of Economics and Statistics 91, 20–32.

Aghion, P., Boustan, L. Hoxby, C., Vandenbussche, J., 2009. Exploiting States' Mistakes to Identify the Causal Effects of Higher Education on Growth. Harvard, mimeo.

Aghion, P., Caroli, E., Garcia-Penalosa, C., 1999. Inequality and economic growth: the perspective of the new growth theories. Journal of Economic Literature 37, 1615–1660.

Aghion, P., Griffith, R., 2006. Competition and Growth: Reconciling Theory and Evidence, MIT Press.

Aghion, P., Harris, C., Howitt, P., Vickers, J., 2001. Competition, imitation and growth with step-by-step innovation. Review of Economic Studies 68, 467–492.

Aghion, P., Harris, C., Vickers, J., 1997. Competition and growth with step-by-step innovation: an example. European Economic Review, Papers and Proceedings, 771–782.

Aghion, P., Howitt, P., 1992. A model of growth through creative destruction. Econometrica 60, 323–351.

Aghion, P., Howitt, P., 1998. Endogenous Growth Theory, MIT Press.

Aghion, P., Howitt, P., 2009. The Economics of Growth, MIT Press.

Aghion, P., Howitt, P., Prantl, S., 2013. Patent Rights, Product Market Reforms and Innovation. NBER Working Paper 18854.

Aghion, P., Howitt, P., Violante, G., 2002. General purpose technology and wage inequality. Journal of Economic Growth 7, 315–345.

Akcigit, U., Hanley, D., Serrano-Velarde, N., 2012. Back to Basics: Basic Research Spillovers, Innovation Policy and Growth. University of Pennsylvania, mimeo.

Akcigit, U., Kerr, W., 2010. Growth Through Heterogeneous Innovations. NBER Working Paper 16443.

Alesina, A., Rodrik, D., 1994. Distributive politics and economic growth. Quarterly Journal of Economics 109, 465–490.

Arrow, K., 1962. The economic implications of learning by doing. Review of Economic Studies 29, 155–173.

Axtell, R., 2001. Zipf distribution of US firm sizes. Science 293, 1818–1820.

Barro, R.J., Sala-i-Martin, X., 2003. Economic Growth. McGraw-Hill.

Bartelsman, E., Doms, M., 2000. Understanding productivity: lessons from longitudinal microdata. Journal of Economic Literature 38, 569–594.

Bartelsman, E., Haltiwanger, J., Scarpetta, S., 2009. Cross-country Differences in Productivity: The Role of Allocation and Selection NBER. Working Paper 15490.

Birch, D., 1981. Who Creates Jobs? The Public Interest 65, 3–14.

Birch, D., 1987. Job Creation in America: How Our Smallest Companies Put the Most People to Work. Free Press.

Blundell, R., Griffith, R., Van Reenen, J., 1995. Dynamic Count Data Models of Technological Innovation. Economic Journal 105, 333–344.

Blundell, R., Griffith, R., Van Reenen, J., 1999. Market share, market value and innovation in a panel of British manufacturing firms. Review of Economic Studies 66, 529–554.

Blundell, R., Preston, I., 1999. Inequality and Uncertainty: Short-Run Uncertainty and Permanent Inequality in the US and Britain. University College London, mimeo.

Boldrin, M., Levine, D., 2008. Against Intellectual Monopoly. Cambridge University Press.

Bresnahan, T., Trajtenberg, M., 1995. General purpose technologies: engines of growth? Journal of Econometrics, 65, 83–108.

Cohen, W., 1995. Empirical studies of innovative activity. In: Stoneman, Paul (Ed.), Handbook of the Economics of Innovations and Technological Change, Blackwell.

Cohen, W., Klepper, S., 1996. Firm size and the nature of innovation within industries: the case of process and product R&D. Review of Economics and Statistics, 232–243.

Corriveau, L., 1991. Entrepreneurs, Growth, and Cycles. PhD Dissertation, University of Western, Ontario.

David, P., 1990. The dynamo and the computer: an historical perspective on the modern productivity paradox. American Economic Review 80, 355–361.

Davis, S., Haltiwanger, J., Schuh, S., 1996. Job Creation and Destruction, MIT Press.

De Loecker, J., Goldberg P., Khandelwal, A., Pavcnik N., 2012. Prices, Markups and Trade Reform. Yale, Mimeo.

Djankov, S., La Porta, R., Lopez-de-Silanes, F., Shleifer, A., 2002. The regulation of entry. Quarterly Journal of Economics 117, 1–37.

Foster, L., Haltiwanger, J., Krizan, C., 2001. Aggregate productivity growth: lessons from microeconomic evidence. In: Dean, E., Harper, M., Hulten, C. (Eds.), New Directions in Productivity Analysis. University of Chicago Press.

Foster, L., Haltiwanger, J., Krizan, C., 2006. Market selection, reallocation, and restructuring in the U.S. retail trade sector in the 1990s. Review of Economics and Statistics 88, 748–758.

Frankel, J., Romer, D., 1999. Does trade cause growth? American. Economic Review 89, 379–399.

Freeman, C., 1982. The Economics of Industrial Innovation. MIT Press.

Galor, O., 2011. Unified Growth Theory. Princeton University Press.

Gerschenkron, A., 1962. Economic Backwardness in Historical Perspective: A Book of Essays. Belknap Press of Harvard University Press.

Gilbert, R., 2006. Looking for Mr Schumpeter: Where Are We in the Competition-Innovation Debate? In: Lerner, J., Stern, S. (Eds.), Innovation Policy and Economy. NBER, MIT Press.

Goldberg, P., Khandelwal, A., Pavcnik, N., Topalova, P., 2010. Imported intermediate inputs and domestic product growth: evidence from India. Quarterly Journal of Economics, 125, 1727–1767.

Gordon, R. 2012. Is U.S. Economic Growth Over? Faltering Innovation Confronts the Six Headwinds. NBER Working Paper 18315.

Gottschalk, P., Moffitt, R., 1994. The growth of earnings instability in the US labor market. Brookings Papers on Economic Activity 2, 217–272.

Greenwood, J., Yorukoglu, M., 1974. Carnegie-Rochester Series on Public Policy 46, 49–95.

Griliches, Z., 1990. Patent statistics as economic indicators: a survey. Journal of Economic Literature 28, 1661–1707.

Grossman, G., Helpman, E., 1991. Quality ladders in the theory of growth. Review of Economic Studies 58, 43–61.

Hall, B., Jaffe, A., Trajtenberg, M., 2001. The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools. NBER Working Paper 8498.

Haltiwanger, J., Jarmin, R., Miranda, J., 2010. Who Creates Jobs? Small vs. Large vs. Young. NBER Working Paper 16300.

Helpman, E., Trajtenberg, M., 1998. A time to sow and a time to reap: growth based on general purpose technologies. In: Helpman, E. (Eds.), General Purpose Technologies and Economic Growth, MIT Press.

Hornstein, A., Krusell, P., 1996. Can Technology Improvements Cause Productivity Slowdowns? NBER Macroeconomics Annual, 11, 209–259.

Howitt, P., 2000. Endogenous growth and cross-country income differences. American Economic Review 90, 829–46.

Hsieh, C., Klenow, P., 2009. Misallocation and manufacturing TFP in China and India. Quarterly Journal of Economics 124, 1403–1448.

Hsieh, C., Klenow, P., 2012. The Life Cycle of Plants in India and Mexico. NBER Working Paper 18133.

Ijiri, Y., Simon, H., 1977. Skew Distributions and the Sizes of Business Firms. North-Holland.

Jones, C., Vollrath, D., 2013. Introduction to Economic Growth. W. W. Norton & Company.

Jovanovic, B., Rousseau, P., 2005. General purpose technologies. In: Aghion, P., Durlauf, S. (Eds.), Handbook of Economic Growth. Elsevier, North-Holland.

Klette, T., Kortum, S., 2004. Innovating firms and aggregate innovation. Journal of Political Economy 112, 986–1018.

Lentz, R., Mortensen, D., 2008. An empirical model of growth through product innovation. Econometrica 76, 1317–1373.

Lipsey, R., Bekar, C., 1995. A structuralist view of technical change and economic growth. Bell Canada Papers on Economic and Public Policy 3, 9–75.

Nelson, R., Phelps, E., 1966. Investment in humans, technological diffusion, and economic growth. American Economic Review 61, 69–75.

Neumark, D., Wall, B., Zhang, J., 2008. Do Small Businesses Create More Jobs? New Evidence for the United States from the National Establishment Time Series. NBER Working Paper 13818.

Nickell, S., 1996. Competition and corporate performance. Journal of Political Economy, 104, 724–746.

Pennings, J.M., Buitendam, A., 1987. New Technology As Organizational Innovation: The Development and Diffusion of Microelectronics. Bollinger.

Peretto, P., 1998. Technological change, market rivalry, and the evolution of the capitalist engine of growth. Journal of Economic Growth 3, 53–80.

Persson, T., Tabellini, G., 1994. Is inequality harmful for growth? American. Economic Review 84, 600–621.

Qian, Y., 2007. Do national patent laws stimulate domestic innovation in a global patenting environment? Review of Economics and. Statistics 89, 436–453.

Romer, P., 1990. Endogenous technical change. Journal of Political Economy 98, 71–102.

Rossi-Hansberg, E., Wright, M., 2007. Establishment size dynamics in the aggregate economy. American Economic Review 97, 1639–1666.

Scherer, F.M., 1984. Innovation and Growth: Schumpeterian Perspectives. MIT Press.

Schmalensee, R., 1989. Inter-industry studies of structure and performance. In: Schmalensee, R., Willig, R.D. (Eds.), Handbook of Industrial Organization, vol. 2. North-Holland.

Schmutzler, A., 2010. Is competition good for innovation? a simple approach to an unresolved question. Foundations and Trends in Microeconomic Analysis 5, 355–428.

Segerstrom, P., Anant, T., Dinopoulos, E., 1990. A schumpeterian model of the product cycle. American Economic Review 88, 1077–1092.

Simon, H., Bonini, C., 1958. The size distribution of business firms. American Economic Review 48, 607–17.

Sivadasan, J., 2009. Barriers to competition and productivity: evidence from India. B.E. Journal of Economic Analysis and Policy 9, 42.

Stanley, M., Buldyrev, S., Havlin, S., Mantegna, R., Salinger, M., Stanley, E., 1995. Zipf plots and the size distribution of firms. Economic Letters 49, 453–57.

Syverson, C., 2011. What determines productivity. Journal of Economic Literature 49, 326–365.

Tirole, J., 1988. The Theory of Industrial Organization. MIT Press.

Topalova, P., Khandelwal, A.K., 2011. Trade liberalization and firm productivity: the case of India. Review of Economics and Statistics 93, 995–1009.

Tushman, M.L., Anderson, P., 1986. Technological discontinuities and organizational environments. Administrative Science Quarterly 31, 439–465.

Vandenbussche, J., Aghion, P., Meghir, C., 2006. Growth, distance to frontier, and composition of human capital. Journal of Economic Growth 11, 97–127.

Vives, X., 2008. Innovation and competitive pressure. Journal of Industrial Economics 56, 419–469.

Weil, D., 2012. Economic Growth. Prentice Hall.

CHAPTER TWO

Technology Diffusion: Measurement, Causes, and Consequences

Diego Comin*[*] and Martí Mestieri[†]
[*]Harvard University, NBER and CEPR, USA
[†]Toulouse School of Economics, France

Abstract

This chapter discusses different approaches pursued to explore three broad questions related to technology diffusion: what general patterns characterize the diffusion of technologies, and how have they changed over time?; what are the key drivers of technology?; and what are the macroeconomic consequences of technology? We prioritize in our discussion unified approaches to these three questions that are based on direct measures of technology.

Keywords

Technology Measurement, Technology Diffusion, Innovation, Medium Term Business Cycles, Innovation

JEL Classification Codes

O31, O33, O41, O57

2.1. INTRODUCTION

> **tech·no·lo·gy**, *noun: a manner of accomplishing a task especially using technical processes, methods, or knowledge.*
>
> ***The Merriam-Webster's Collegiate Dictionary***

New technologies take the form of new production processes, new tools, and new and higher quality goods and services. Following the seminal work of Solow (1956), there is a wide consensus that advances in technology are a key source of economic growth over the long term. Many of these advances result, directly or indirectly, from purposeful investments in research and development (R&D), as pointed out by the endogenous growth literature (e.g. Arrow, 1962; Romer, 1990; Aghion and Howitt, 1992).

R&D is not the only (or even the main) type of investment to upgrade technology. In fact, R&D investments are concentrated in a few countries (e.g. Keller, 2004). The overwhelming majority of governments and companies around the world do not engage in any significant R&D expenditures. Instead, most companies in the vast majority of countries are well behind the technology frontier. Their fundamental concern when upgrading their technology is to obtain access to better technologies that already exist

ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53540-5.00002-1

but they do not use yet. Hence, it is very important to understand technology adoption patterns for companies and countries.

Technology diffusion is the dynamic consequence of adoption. It characterizes the accumulation of technology across adopters and over time, which arises from individual adoption decisions. This chapter discusses different approaches pursued to explore three broad questions related to technology diffusion: first, what the patterns of technology diffusion are, and how they have changed over time; second, what factors affect technology diffusion; and third, what the macroeconomic consequences of technology diffusion are.

Several vast literatures that expand various disciplines have addressed some of these questions. Therefore, it is impossible to do justice to all this work in just a chapter. Rather than focusing on being comprehensive in answering one question (which has been done elsewhere),[1] we see greater value in presenting empirical strategies that have explored the three questions in a unified way. The other principle we use to guide our choice is to focus on works that use direct measures of technology.[2] Because these conditions are restrictive, our chapter does not intend to be a comprehensive survey.

The chapter is organized in three sections that coincide with the three questions we have outlined. Section 2.2 describes various approaches followed to measure technology diffusion and discusses their value and shortcomings. We pay special attention to attempts made to explore the evolution of adoption patterns over time as well as how they differ across countries. Section 2.3 explores factors identified as drivers of technology. Section 2.4 explores the macroeconomic consequences of technology, focusing mostly on how technology affects income dynamics at different frequencies. Section 2.5 concludes with some open questions for future research.

2.2. MEASUREMENT

Prior to studying diffusion patterns, we need to measure technology diffusion. The approaches developed to measure technology diffusion differ in terms of (i) the dimensions of technology they intend to measure, and (ii) the level at which they try to measure diffusion. In this section, we describe different existing measures of diffusion, as well as the main lessons from each approach.

2.2.1 Extensive Measures at the Country Level

Probably the simplest way to think about technology consists in tracking whether a specific technology is present or not in a given country at a moment in time. The data requirements to construct such measures are minimal. Country-level extensive measures are informative of the overall level of technology in a country if there is large cross-country

[1] See, for example, Metcalfe (1981, 1998), Stoneman (1983, 1987), Stoneman et al. (1995), Thirtle and Ruttan (1987), Karshenas and Stoneman (1993), and Vickery and Northcott (1995).

[2] See Coe and Helpman (1995) and Keller (2004) for analyses based on indirect technology measures.

variation in adoption lags. However, country-level extensive measures of adoption do not capture how intensively a technology is used once it is present in the country. As we show below, this condition makes country-level extensive measures of technology more relevant to study technology adoption patterns until around the beginning of the 20th century.

We know from the work of Maddison (2004) that cross-country income differences were relatively small until the Industrial Revolution. How large were cross-country differences in technology adoption in the distant past? Comin et al. (2010) take on this question by assembling three data sets with country-level extensive measures of technology adoption. Each data set reports the adoption patterns of the inhabitants of modern-day territories in different historical moments: 1000 BC, 0 AD, and 1500 AD. The first two are coded using 12 technologies from the *Atlas of Cultural Evolution* (Peregrine, 2003). The data set for 1500 AD covers 24 technologies coded by Comin et al. (2010). The technologies considered satisfy three criteria. First, they were state-of-the-art technologies (at the time considered); second, they were used in productive activities (i.e. activities that entered GDP); and third, it has been possible to document its presence or absence for a wide range of countries. In all three periods, the technologies can be classified in five broad sectors: agriculture, industry, transportation, communication, and military. For each technology, the data set measures whether it was present (1) or absent (0) from the relevant territory in the relevant period of time. Comin et al. (2010) compute country-sector adoption levels as the simple average of the binary adoption values across the technologies in the sector. Then, the overall adoption level is computed as the simple average of the sectoral adoption levels.

Table 2.1 presents the variation across continents in overall technology adoption. In all three historical periods, Europe and Asia present the highest average levels of overall technology adoption, while America and Oceania present the lowest, with Africa in between. The range of variation in the average adoption levels across continents suggests that technological differences were significant despite the wide consensus that cross-country variation in living standards was limited until the 19th century (e.g. Maddison, 2004). Similarly, there was significant within-continent variation in technology levels. Note that, given the binary nature of the underlying data, the maximum level the standard deviation can achieve is 0.5. The median standard deviation within continents (in all three periods) is 0.15. Table 2.2 shows that the cross-country variation in technology is larger than the cross-continent variation with a level for the standard deviation close to 0.3 in all three periods.

Finally, one relevant empirical question is whether all variation in technology is captured by the variation in the average technology levels in the country or whether there is significant variation in technology across sectors (within a country). Table 2.2 explores this question. In particular, it reports the cross-country dispersion of the deviation between the sectoral and the overall adoption levels. This dispersion ranges from 0.12 to 0.35 with a median value of 0.2. These magnitudes suggest that a significant fraction of the variation in technology adoption is driven by within-country differences in technology across sectors.

Table 2.1 Descriptive statistics of overall technology adoption by continent

Period	Continent	Obs.	Average	Std. Dev.	Min	Max
1000 BC	Europe	30	0.66	0.16	0.5	1
	Africa	34	0.36	0.31	0	1
	Asia	23	0.58	0.25	0.1	1
	America	24	0.24	0.12	0	0.4
	Oceania	2	0.2	0.14	0.1	0.3
0 AD	Europe	33	0.88	0.15	0.7	1
	Africa	40	0.77	0.2	0.6	1
	Asia	34	0.88	0.15	0.6	1
	America	25	0.33	0.17	0	0.6
	Oceania	3	0.17	0.11	0.1	0.3
1500 AD	Europe	26	0.87	0.074	0.69	1
	Africa	39	0.32	0.2	0.1	0.78
	Asia	25	0.66	0.19	0.07	0.88
	America	24	0.14	0.07	0	0.13
	Oceania	9	0.12	0.04	0	0.13

Table 2.2 Variation in technology adoption within countries vs. across countries

Period	Obs.	STD. across countries	STD. of deviations of sector level technology from overall technology adoption within countries				
		Overall	Agri.	Ind.	Military	Transp.	Comm.
1000 BC	114	0.28	0.35	0.18	0.16	0.22	0.23
0	136	0.28	0.25	0.18	0.26	0.24	0.32
1500 AD	125	0.32	0.2	0.19	0.13	0.12	0.17

Note: STD. Overall is the cross-country standard deviation in overall technology adoption level STD. of deviations of sector level technology from overall technology adoption is computed as follows: $\sigma(xsct - xct)$ where $\sigma(z)$ represents the standard deviation of z across countries, $xsct$ is the level of technology in sector s, country c, and period t, and xct denotes the overall adoption level in country c in period t, the average of the adoption levels by sector for country c in period t.

2.2.2 Traditional Measures of Technology Diffusion

It is possible to extend extensive measures of technology diffusion to more disaggregated levels to study how producers have access to a technology once it has arrived to a country. Let's suppose that potential adopters have a binary choice of whether to incur in a sunk cost of adopting the technology. After they incur in such a cost, they can use the technology indefinitely at no extra cost. Let's define Y_t as:

$$Y_t = \frac{m_t}{M}, \tag{2.1}$$

where M is the (fixed) number of potential adopters and m_t is the number of producers that have adopted the technology at time t. This is how the diffusion literature has measured diffusion traditionally.

The traditional diffusion literature has fitted S-shaped diffusion curves (like the logistic function) to diffusion measures such as Y_t (Griliches, 1957; Mansfield, 1961; Gort and Klepper, 1982). For future reference, the logistic is defined by:

$$L_t = \frac{\delta_1}{1 + e^{-(\delta_2 + \delta_3 t)}}, \tag{2.2}$$

where t represents time, δ_3 reflects the speed of adoption, δ_2 is a constant of integration that positions the curve on the time scale, and δ_1 is the long-run outcome.

Several features of this curve are relevant. The logistic curve summarizes the process of technology diffusion in just three parameters (δ_1, δ_2, and δ_3). It asymptotes to 0 when t goes to minus infinity and to δ_1 when t goes to infinity. Finally, it is symmetric around the inflection point of $L_t = \delta_1/2$ which occurs at $t = -\delta_2/\delta_3$.

Logistic or S-shaped curves have been fitted to technology measures such as (2.1) for technologies in many sectors and various countries. Examples of technologies explored in diffusion studies include the hybrid corn in US states (Griliches, 1957), β-blockers in US states (Skinner and Staiger, 2007), tetracycline among physicians in four US cities (Coleman and Menzel, 1966), 22 manufacturing processes and machines in the UK (Davies, 1979), and various consumer durables in the US (Cox and Alm, 1996). The main finding of the traditional diffusion literature is that S-shaped curves such as (2.2) provide a good fit to traditional diffusion measures of the form (2.1).

The slow initial pace that characterizes logistic diffusion patterns has motivated a number of theories about the drivers of diffusion.[3] Epidemic models (e.g. Griliches, 1957; Mansfield, 1961, 1963; Romeo, 1975; Dixon, 1980; Davies, 1979; Levin et al. 1987; Rose and Joskow, 1990) build on the premise that the lack of information on the technology prevents potential adopters from adopting profitable technologies. Information, in turn, is spread slowly because it only flows from those agents that have already adopted the technology. The so-called probit model builds on firms' heterogeneity in adoption costs or profits to generate heterogeneity in the timing of adoption.[4] A third class of models that deliver S-shaped dynamics is based on the interaction of competition and legitimation forces (Hannan and Freeman, 1989). Legitimation is the process by which certain types of technologies become accepted as more agents adopt them. Competition forces limit the maximum level of diffusion as competition for resources limits the number of agents that an ecosystem can support. Finally, information cascades are another mechanism that may lead to S-shaped diffusion curves. In Banerjee (1992) and Arthur (1989), initially, agents may adopt slowly because they are experimenting with various technological options.

[3] See Geroski (2000) for an insightful survey and Skinner and Staiger (2007) for a review of the historical discussion as well as for some evidence to settle it.

[4] See, for example, the vintage human capital of Chari and Hopenhayn (1991) for a beautiful example.

After some initial precursors have decided to adopt one technology, followers may find optimal to copy their predecessors as in a herd leading to an acceleration of the speed of diffusion.

Because most studies of technology diffusion that use traditional measures focus on one single technology and one or a few countries, traditional measures have not been able to shed light on significant general patterns in technology diffusion. One exception is Cox and Alm (1996) who show that in the US, the time it takes for 25% of potential adopters to adopt a technology (mostly consumer durables) has declined over the 20th century.

2.2.3 The Intensive Margin

Despite its great intuitive appeal, traditional diffusion measures have two important drawbacks. First, their computation requires the use of micro-level data sets which are hard to assemble. The limits imposed by this requirement may explain why, after 50 years of research, we still lack comprehensive data sets that cover the diffusion of many technologies, in many countries over protracted periods. Second, traditional diffusion measures do not capture the intensity with which each adopter uses the technology.[5] For example, a company in the traditional measure will be coded as an adopter both when only one worker uses the technology and when all the workers have access to the technology. Similarly, traditional measures do not reflect how many units of a given technology a worker uses. Indeed, technological change is sometimes directed to increasing the number of technological goods that a worker can use at the same time. These concerns may be significant from a quantitative perspective. Clark (1987) shows that, circa 1910, the intensity of use of spindles and looms accounted for the bulk of cross-country productivity differences in cotton mills.

Since micro-level data sets do not tend to collect information on the intensity of use of technologies, it is difficult to extend traditional diffusion measures to include the intensive margin of adoption. An alternative approach consists in building these measures using country-level data. Comin and Hobijn (2004, 2009a) and Comin et al. (2006, 2008a) constructed the CHAT data set under this premise. CHAT covers the diffusion of 104 technologies (from most sectors of economic activity), for over 150 countries over the last 200 years. The measures of technology in CHAT are ratios for which the numerator reflects the intensity with which producers or consumers employ a technology at a given moment in time and the denominator scales that by the size of the economy (typically measured by the population or by GDP). For example, the diffusion of credit and debit cards is measured by the number of credit and debit card transactions per capita or by the number of points of service per capita, instead of by the share of people that has at least one credit card. Conceptually, a measure such as the number of card transactions per capita can be expressed as the product of two variables: The fraction of people with credit cards, and the average number of transactions of credit card users per user. The first

[5] This is what Mansfield (1968), Davies (1979), and Stoneman (1981) call intra-firm diffusion.

variable captures the extent of diffusion of credit cards, while the second captures the intensity with which they are used once they have diffused.

Because technology is often embodied in capital goods, some of the measures correspond to the number of specific capital goods per capita (e.g. computers and telephones). Other technologies take the form of new production techniques. In these cases, the technology is measured by the output produced with the technique per capita (e.g. tons of steel produced with electric arc furnaces per capita). One can make these measures unit free by taking the logs of the adoption ratios (i.e. log of the number of MRI units per capita).

2.2.3.1 Usage Lags

Measures of adoption that incorporate the intensive margin are hard to compare across technologies because they have different units. This difference in units makes it also difficult to assess the magnitude of the cross-country variation in technology and its comparison with cross-country differences in income. Comin et al. (2008b) transform cross-country differences in adoption intensity to time lags. Time lags have the advantage that they have a common unit across technologies (e.g. years). They define the usage lag of technology x in country c at year t as the answer to the following question: How many years before year t did the United States last have a usage intensity of technology x that country c has in year t?[6]

For example, the amount of kWh of electricity (per capita) produced in Uruguay in 1990 was last observed in the United States in 1949. Thus, the electricity usage lag in Uruguay in 1990 is 41 years. Similarly, the number of personal computers per capita in Spain in 2002 was comparable to that in the United States in 1989. Hence, the 2002 PC usage lag of Spain is 13 years.

Comin et al. (2008b) compute the usage lags of 10 production technologies in periods where they are cutting-edge and for which CHAT covers at least 95 countries. These technologies include electricity production, transportation, communication, IT, and agriculture. In addition, they also compute the time usage lags for per capita GDP. As illustrated by Figure 2.1, most of the world population is living in countries with real GDP per capita levels that have not been observed in the United States in the post World War II era. Moreover, most of Sub-Saharan Africa, as well as Afghanistan and Mongolia, have per-capita income levels that have not been observed in the United States since 1820.

With respect to technology usage lags, their main findings are that (i) Technology usage lags are large, often comparable to lags in real GDP per capita; (ii) usage lags are highly correlated across countries with lags in per-capita income; and (iii) usage lags are highly correlated across technologies. These results are presented in Table 2.10 in the Appendix.

[6] An alternative way to deal with the differences in units is to take logarithms of the technology measures. This is the approach followed by much of the work discussed below.

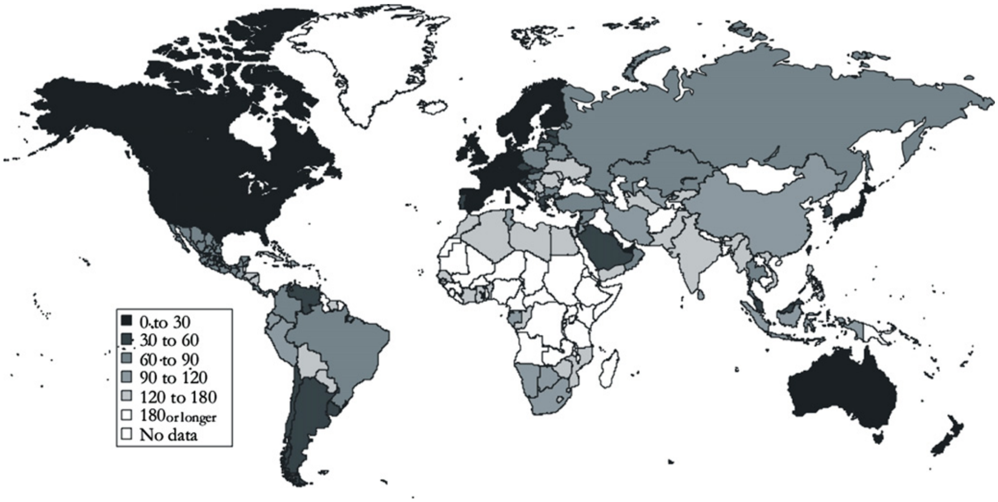


Figure 2.1 Real GDP per capita lags in year 2000.

2.2.3.2 The Shape of Diffusion Curves Once the Intensive Margin is Included

After documenting the magnitude of cross-country differences in technology adoption measures, one natural question is how do the measures of technology that incorporate the intensive margin evolve. In particular, do they follow a logistic curve?

Comin et al. (2008a) study this question using an early version of CHAT with 115 technologies that cover 5678 technology–country pairs.[7] They fit function (2.2) separately to each technology–country pair. For 1291 cases it is not possible to fit the logistic curve due to the lack of curvature in the data since it covers the late stages of diffusion. For 466 cases, the estimate of the speed of diffusion (δ_3) is negative because the technology has become obsolete.[8]

[7] This version of CHAT included some measures of the diffusion of agricultural technologies (typically high-yield seeds) measured as the fraction of agricultural land that used a specific high-yield variety. These series came from Evenson and Gollin (2003).

[8] A negative δ_3 can result either from the substitution by a superior technology or because the logistic is a poor fit. To compute how many of the negative estimates of δ_3 are due to the former, Comin et al. (2008a) recognize that the presence of competing technologies is likely to have similar effects in the estimates of δ_3 across countries. Therefore, in those cases where the negative estimate of δ_3 is produced by the replacement of dominated technologies, we should observe a large number of negative estimates across countries. Comin et al. (2008a) find that 15 out of 115 technologies considered have negative estimates of δ_3 for at least 50% of the technology–country pairs. They identify these as the cases where the estimates of δ_3 are driven by the obsolescence of technology, and therefore are cleared from the count. These technologies include open hearth and Bessemer steel production and the number of sail ships, hospital beds, and checks, all of which have been dominated by another technology.

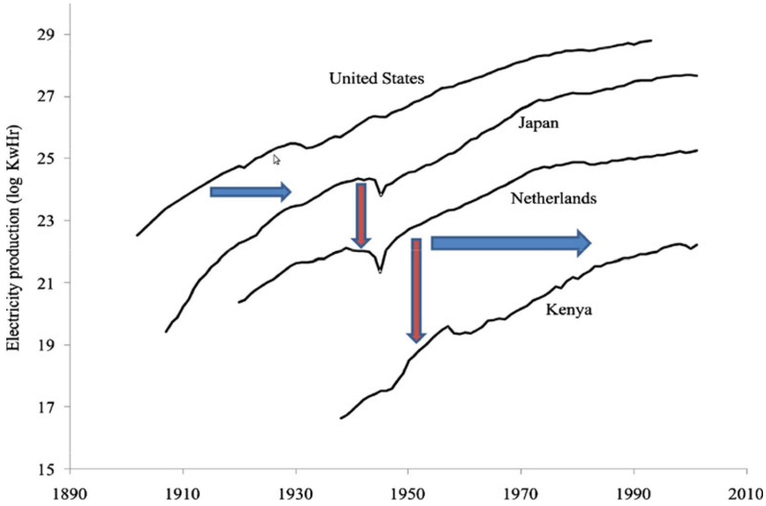


Figure 2.2 Example of diffusion curve.

This leaves 3921 technology–country cases where we can evaluate whether the logistic fits well the evolution of technology measures that include the intensive margin of adoption. For 454 cases, Comin et al. (2008a) still find a negative estimate of δ_3 despite not being a dominated technology. This is, for example, the case of cars per capita in Tanzania, where population grew faster than the number of cars. For 202 cases, the predicted initial adoption is previous to the invention date of the technology. For 336 cases, the predicted adoption date is unrealistically late (either 150 years later than the invention of the technology or 20 years after the first for the country). Finally, 1098 cases correspond to technologies that have a growing ceiling which contradicts the notion that δ_1 is fixed.[9] Adding these up, it turns out that for 53% of the technology–country cases (2084 of 3921), the logistic does not provide a good fit to technology diffusion measures that incorporate the intensity of use.

So, if technology measures do not follow a logistic pattern, what do they follow? Figure 2.2 plots one typical technology measure in CHAT, the production of electricity measured as the log of MWh produced in the US, Japan, Netherlands, and Kenya.

There are a number of features worth noting of these curves. First, they have a concave shape. Second, the shape of these curves is fairly similar. They look as if the same curve, say the one corresponding to the US, had been shifted left and down by different amounts. These two observations motivate us to conjecture that the curvature of the diffusion curve is related to technological characteristics common across countries, while horizontal and

[9] These include: steam and motor ship tonnage; rail passengers-kilometers; railway freight tonnage; tons of blast oxygen furnace, electric-arc furnace, and stainless steel produced; cars; trucks; aviation freight ton-kilometers; TVs; PCs; credit and debit card points of service; ATMs; and checkers.

vertical shifts of the diffusion curves are informative about cross-country differences. One implication of this characterization of diffusion curves is that we just need two parameters to characterize differences across countries in the diffusion of a given technology.

Of course, this raises two questions: How do we interpret these two shifters? And, how can they be identified in the data? Comin and Hobijn (2010) and Comin and Mestieri (2010) explore these two questions.

To start thinking about the shapes of diffusion curves, let $y^c_{\tau,t}$ denote the log-output produced with technology τ at time t in country c. Based on the previous discussion about the shape of diffusion curves, one could conjecture that the diffusion curve could be approximately described by the following expression:

$$y^c_{\tau,t} = \underbrace{\beta^c_{\tau 1}}_{\text{Vtcal Shift}} + \beta_{\tau 2} t + \beta_{\tau 3} \overbrace{\ln(t - \tau - \underbrace{\beta^c_{\tau 4}}_{\text{Hztal Shift}})}^{\text{Concave Shape}} + \varepsilon^c_{\tau t}. \tag{2.3}$$

The left-hand side is the log level of technology. The intercept $\beta^c_{\tau 1}$ captures the vertical shifts in the diffusion curve. We hypothesize a simple concave function such as the log function to introduce curvature in the diffusion curve, as can be seen in the third term of (2.3). The term inside the brackets, $t - \tau$, is the time elapsed since a technology has been invented (we denote a technology τ by its invention date). $\beta^c_{\tau 4}$ is a shifter of the concave curve. The larger $\beta^c_{\tau 4}$ is, the more to the right the diffusion curve shifts. Note that $\ln(t-\tau-\beta^c_{\tau 4})$ is only well defined for $t-\tau-\beta^c_{\tau 4} > 0$. Hence, a higher $\beta^c_{\tau 4}$ captures a delay in the arrival date of the technology τ to country c. Finally, we add a linear time trend that ensures that the technology measure asymptotically behaves log-linearly, as Figure 2.7 suggests.

This statistical characterization of the diffusion curves seems intuitive but it also raises some questions. For example, what role does income play in technology diffusion? A priori, there are two clear roles income can play in the diffusion measures contained in CHAT. First, richer countries should observe larger demand for the goods and services that embody or use technology. Hence, the Engel curve effect should induce a positive effect of income on technology. Second, the costs of producing the goods and services that embody technology tend to increase with the wage rate. Expression (2.3) ignores these effects. To incorporate them properly, it is necessary to develop a model of production and demand for technology. Next, we develop one such model based on Comin and Mestieri (2013). The model provides a microfoundation for a version of (2.3) as well as an interpretation for the vertical and horizontal shifters in Figure 2.2. In particular, it relates the horizontal shifts to the lag with which new vintages of technology (including the first one) on average arrive in a country. The vertical shifters capture the intensity (relative to GDP) with which the technology is used asymptotically.

2.2.3.3 A Microfoundation for the Diffusion Curve

Consider the following economic environment. There is a unit measure of identical households in the economy. Each household supplies inelastically one unit of labor, for which they earn a wage w. Households can save in domestic bonds which are in zero net supply. The utility of the representative household is given by:

$$U = \int_{t_0}^{\infty} e^{-\rho t} \ln(C_t) dt, \tag{2.4}$$

where ρ denotes the discount rate and C, consumption. The representative household maximizes its utility subject to the budget constraint (2.5) and a no-Ponzi scheme condition (2.6):

$$\dot{B}_t + C_t = w_t + r_t B_t, \tag{2.5}$$

$$\lim_{t \to \infty} B_t e^{\int_{t_0}^{t} -r_s ds} \geq 0, \tag{2.6}$$

where B denotes the bond holdings of the representative consumer, $\dot{B}$ is the increase in bond holdings over an instant of time, and r_t its return on bonds.

World technology frontier—At a given instant of time, t, the world technology frontier is characterized by a set of technologies and a set of vintages specific to each technology. To simplify notation, we omit time subscripts, t, whenever possible. Each instant, a new technology, τ, exogenously appears. We denote a technology by the time it was invented. Therefore, the range of invented technologies is $(-\infty, t]$.

For each existing technology, a new, more productive vintage appears in the world frontier every instant. We denote vintages of technology-τ generically by v_τ. Vintages are indexed by the time in which they appear. Thus, the set of existing vintages of technology-τ available at time $t(> \tau)$ is $[\tau, t]$. The productivity of a technology–vintage pair has two components. The first component, $Z(\tau, v_\tau)$, is common across countries and it is purely determined by technological attributes. In particular,

$$\begin{aligned} Z(\tau, v) &= e^{(\chi+\gamma)\tau+\gamma(v_\tau-\tau)} && (2.7)\\ &= e^{\chi\tau+\gamma v_\tau}, && (2.8) \end{aligned}$$

where $(\chi + \gamma)\tau$ is the productivity level associated with the first vintage of technology τ and $\gamma(v_\tau - \tau)$ captures the productivity gains associated with the introduction of new vintages $(v_\tau \geq \tau)$.[10]

The second component is a technology–country specific productivity term, a_τ, which we further discuss below.

[10] In what follows, whenever there is no confusion, we omit the subscript τ from the vintage notation and simply write v.

Adoption lags—Economies typically are below the world technology frontier. Let D_τ denote the age of the best vintage available for production in a country for technology τ. D_τ reflects the time lag between when the best vintage in use was invented and when it was adopted for production in the country; that is, the adoption lag. The set of technology-τ vintages available in this economy is $V_\tau = [\tau, t - D_\tau]$.[11] Note that D_τ is both the time it takes for an economy to start using technology τ and its distance to the technology frontier in technology τ.

Intensive margin—New vintages (τ, v) are incorporated into production through new intermediate goods that embody them. Intermediate goods are produced competitively using one unit of final output to produce one unit of intermediate good.

Intermediate goods are combined with labor to produce the output associated with a given vintage, $Y_{\tau,v}$. In particular, let $X_{\tau,v}$ be the number of units of intermediate good (τ, v) used in production, and $L_{\tau,v}$ be the number of workers that use them to produce services. Then, $Y_{\tau,\,v}$ is given by:

$$Y_{\tau,v} = a_\tau Z(\tau, v) X_{\tau,v}^{\alpha} L_{\tau,v}^{1-\alpha}. \tag{2.9}$$

The term a_τ in (2.9) represents factors that reduce the effectiveness of a technology in a country. These may include differences in the costs of producing the intermediate goods associated with a technology; taxes; relative abundance of complementary inputs or technologies; frictions in capital, labor, and goods markets; barriers to entry for producers that want to develop new uses for the technology, etc.[12] As we shall see below, a_τ determines the long-run penetration rate of the technology in the country. Hence, we refer to a_τ as the intensive margin of adoption of a technology.

Production—The output associated with different vintages of the same technology can be combined to produce competitively sectoral output, $Y_{\tau,}$ as follows:

$$Y_\tau = \left(\int_\tau^{t-D_\tau} Y_{\tau,v}^{\frac{1}{\mu}} \, dv \right)^{\mu}, \quad \text{with } \mu > 1. \tag{2.10}$$

Similarly, final output, Y, results from aggregating competitively the sectoral outputs $\{Y_\tau\}$ as follows:

$$Y = \left(\int_{-\infty}^{\bar{\tau}} Y_\tau^{\frac{1}{\theta}} \, d\tau \right)^{\theta}, \quad \text{with } \theta > 1, \tag{2.11}$$

where $\bar{\tau}$ denotes the most advanced technology adopted in the economy, that is the technology τ for which $\tau = t - D_\tau$.

[11] Here, we are assuming that vintage adoption is sequential. Comin and Hobijn (2010) provide a microfounded model in which this is an equilibrium result rather than an assumption.

[12] Comin and Mestieri (2010) discuss how a wide variety of distortions result in wedges in technology adoption that imply a reduced form as in (2.9).

Factor demands and final output—We take the price of final output as numéraire. The demand for output produced with a particular technology is:

$$Y_\tau = Y p_\tau^{-\frac{\theta}{\theta-1}}, \tag{2.12}$$

where p_τ is the price of sector τ output. Both the income level of a country and the price of a technology affect the demand of output produced with a given technology. Because of the homotheticity of the production function, the income elasticity of technology τ output is one. Similarly, the demand for output produced with a particular technology vintage is:

$$Y_{\tau,v} = Y_\tau \left(\frac{p_{\tau,v}}{p_\tau}\right)^{-\frac{\mu}{\mu-1}}, \tag{2.13}$$

where $p_{\tau,v}$ denotes the price of the (τ, v) intermediate good.[13] The demands for labor and intermediate goods at the vintage level are:

$$(1-\alpha)\frac{p_{\tau,v} Y_{\tau,v}}{L_{\tau,v}} = w \tag{2.14}$$

$$\alpha\frac{p_{\tau,v} Y_{\tau,v}}{X_{\tau,v}} = 1 \tag{2.15}$$

Perfect competition in the production of intermediate goods implies that the price of intermediate goods equals their marginal cost,

$$p_{\tau,v} = \frac{w^{1-\alpha}}{Z(\tau,v)a_\tau}(1-\alpha)^{-(1-\alpha)}\alpha^{-\alpha}. \tag{2.16}$$

Combining (2.13)–(2.15), the total output produced with technology τ can be expressed as:

$$Y_\tau = Z_\tau L_\tau^{1-\alpha} X_\tau^\alpha, \tag{2.17}$$

where L_τ denotes the total labor used in sector τ,

$$L_\tau = \int_\tau^{t-D_\tau} L_{\tau,v}\,dv, \tag{2.18}$$

X_τ is the total amount of intermediate goods in sector τ,

$$X_\tau = \int_\tau^{t-D_\tau} X_{\tau,v}\,dv, \tag{2.19}$$

[13] Even though older technology–vintage pairs are always produced in equilibrium, the value of its production relative to total output is declining over time.

and the productivity associated to a technology is:

$$\begin{aligned} Z_\tau &= \left(\int_\tau^{\max\{t-D_\tau,\tau\}} Z(\tau,\nu)^{\frac{1}{\mu-1}} d\nu \right)^{\mu-1} \\ &= \left(\frac{\mu-1}{\gamma} \right)^{\mu-1} \underbrace{a_\tau}_{\text{Intensive Mg}} \underbrace{e^{(\chi\tau+\gamma\max\{t-D_\tau,\tau\})}}_{\text{Embodiment Effect}} \underbrace{\left(1 - e^{\frac{-\gamma}{\mu-1}(\max\{t-D_\tau,\tau\}-\tau)}\right)^{\mu-1}}_{\text{Variety Effect}}. \end{aligned} \tag{2.20}$$

This expression is quite intuitive. The productivity of a technology, Z_τ, is determined by the intensive margin, the productivity level of the best vintage used (i.e. embodiment effect), and the productivity gains from using more vintages (i.e. variety effect). Adoption lags have two effects on Z_τ. The shorter the adoption lags, D_τ, the more productive are, on average, the vintages used. In addition, because there are productivity gains from using different vintages, the shorter the lags, the more varieties are used in production and the higher Z_τ is.

The price index of technology-τ output is:

$$\begin{aligned} p_\tau &= \left(\int_\tau^{t-D_\tau} p_{\tau,\nu}^{-\frac{1}{\mu-1}} d\nu \right)^{-(\mu-1)} \\ &= \frac{w^{1-\alpha}}{Z_\tau} (1-\alpha)^{-(1-\alpha)} \alpha^{-\alpha}. \end{aligned} \tag{2.21}$$

Diffusion equation—Combining the demand for sector τ output, (2.12), the sectoral price deflator (2.21), the expression for the equilibrium wage rate (2.14), the expression for Z_τ, (2.20) and denoting logs with lower-case letters, we obtain:

$$y_\tau = y + \frac{\theta}{\theta-1} \left[z_\tau - (1-\alpha)(y-l) \right]. \tag{2.22}$$

From expression (2.20) we see that, to a first-order approximation, γ only affects y_τ through the linear trend. This allows us to do a second-order approximation of log Z_τ around the starting adoption date as:

$$z_\tau \approx \ln a_\tau + (\chi+\gamma)\tau + (\mu-1)\ln(t-\tau-D_\tau) + \frac{\gamma}{2}(t-\tau-D_\tau). \tag{2.23}$$

Substituting (2.23) in (2.22) gives us the following estimating equation[14]:

$$y^c_{\tau t} = \beta^c_{\tau 1} + y^c_t + \beta_{\tau 2} t + \beta_{\tau 3} \left((\mu-1)\ln(t - D^c_\tau - \tau) - (1-\alpha)(y^c_t - l^c_t) \right) + \varepsilon^c_{\tau t}, \tag{2.25}$$

[14] When bringing the model to the data, we shall see that some of the technology measures we have in our data set correspond to the output produced with a specific technology, and therefore Equation (2.25) is the appropriate model counterpart. Other technology measures, instead, capture the number of units of the input that embody the technology (e.g. number of computers). The model counterpart to those measures is X_τ. To derive an estimating equation for these measures, we integrate (2.15) across vintages

where $y^c_{\tau t}$ denotes the log of the output produced with technology τ, y^c_t is the log of output, $y^c_t - l^c_t$ is the log of output per capita, $\varepsilon^c_{\tau t}$ is an error term, and the country-technology specific intercept, β^c_1, is equal to:

$$\beta^c_{\tau 1} = \beta_{\tau 3}\left(\ln a^c_\tau + \left(\chi + \frac{\gamma}{2}\right)\tau - \frac{\gamma}{2}D^c_\tau\right). \tag{2.26}$$

Equation (2.25) shows that the adoption lag D^c_τ is the only determinant of shifts in the curvature of the diffusion curve. Intuitively, longer lags imply that fewer vintages are available for production and, because of the diminishing gains from variety, the steepness of the diffusion curve declines faster than if more vintages had been already adopted. Equation (2.26) shows that, for a given adoption lag, the only driver of cross-country differences in the intercept $\beta^c_{\tau 1}$ is the intensive margin, a^c_τ. A lower level of a^c_τ generates a downward shift of the diffusion curve which, *ceteris paribus*, leads to lower output associated with technology τ throughout its diffusion and, in particular in the long run.[15]

Formally, we can identify differences in the intensive margin relative to a benchmark, which we take to be the average value for 17 Western countries (defined by Maddison, 2004)[16] as:

$$\ln a^c_\tau = \frac{\beta^c_{1,\tau} - \beta^{Western}_{1,\tau}}{\beta_{3,\tau}} + \frac{\gamma}{2}(D^c_\tau - D^{Western}_\tau). \tag{2.27}$$

2.2.3.4 The Intensive and Extensive Margin

Estimation—Comin and Hobijn (2010) and Comin and Mestieri (2010, 2013) develop a two step procedure to estimate (2.24) and (2.25). First, they estimate the equation jointly for a few countries for which the data series are longest and the data quality is highest. Here, we follow Comin and Mestieri (2013) and use the US, the UK, and France. Then, imposing the estimates of $\hat{\beta}_{2\tau}$ and $\hat{\beta}_{3\tau}$, which are in principle common across countries, they re-estimate the equation to obtain the country-technology estimates of D^c_τ and a^c_τ.

We focus on a subsample of 25 technologies that have a wider coverage over rich and poor countries and for which the data captures the initial phases of diffusion (see

to obtain (in logs) $x^c_\tau = y^c_\tau + p^c_\tau + \ln(\alpha)$. Substituting in for Equation (2.25), we obtain an analogous expression to the one used in the main text:

$$x^c_{\tau t} = \beta^c_{\tau 1} + y^c_t + \beta_{\tau 2}t + \beta_{\tau 3}\left((\mu - 1)\ln(t - D^c_\tau - \tau) - (1-\alpha)(y^c_t - l^c_t)\right) + \varepsilon^c_{\tau t}. \tag{2.24}$$

[15] The intuition for why using a second-order approximation of productivity growth suffices is that identification of adoption lags comes through the initial stages of diffusion, where the diffusion curve has more curvature than a log-linear trend (as when it becomes log-linear, it is impossible to separately identify it from embodied productivity growth). Hence, the approximation of the diffusion curve around the initial stages.

[16] These countries are Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Italy, Japan, Netherlands, New Zealand, Norway, Sweden, Switzerland, United Kingdom, and the United States of America.

Appendix A). These technologies cover a wide range of sectors in the economy (transportation; communication and IT; industrial; agricultural; and medical sectors). Their invention dates also span quite evenly over the last 200 years.

As in Comin and Hobijn (2010), we use the plausibility and precision of the estimates of the adoption lags from Equation (2.25) as a pre-requisite to utilize the technology-country pair in our analysis. We find that these two conditions are met for the majority of the technology–country pairs (67%).[17] For these technology country-pairs, we find that Equation (2.25) provides a very good fit for the data with an average detrended R^2 of 0.79 across countries and technologies (Table 2.11).[18]

Statistics—Tables 2.3 and 2.4 report summary statistics for the estimates of the adoption lags and the intensive margin for each technology. The average adoption lag across all technologies (and countries) is 44 years. We find significant variation in average adoption lags across technologies. The range goes from 7 years for the Internet to 121 years for steam and motor ships. There is also considerable cross-country variation in adoption lags for any given technology. The range for the cross-country standard deviations goes from 3 years for PCs to 53 years for steam and motor ships.

We also find significant cross-country variation in the intensive margin. The intensive margin is reported as log differences relative to the average adoption of Western countries.[19] The average intensive margin is -0.62, which implies that the level of adoption of the average country is 54% of the Western countries. More generally, there is significant cross-country dispersion in the intensive margin. The range goes from 0.3 for mail to 1.1 for cars and the Internet. These summary statistics for the estimates of adoption lags and the intensive margin of adoption are consistent with those in Comin and Hobijn (2010) and Comin and Mestieri (2010) which use smaller technology samples and estimate other versions of the diffusion Equation (2.25).

Evolution—The long time spans and cross-country coverage of the technologies in CHAT allow us to explore the presence of cross-country trends in adoption patterns. Comin and Hobijn (2010) explored whether there has been any trend in adoption lags

[17] Plausible adoption lags are those with an estimated adoption date of no less than 10 years before the invention date (this is to allow for some inference error). Precise are those with a significant estimate of adoption lags and the intercept $\beta_{1\tau}^{c}$ at a 5% level. Following Comin and Hobijn (2010), we relax this condition and include in the "precise" category those estimates that have a standard error of adoption lags smaller than $\sqrt{2003 - \text{invention date}}$. The idea is to allow for some older technologies to be more imprecisely estimated. However, this additional margin hardly expands the set of precise estimates. Only 15 additional estimates are included with this condition, which represent 1.2% of our precise observations. Most of the implausible estimates correspond to diffusion curves that do not have the initial phases of diffusion. This makes it very hard to separately identify the log-linear trend from the log component of (2.25).

[18] To compute the detrended R^2, we partial out the linear trend γt and compute the R^2 of the detrended data.

[19] To compute the intensive margin we follow Comin and Mestieri (2013) and calibrate $\gamma = (1 - \alpha) \cdot 1\%, \alpha = 0.3$, and use a value of $\beta_{3,\tau}$ that results from setting the elasticity across technologies θ to be the mean across our estimates, which is $\theta = 1.28$.

Table 2.3 Estimated adoption lags

	Invention Year	Obs.	Mean	SD	P10	P50	P90	IQR
Spindles	1779	31	119	48	51	111	171	89
Steam and Motor Ships	1788	45	121	53	50	128	180	104
Railways-Freight	1825	46	74	34	31	74	123	50
Railways-Passengers	1825	39	72	39	16	70	123	63
Telegraph	1835	43	45	32	10	40	93	43
Mail	1840	47	46	37	8	38	108	62
Steel (Bessemer, Open Hearth)	1855	41	64	34	14	67	105	51
Telephone	1876	55	50	31	8	51	88	51
Electricity	1882	82	48	23	15	53	71	38
Cars	1885	70	39	22	11	34	65	36
Trucks	1885	62	36	22	9	34	62	32
Tractor	1892	88	59	20	18	67	69	12
Aviation-Freight	1903	43	40	15	26	42	60	19
Aviation-Passengers	1903	44	28	16	9	25	52	18
Electric Arc Furnace	1907	53	50	19	27	55	71	34
Fertilizer	1910	89	46	10	35	48	54	7
Harvester	1912	70	38	18	10	41	54	17
Synthetic Fiber	1924	48	38	5	33	39	41	2
Blast Oxygen Furnace	1950	39	14	8	7	13	26	11
Kidney Transplant	1954	24	13	7	3	13	25	5
Liver Transplant	1963	21	18	6	14	18	24	3
Heart Surgery	1968	18	12	6	8	13	20	4
Cellphones	1973	82	13	5	9	14	17	6
PCs	1973	68	16	3	12	15	19	3
Internet	1983	58	7	4	1	7	11	3
All Technologies		1306	44	35	9	38	86	46

over the last 200 years. They find that the average lag with which countries adopt technologies has dropped with the invention date of technologies. In particular, they find that technologies invented 10 years later, on average, have been adopted 4 years earlier (relative to the invention date).

The first column of Table 2.5 extends this finding to our 25 technologies. More specifically, it reports the estimates of regressing the (log) adoption lags on the invention date (minus 1820) and a constant. The first column reports the results from this regression for the whole sample. The constant term shows the average (log) adoption level in 1820. The negative coefficient in the invention date illustrates the finding in Comin and Hobijn (2010) that new technologies have diffused on average faster.

Table 2.4 Estimated intensive margin

	Invention Year	Obs.	Mean	SD	P10	P50	P90	IQR
Spindles	1779	31	−0.02	0.6	−0.8	−0.1	0.8	0.7
Steam and Motor Ships	1788	45	−0.01	0.6	−0.6	0.0	0.7	0.6
Railways-Freight	1825	46	−0.17	0.4	−0.6	−0.2	0.4	0.6
Railways-Passengers	1825	39	−0.24	0.5	−0.9	−0.2	0.2	0.5
Telegraph	1835	43	−0.26	0.5	−1.0	−0.2	0.3	0.7
Mail	1840	47	−0.19	0.3	−0.6	−0.1	0.1	0.4
Steel (Bessemer, Open Hearth)	1855	41	−0.22	0.4	−0.7	−0.1	0.2	0.6
Telephone	1876	55	−0.91	0.9	−2.2	−0.8	0.1	1.2
Electricity	1882	82	−0.58	0.6	−1.2	−0.5	0.1	0.9
Cars	1885	70	−1.13	1.1	−2.1	−1.1	0.1	1.6
Trucks	1885	62	−0.86	1.0	−1.7	−0.8	0.1	1.1
Tractor	1892	88	−1.02	0.9	−2.3	−0.9	0.1	1.5
Aviation-Freight	1903	43	−0.39	0.6	−1.3	−0.2	0.2	0.9
Aviation-Passengers	1903	44	−0.45	0.7	−1.3	−0.4	0.2	0.9
Electric Arc Furnace	1907	53	−0.29	0.5	−0.9	−0.2	0.3	0.8
Fertilizer	1910	89	−0.83	0.8	−1.9	−0.7	0.1	1.3
Harvester	1912	70	−1.10	1.0	−2.7	−1.0	0.2	1.5
Synthetic Fiber	1924	48	−0.52	0.7	−1.6	−0.4	0.2	0.9
Blast Oxygen Furnace	1950	39	−0.81	0.9	−2.3	−0.4	0.1	1.8
Kidney Transplant	1954	24	−0.19	0.4	−0.8	−0.1	0.1	0.3
Liver Transplant	1963	21	−0.33	0.7	−1.6	−0.1	0.1	0.5
Heart Surgery	1968	18	−0.44	0.8	−1.7	−0.1	0.2	0.6
Cellphones	1973	82	−0.75	0.7	−1.8	−0.6	0.1	1.2
PCs	1973	68	−0.60	0.6	−1.4	−0.6	0.1	0.9
Internet	1983	58	−0.96	1.1	−2.1	−0.8	0.1	1.5
All Technologies		1306	−0.62	0.8	−1.7	−0.4	0.2	1.0

Comin and Mestieri (2013) go one step further and ask whether the trend in adoption lags is uniform across countries. In particular, has it been the same for Western leaders and for non-Western followers? Column 2 of Table 2.5 reports the regression for Western countries and column 3 for non-Western countries. In 1820, adoption lags were significantly shorter in Western countries than in non-Western countries. However, the rate of decline of adoption lags has been significantly larger in non-Western countries than in Western countries (1.12% vs. 0.81%). Therefore, adoption lags have converged across countries.[20]

[20] Comin and Mestieri (2013) show that this finding extends to considering alternative country groupings such as bottom 10% and 20% of countries according to their income.

Table 2.5 Evolution of the adoption lag

Dependent variable is:	(1) Log(Lag) World	(2) Log(Lag) Western countries	(3) Log(Lag) Rest of the world
Year-1820	−0.0106*	−0.0081*	−0.0112*
	(0.0004)	(0.0006)	(0.0004)
Constant	4.27*	3.67*	4.48*
	(0.06)	(0.07)	(0.05)
Observations	1274	336	938
R-squared	0.45	0.34	0.53

Note: robust standard errors in parentheses.
Each observation is re-weighted so that each technology carries equal weight.
*Denotes 1% significance.

We conduct a similar exercise for the intensive margin of adoption. Given that the intensive margin is defined relative to a benchmark, the evolution of the average intensive margin is not very meaningful, but we can still ask the question of whether there has been convergence in the intensive margin across countries. Comin and Mestieri (2013) explore this question by regressing the intensive margin on the invention date of the technology (minus 1820) and a constant. Table 2.6 reports their main finding. As shown in column 3, the intensive margin in non-Western countries (relative to the Western average) has declined (with the invention date) at a rate of 0.54% per year. This estimate implies that the gap in the intensity of technology adoption between rich and poor countries is larger for newer than for old technologies. So, there has been divergence in the intensive margin

Table 2.6 Evolution of the intensive margin

Dependent variable is:	(1) Intensive World	(2) Intensive Western countries	(3) Intensive Rest of the world
Year-1820	−0.0029*	0.0000	−0.0054*
	(0.0005)	(0.0002)	(0.0005)
Constant	−0.32*	−0.00	−0.39*
	(0.05)	(0.06)	(0.07)
Observations	1306	350	956
R-squared	0.042	0	0.13

Note: robust standard errors in parentheses.
Each observation is re-weighted so that each technology carries equal weight.
*Denotes 1% significance.

over the last 200 years. In Section (2.4.3), we review the implications that this has had on the cross-country dynamics of income.

Robustness checks—One important identification assumption is that the curvature of the diffusion curve (2.25), $\beta_{3\tau}$, is common across countries for a given technology τ. Comin and Hobijn (2010) evaluate this hypothesis by allowing it to vary by technology-country pair and then testing the null that the common and the country-specific estimate of $\beta_{3\tau}$ are the same. Reassuringly, they find that they cannot reject the null that both estimates are the same for 69% of the technology–country pairs.

A second restriction used in the estimation—this one imposed by the model—is that the elasticity of technology with respect to income is one. The homotheticity of technology may be a restrictive constraint in reality. To evaluate the robustness of the findings to alternative formulations of the demand for technology, Comin and Mestieri (2013) propose a method to estimate the income elasticity of technology. Specifically, they estimate the income elasticity of technology in the first stage (along with β_2 and β_3) for the three baseline countries (US, UK, and France). Effectively, this implies that the income elasticity of technology is identified from the time-series variation of technology and income for these countries. Since the time span of the diffusion for most technologies in these countries is quite long, it covers periods when their income was far lower than today. Hence, this estimate seems a reasonable proxy for the income elasticity of technology in developing countries too. They find that both the estimates and the trends in adoption described above are robust to allowing for non-homotheticities in demand.

2.2.4 Other Approaches

We conclude our discussion of the measurement of technology by mentioning one recent approach proposed by Alexopoulos (2011). Her approach consists of measuring technology by the number of books published in the field of a particular technology. The rationale of this measure is that technology books are published when new discoveries (relevant for the industry) are made. One advantage of this measure is that, because the topics covered by books are classified into narrow fields, it is possible to collect time-series measures for relatively disaggregated fields.

One important question is whether these measures capture innovations or diffusion of the innovations. To explore this issue, Alexopoulos shows that the number of new books on a given technological field peaks in the early stages of diffusion of a new technology, and leads other measures of diffusion of the technology. She argues, based on this evidence, that books measure innovation rather than diffusion. However, Alexopoulos also shows that both R&D expenditures and patent applications lead the number of science books published. This would suggest that the number of technology books published in a discipline does not reflect innovation but measure technology some time after the innovation has taken place. One plausible hypothesis is that the number of books published reflects the expected value of the technology at the early stages of diffusion, which is when it may be optimal to publish a book.

2.3. DRIVERS OF TECHNOLOGY ADOPTION

After showing the magnitude of the existing cross-country differences in technology, one can only wonder about what factors explain the large cross-country differences in technology. At this point, it may be safe to conjecture that there may be a large number of factors that drive cross-country differences in technology. Many of them may still be unexplored, while we are just beginning to have direct evidence of the relevance of a few others.

As before, in this section, we will tend to focus on studies that have explored cross-country differences in technology as opposed to within country differences. In part, because it is not clear that the drivers of adoption within country are the same as those across countries. However, when relevant, we describe within-country evidence. As in Section 2.2, we also prioritize studies that consider direct measures of multiple technologies because of our interest in uncovering general patterns in the data.

We organize our exposition by classifying the drivers into three broad categories. The first two (knowledge, and institutions/policies) affect technology from the supply side, while the third (aggregate demand) represents the pull forces of technology.[21]

2.3.1 Knowledge

New technology brings new production processes, machines, products, and services which typically are not straightforward to implement (Comin and Hobijn, 2007). A significant part of the cost of adopting new technologies is the cost of figuring out what technology is needed to produce the desired good or service and how to use it individually or as part of an existing production process. Therefore, any prior knowledge that reduces the magnitude of these costs should foster technology adoption.

Knowledge may take a variety of forms depending on who has it, and its nature. Nelson and Phelps (1966) focused on human capital; that is, formal knowledge embodied in people.[22] Human capital has typically been measured as the fraction of population that has attained a certain schooling level or as the fraction of population in schooling age that is enrolled in certain schooling level. Formal schooling may not be the only (or even the most relevant) source of knowledge for the adoption of new technologies since workers may learn on the job.[23]

[21] As we show below, both profitability and spread of information—the traditional drivers of adoption for the economics and marketing literatures—are comprised in these categories.

[22] See Benhabib and Spiegel (2005) for a more comprehensive survey of work exploring this hypothesis.

[23] See, for example, Seshadri and Manuelli (2005), Erosa et al. (2010), and the references therein.

In addition to knowledge embodied in people, knowledge may be collectively embodied in organizations or in sectors. The concept of organizational knowledge captures the notion that there may be complementarities between the knowledge of workers which increase the organization capacity to adopt new technologies beyond the sum of the workers' individual capacities. Finally, a company's capabilities to adopt or use a new technology may be positively affected by the capabilities of other agents. These may be similar companies in the same geography (clusters), e.g. Porter (1998), or distinct organizations with which it interacts directly or indirectly. For example, a company may seek technological advice from public organizations that have prior experience in the technology (e.g. Fraunhofer in Germany, Comin et al. 2012). Finally, a company's adoption potential may be affected by the technological experience of companies in other geographies with which it has some contact. This implication would follow from a simple extension of epidemic diffusion models (to allow for multiple geographies).

Next, we review some evidence about the role of the different sources of knowledge on technology adoption.

2.3.1.1 Human Capital

Caselli and Coleman (2001) explore the role of human capital in the diffusion of computers. Using data on the value of computer imports for 90 countries between 1970 and 1990, they study whether imports are affected by various measures of human capital. In their specification, they control for per-capita income, year-dummies, continent dummies, and a country-level random effect. They find that an increase by 1 percentage point in the fraction of the population with more than primary schooling is associated with an increase in the value of computers imported by 1%.

Riddell and Song (2012) use Canadian micro-level data from the Workplace and Employee Survey to explore the same question. More specifically, these authors use time and state variation in compulsory education laws to instrument the education attainment of workers. Their main findings are that graduating from high-school increases the probability of using a computer in the job by 37 percentage points. Similarly, an additional year of schooling increases this probability by 7 percentage points. In contrast, they do not find any significant effect of education on the probability that a worker uses computer-controlled machines. There are a few remarks worth making about the findings in Riddell and Song (2012). First, the fact that a worker's own human capital does not affect his probability of adopting numerically controlled machines does not imply that human capital is irrelevant for the diffusion of this technology. It may well be the case that the human capital of other relevant agents is important (technicians, managers, importers,...). A second remark made by Riddell and Song concerns the significantly higher estimates (almost three times) for the effect of human capital on computer adoption when instrumenting education than with OLS. This result suggests that, with the instrumentation, the authors are probably capturing the local average treatment effect

(LATE) rather than the average treatment effect (ATE) which is the relevant measure for the question posed.

One would like to explore whether the importance of human capital for technology adoption extends beyond computers. Benhabib and Spiegel (1994) find evidence that the stock of human capital affects the growth rate of productivity (i.e. TFP) which they interpret in the light of the Nelson and Phelps (1966) model. Comin and Hobijn (2004) look at the predecessor of CHAT (the HCCTAD) which contains information on the diffusion of 25 major technologies in 15 advanced countries over the last 200 years.

The specification used by Comin and Hobijn (2004) is similar to the one used by Caselli and Coleman (2001). In particular, they consider the following regression:

$$y_{jt}^{c} = \eta_{jt} + \beta X_{jt}^{c} + \epsilon_{jt}^{c}, \tag{2.28}$$

where y_{jt}^{c} denotes the adoption level of technology j in country c in year t, η_{jt} is a full set of technology-time dummies, and X_{jt}^{c} is a matrix of (possibly technology-specific) controls. In particular X_{jt}^{c} always include the log of GDP per capita and may include controls for the openness of the country, quality of political institutions, measures of adoption of general technologies (i.e. electricity) and of predecessor technologies. The regression results are reported in Table 2.7.

Table 2.7 Technology pooled regressions

	Dependent variable is: Technology_{cjt}								
ln(RGDPpc)	1.15 (0.03)*	1.12 (0.03)*	0.57 0 (0.07)*	1.10 (0.03)*	1.05 (0.04)*	0.93 (0.05)*	1.04 (0.35)*	1.04 (0.03)*	1.2 (0.09)*
Prim.enr. 70−		0.09 (0.06)		0.06 (0.07)	0.08 (0.07)	1.23 (0.18)*	0.10 (0.07)	0.09 (0.07)	1.69 (0.26)*
Prim.enr. 70+		0.35 (0.21)		0.39 (0.23)	0.22 (0.23)	−0.11 (0.2)			−0.48 (0.4)
Sec.enr. 70−		0.30 (0.08)*		0.36 (0.08)*	0.31 (0.08)*	0.37 (0.09)*	0.27 (0.08)*	0.3 (0.08)*	0.22 (0.12)
Sec.enr 70+		0.08 (0.128)		0.05 (0.15)	−0.01 (0.15)	0.13 (0.27)			−0.36 (0.36)
Prim.Att.			0.01						
Prim.Att.			(0.00)*						
Sec.Att			0.01 (0.00)*						
Tert.Att			0.01 (0.00)*						
Openness				0.06 (0.02)*	0.06 (0.02)*	0.24 (0.11)	0.07 (0.02)*	0.31 (0.09)*	0.35 (0.15)*

(*Continued*)

Table 2.7 Continued

Dependent variable is: Technology$_{cjt}$

TwtGDP							−0.22		
							(0.06)*		
Open. · TwtGDP								−0.15	
								(0.05)*	
Ex.mon.					0.16		0.13	0.14	
					(0.07)		(0.07)	(0.07)	
Ex.prem.					−0.11	0.06	−0.14	−0.12	−0.05
					(0.04)	(0.06)	(0.04)*	(0.04)*	(0.08)
Ex.Other					−0.33	−0.17	−0.36	−0.33	−0.53
					(0.06)*	(0.08)	(0.06)*	(0.06)*	(0.11)*
Mil.Reg.					−0.42	−0.46	−0.45	−0.43	−1.17
					(0.08)*	(0.15)*	(0.08)*	(0.08)*	(0.19)
Legislat. Eff.						−0.16			−0.31
						(0.05)*			(0.07)*
Party						0.08			0.75
						(0.04)			(0.05)
ln(MWHR)									0.06
									(0.04)*
Prev.tech									0.16
									(0.03)*
No. of obs.	5488	5417	2341	4986	4986	2118	5057	5057	1000
R^2 (within)	0.24	0.24	0.17	0.23	0.25	0.33	0.25	0.24	0.48

Notes: Standard errors in parentheses,
The technology measures included from CHAT are: Fraction of spindles that are ring spindles, Fraction of tonnage of steel produced using Bessemer method, Fraction of tonnage of steel produced using Open Hearth furnaces, Fraction of tonnage of steel produced using Blast Oxygen furnaces, Fraction of tonnage of steel produced using Electric Arc furnaces, Mail per capita, Telegrams per capita, Telephones per capita, Mobile phones per capita, Newspapers per capita, Radios per capita, Televisions per capita, Personal computers per capita, Industrial robots per unit of real GDP, Freight traffic on railways (TKMs) per unit of real GDP, Passenger traffic on railways (PKMs) per capita, Trucks per unit of GDP, Passenger cars per capita, Aviation cargo (TKMs) per unit of real GDP, Aviation passengers (PKMs) per capita, Transportation (shipping), Fraction of merchant fleet (tonnage) made up of steamships and motorships, MWhr of electricity produced per unit of real GDP. TwtGDP and Previous technology have been instrumented for 5-year lagged values.
* Denotes significance at 1% level.

Because of data constraints, Comin and Hobijn (2004) allow for different effects of enrollment rates before and after 1970. The most robust result they find concerning human capital is that, until 1970, secondary enrollment is positively associated with technology adoption. This effect does not diminish after including all these controls with the exception of electricity production and the predecessor technologies which reduces significantly the sample (from over 5000 to 1000 observations) and reduces the regression coefficient by a fourth (from 0.3 to 0.22). After 1970, however, they find no significant effect of secondary enrollment on technology adoption. Attainment rates (in all schooling level) are also positively associated with technology adoption.

Comin and Hobijn (2004) also explore the association between education and adoption of specific technologies. Consistent with Riddell and Song (2012), they find heterogeneity in the coefficients. The positive impact of secondary schooling on adoption is driven by mass communication technologies (newspapers, radio, and TV) and by electricity. For the other technologies the association with secondary enrollment is insignificant. For transportation technologies they find a positive association between primary enrollment and technology diffusion; and a negative one for steel production technologies. Finally, and consistent with the previous evidence, they find a positive and significant association between the rate of tertiary attainment and the adoption of computers.

2.3.1.2 Adoption History

Vintage capital models, either based on human or physical capital (e.g. Johansen, 1959; Solow, 1960; Chari and Hopenhayn, 1991; Caselli, 1999), predict some form of leapfrogging because of the difficulty to transfer technology-specific human or physical capital from old to new technologies. Comin and Hobijn (2004) test this prediction by matching technologies in HCCTAD to their predecessor technologies. In particular, they use information on the diffusion of 11 technologies for which they have information for both new and predecessor technologies. Contrary to the vintage capital models, they find that there is a positive association between the adoption of predecessor and new technologies. This effect is robust to controlling for variables that affect the overall return to adopting new technologies in the country such as income, education, trade openness, and the institutional environment.[24]

This finding suggests that there are inputs in the adoption process that are transferable across technologies within a sector. These inputs are not formal human capital since this is one of the controls. They do not capture institutional quality, openness, or other variables that are likely to have a symmetric effect across technologies since income is also in the set of controls. What can they capture then?

Comin et al.'s (2010) investigation shed some light on this question. Combining the data set on country-level measures of historical technology adoption described in Section 2.2.1 and measures of adoption for current times from CHAT, Comin et al. (2010) explore the effect of historical adoption on current adoption. This exercise is distinct from Comin and Hobijn (2004) in at least two respects. First, it covers all countries, not just 15 rich countries. Second, the periods they considered are 1000 BC, 0 AD, 1500 AD, and 2000 AD. Therefore, the horizons over which they estimate the persistence of technology adoption are much longer than in Comin and Hobijn (2004).

Figure 2.3 presents one of the key findings. The overall technology adoption level in 1500 AD is positively and significantly associated with current income per capita. This R^2

[24] For steel production technologies, they find a negative partial association between the adoption of new and predecessor technologies.

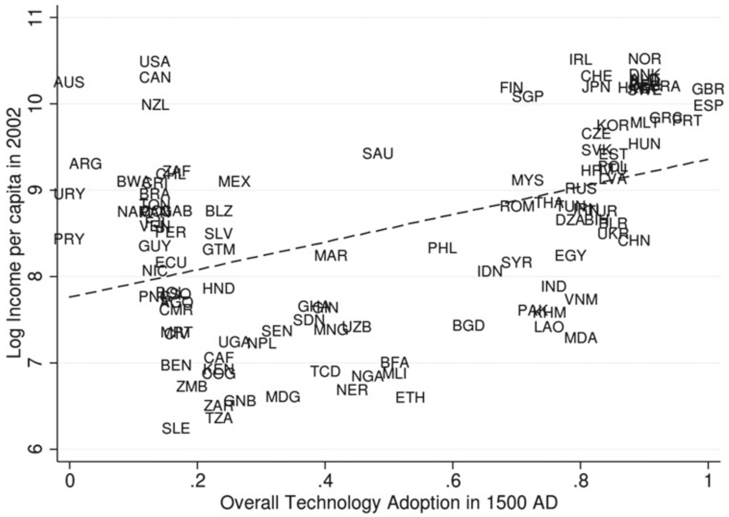


Figure 2.3 Technology in 1500 AD and current development.

indicates that this measure of technology in 1500 AD accounts for 18% of the variation in log-per capita GDP in 2002. Changing from the maximum (i.e. 1) to the minimum (i.e. 0) the overall technology adoption level in 1500 AD is associated with a reduction in the level of income per capita in 2002 by a factor of 5. The authors also find a similar association between past and current technologies. The association is robust to including continent dummies, and controlling for geographical variables.

This persistence of technology adoption may well be the result of some persistent factor that affects (contemporaneously) technology adoption. The literature has suggested a few, such as genetic endowment (Ashraf and Galor, 2013; Spolaore and Wacziarg, 2009), culture (Tabellini, 2007; Guiso et al. 2008), and institutions (Acemoglu et al. 2002; Bockstette et al. 2002). Comin et al. (2010) note that these factors are likely to have a symmetric effect on technology adoption across sectors. In contrast, sector-specific knowledge is likely to have a larger effect on subsequent adoption within a given sector than in other sectors. This variation provides a natural identification strategy for the source of the persistence in technology adoption. In particular, one could compare the persistence in technology adoption at the sector level before and after including country time-varying effects. If the inclusion of country effects (which is equivalent to looking at deviations in adoption from the country mean in each period) does not affect dramatically the estimated persistence, we can conclude that it is unlikely to result from country-wide factors that affect symmetrically technology adoption across sectors.

Table 2.8 Persistence of technology within countries

	Dependent Variable: Technology$_{cst}$					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Technology$_{cst-1}$	0.4* (4.85)	0.29* (4.84)	0.25* (3.71)	0.25* (3.11)	0.39* (4.54)	0.28* (4.12)
Country-time fixed effects	NO	YES	YES	YES	YES	YES
Sectors excluded other than military	–	–	Agri.	Comm	Transp.	Indust.
N	417	417	315	307	312	317
R^2	–	0.48	0.43	0.56	0.32	0.56

Notes: t-statistics in parentheses.
Panel regressions using (migration-adjusted) technology level a in sector s, country c in year $t, a_{cst} = \beta_{cs} + \beta_{ct} + \beta_s \cdot t + \gamma \cdot a_{cst-1} + \varepsilon_{cst}$, where β_{cs} is a country-sector fixed effect, β_{ct} is a country-time fixed effect, β_s is a technology-specific trend, and ε_{cst} is an error term. Regressions are estimated in first differences instrumenting $a_{cst} - a_{cst-1}$ with a_{cst-2} to avoid lagged dependent variable bias, where $t = 2000,\ t-1 = 1500$ AD, $t-2 = 0$.
* Denotes significance at 1%.

Table 2.8 presents the results from this exercise.[25] In it, we can see that including country-level effects does not change much the persistence of technology at the sector level. The point estimate declines from 0.4 to 0.29 and both are significant at the 1% level (see columns 1 and 2). Furthermore, the results are not driven by the persistence of adoption in any single sector, as results in columns (3)–(6) show. These results suggest that the most likely driver of persistence in technology is the learning of sector-specific technological knowledge, which follows from adopting and using new technologies. Comin et al. (2010) further note that these dynamics are very pervasive and, based on their findings, can lead to large income differences across countries.

2.3.1.3 Geographic Interactions

Most empirical studies on technology adoption have treated adoption units as independent. Consequently, they have tried to link a country's technology adoption patterns to the country's characteristics (e.g. human capital, institutions, policies, adoption history, etc.). This empirical approach to the drivers of technology adoption ignores the possibility of cross-country interactions in the adoption process. This assumption might be restrictive. Adopting a technology requires acquiring knowledge which often comes from interactions with other agents. The frequency and success of these interactions is likely

[25] The specification used in Comin et al. (2010) allows for country-sector fixed effects in the level of adoption. They instrument the change in adoption with the lag level of adoption following Arellano and Bond (1991).

to be shaped by geography. Technological knowledge is likely to be more easily transmitted between agents in countries that are close than between agents located far apart. Similarly, the payoff to adopting a given technology (e.g. railways) may be affected by the adoption experience of neighboring countries. As argued in Diamond (1999), some technologies may be geography-specific. All these mechanisms may generate correlated adoption patterns across nearby countries.

In the development literature, several studies have explored how the neighbors' adoption decision affects an agent's own decision. Foster and Rosenzweig (1995) study the adoption of high-yield varieties in Indian villages. They find that the profitability of this technology was increasing and concave in the neighbors' experience with the seeds. Bandiera and Rasul (2006) study the diffusion of sunflowers in Mozambique finding positive effects of neighbors' adoption decisions on a farmer's adoption when few neighbors have adopted, but negative when a significant number of neighbors have. Conley and Udry (2010) study the fertilizer behavior of pineapple farmers in Ghana. They observe that a farmer will tend to imitate neighbors' fertilizer behavior when the neighbor has been successful in the past. This effect is stronger when the farmer has little experience of his own.[26]

Despite its importance, it is still difficult to ascertain the generality of the findings from existing micro studies. In particular, are informational frictions and social interactions relevant for other technologies (e.g. in other sectors, more complex, or more capital intensive)? And how relevant are informational frictions and interactions once the focus moves from explaining adoption differences among individuals to cross-country differences?

To explore the empirical importance of these mechanisms, Comin et al. (2013) (CDR, henceforth) measure how far a country is from the high-density points in the distribution of technology adoption in the other countries. They denote this measure of the spatial distance from other country's technology SDT. A negative correlation between SDT and adoption, after controlling for country and time fixed effects, implies that countries that are further away from those where the technology diffuses faster tend to experience a slower adoption of the technology.

Using data on 20 technologies from CHAT, they find a strong and significant negative partial correlation between SDT and a country's adoption. The estimates imply that spatial interactions that facilitate technology adoption decline by 73% every 1000 Kms. The estimates are robust to controlling for income, human capital, trade openness, institutions and for the spatial distance from other countries per-capita income (SDI), constructed in a way parallel to SDT.

[26] Similar neighbor effects have been observed in bed nets Dupas (2009). See Foster and Rosenzweig (2010) for a survey of the development literature.

To further explore the nature of the interactions that is causing the effect, CDR also explore whether the effect of other countries' technology evolves as the technology diffuses. Note that, interactions mediated by the flow of people or of goods and services would tend to persist over time. In contrast, interactions driven by the diffusion of knowledge should tend to vanish over time as knowledge is easier to replicate within one location. Consistent with this later hypothesis, CDR find that the effect of SDT on technology adoption diminishes as the diffusion process unfolds.

A final question CDR take on is Jared Diamond's hypothesis that technologies diffuse along latitudes. To explore this, they decompose SDT between two components one based on distances along latitudes and another based on distances along longitudes. Consistent with Diamond (1999), they find that latitude component of SDT has a stronger association with technology adoption than longitude component of SDT, although both have a significant effect. This finding is remarkable since their sample does not include any technology where climatic reasons might suggest that distance across latitudes is a larger impediment for diffusion than distance across longitudes.

One last form of geographic interaction considered in the literature is migration flows. International migration may have brought significant technological knowledge from areas with more advanced technologies to others where advanced technologies were rare. If that is the case, one would expect that adjusting knowledge flows by the movement of people should provide a more accurate account of the cross-country dynamics of productivity. With this idea in mind, Putterman and Weil (2010) explore whether the history of people matters more than history of places by measuring, for each country, the origin of the ancestors of today's population, going back to 1500 AD. Putterman and Weil show that, after adjusting for historical migration flows, variables such as the antiquity of states and years of experience with agriculture have a greater explanatory power of current development. Comin et al. (2010) extend Putterman and Weil's analysis by applying the Putterman-Weil migration matrix to the historical technology adoption in 1500 AD. This yields a measure of the historical adoption level in 1500 AD of the ancestors to the people that live today in each country. With this migration-adjusted measure of technology, they re-examine their exploration of the persistence of technology adoption.

Figure 2.4 shows the simple scatter plot between migration-adjusted technological heritage from 1500 AD and per-capita income today. Comparing Figure 2.4 with Figure 2.3, it is clear that long-run technological persistence is stronger overall if we base technology on peoples rather than places (the R-squared increases from .18 to .50). A movement from 0 to 1 is associated with an increase in per-capita income today by a factor of 26.1! Similarly, this regression implies that 78% of the log difference in income today between sub-Saharan Africa and Western Europe is associated with the technology differences in 1500 AD. Based on this evidence, it is clear that the propagation of technological knowledge through migration flows is an important source of cross-country differences in technology adoption.

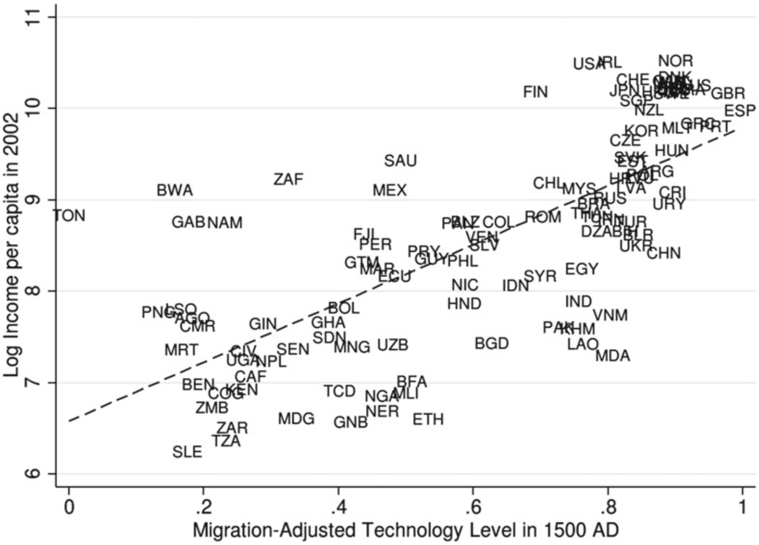


Figure 2.4 Migration-adjusted technology in 1500 AD and current development.

2.3.2 Institutions and Policies

In the same way that insufficient technological knowledge constrains the agents' ability to use a new technology in a productive way, inadequate political institutions may reduce the agents' incentive to incur in the costs of using a new technology. There is no lack of theoretical arguments and anecdotal evidence that point to specific mechanisms by which inadequate institutions may effectively block the diffusion of new technologies.[27] Broadly speaking, we can classify these arguments into two groups. One common argument is that bad institutions may not effectively protect the rights of adopters over their technologies or the income they generate. The threat of this risk of expropriation suffices to deter agents from investing in new technologies. A second argument stresses the redistributive consequences of the diffusion of new technologies. For example, Olson (1982) argues that new technologies may eliminate the rents of producers that have significant physical or human capital invested in older technologies. Acemoglu and Robinson (2000) emphasize that the diffusion of some new technologies that facilitate transportation and communication may reduce the political power of some elites. Bad institutions may enable threatened political or economic incumbents to raise barriers to the diffusion of the technologies that jeopardize their economic or political rents.

Despite the abundance of theories that model these logic and narratives that anecdotally provide some evidence, we have few systematic analyses that evaluate the general relevance of political constraints on technology diffusion. Comin and Hobijn (2004) explore the first hypothesis using their sample of 25 technologies and 23 advanced

[27] See, for example, the review by Acemoglu et al. (2005) and the references therein.

countries. They consider various characteristics of the political institutions of each country as regressors in (2.28). These include a set of dummies for the type of effective executive (monarch, president, premier, or lack of effective executive); a dummy for whether the regime is military; and an index of the legitimacy of the party system that measures whether no party is excluded from participating in the political process. Table 2.7 reproduces their findings. There are two main observations. First, both not having an effective executive and having a military regime are associated with a lower level of technology in the country. This is consistent with the notion that property right protection is a necessary condition for adopting technologies. Second, an effective legislature is associated with a less intense adoption of technologies. This finding may be surprising, but we try to rationalize it below.

The role of redistributive politics on technology diffusion has also been explored recently. Comin and Hobijn (2009b) bring to the data Olson (1982) theory and explore the significance of barriers erected by incumbent producers on the speed of diffusion of new technologies. Their identification strategy consists of two parts. First, certain institutional attributes affect the political cost faced by the legislature when raising barriers to the diffusion of a new technology. In particular, the cost lobbies must incur to induce legislators to raise diffusion barriers is higher when legislators are not independent, the judicial system is effective, and the regime is democratic and non-military.[28]

Second, the benefits old technology producers enjoy from raising barriers against the diffusion of a new technology depend on certain attributes of the new and old technologies. There are some new technologies that are so superior to the old technology that, even with political barriers, consumers prefer the new technology to the old one. In these cases, old technology producers find no benefit in lobbying for barriers. Thus, the new technology will diffuse quickly regardless of the costs of lobbying. Other new technologies do, however, have close predecessor technologies because the productivity differential between old and new technologies is relatively small. In these cases, old technology producers may benefit from barriers to the new technology because in the presence of barriers consumers may prefer to use the old technology. The speed of diffusion of these new technologies depends, therefore, on the cost of erecting barriers. When it is costly to raise political barriers, lobbying is unsuccessful, barriers are not raised, and new technologies diffuse quickly. Conversely, when the cost of raising barriers is low, the legislative authority accepts the old technology's lobbying bribes and raises barriers that slow down the diffusion of the new technology.

It follows from these two premises that, the effect of lobbies on technology diffusion can be identified by the differential effect of institutions. If lobbies matter, institutions that affect the political cost of erecting barriers should have a differential effect on the diffusion of technologies with close predecessors relative to those without close predecessors.

[28] See, Myerson (1993), Ferejohn (1986), Persson et al. (2000), Kunicová and Rose-Ackerman (2005), Persson et al. (2003), and Besley and Case (1995).

Comin and Hobijn use a sample of 20 technologies, and 23 countries over the last 200 years from HCCTAD. Their main finding is that variables that affect the cost of lobbying such as how democratic is a country, the judicial effectiveness, and whether the regime is military, have a differential effect on the diffusion of technologies with a close predecessor of the expected sign (higher cost is associated with higher differential adoption). Similarly, Comin and Hobijn (2009b) find that a more independent legislative power is associated with a negative differential diffusion of the technologies with a close predecessor. They rationalize as evidence that, other things equal, a more independent legislature faces less constraints to pass regulations that favor powerful lobbies, which tend to be those of the incumbent technologies. It is important to stress that these results are robust to including country and year-dummies as well as country-dummies that are specific to incumbent and new technology groupings.

Though the evidence presented in Comin and Hobijn (2009b) is supportive of the role of distributive politics on technology diffusion, it is important to be aware of two important limitations of this study. First, it evaluates only one particular theory of redistributive politics and technology diffusion (Olson, 1982). Second, the sample used covers only advanced economies. Hence, more studies are necessary to make an accurate assessment of the global significance of political institutions on technology adoption patterns.

Trade Openness One of the reasons why institutions may matter is because they affect the policies implemented by governments. Among those, trade policies have probably received most attention. Sachs and Warner (1995), Frankel and Romer (1999), and Feyrer (2009a,b) showed that trade has a significant impact on income growth. A natural question is whether the channel by which trade affects growth is technology adoption. This question is still largely unexplored. All the existing evidence we are aware off basically consists in including measures of trade openness in specification (2.28).[29] Comin and Hobijn (2004) find that countries whose trade makes up a larger part of its GDP are the front runners in technology adoption. The coefficient on openness is significant for the bulk of our specifications and its magnitude implies that countries that are 12–15% more open than others will be 1% ahead in the adoption of technologies.

Related to this, Coe and Helpman (1995) find a strong effect of the technological advancement of the trading partner on TFP. Again, it is natural to inquire whether this effect operates through the adoption of new technologies. To explore this hypothesis, Comin and Hobijn (2004) include as regressors the trade-weighted averages of GDP and the trade-weighted level of technology adoption for the trading partners. Somewhat surprisingly, they obtain a negative coefficient which they interpret as evidence that the effect of trade on TFP may be affecting factors other than a more intensive adoption of technology. Again, this seems an area where there is room for more research in the future.

[29] Lucas et al. (2011) and Perla et al. (2012) provide structural models of trade and growth that account for the diffusion process. They rely on calibrations to assess the link between trade and growth.

2.3.3 Demand

The level of demand is an important determinant of the return to adopting a technology. A higher demand allows adopters to cover the sunk costs of adoption among more buyers of the goods and services produced with the technology. Therefore, increasing the profitability of the investment. Even when the costs of adoption are negligible, we should expect larger demand for the goods and services that embody a technology in places and times where aggregate demand is higher. This is clear from Equation (2.22).

The notion that demand is an important driver of technology has been recognized at least since Schmookler (1966). Schmookler argued that demand should play a key role both in the amount of innovation activity as well as in the sectors where it was concentrated. He brought this hypothesis to the data by exploring how patenting activity in capital intensive sectors correlated with lagged investment (his measure of demand pull). Both in cross-sections of sectors and in the time series within sectors, Schmookler (1966) found a strong co-movement between lagged investment and patenting activity.[30] Subsequent research has explored the cyclicality of R&D activities (Griliches, 1990; Fatas, 2000; Comin and Gertler, 2006). The robust finding is that R&D expenditures positively co-move with output at business cycle frequencies and that the co-movement increases when we consider lower frequencies.[31] Most of this evidence is at the aggregate level. The exception is Barlevy (2007), who found a positive co-movement between firm-level growth in real R&D expenditures and 4-digit sector level growth in aggregate demand.

Nevertheless, R&D and technology adoption are distinct activities that are undertaken by different companies and that also differ in their geography. Is there evidence of the importance of demand pull forces for the adoption of new technologies?

In Table 2.7, Comin and Hobijn (2004) introduce the log of income per capita as a control (see Equation (2.28)). They find that the elasticity of technology with respect to income is around 1 (and significant at the 1% level). Of course, the concern that per-capita income captures variables other than demand is legitimate. This concern, however, should be mitigated to some extent by the fact that the estimate of the income elasticity does not decline after controlling for potential omitted variables such as institutions, openness, human capital, and adoption of predecessor and complementary technologies.

Comin (2009) follows a different approach to estimating the elasticity of technology with respect to income. Following the traditional diffusion literature, he poses an S-shaped diffusion process modified to allow for the speed of diffusion to depend on deviations from trend of GDP. In particular, consider the (log) ratio of adopters (m_{jt}) to non-adopters ($M - m_{jt}$) for a generic technology j:

$$r_{j_t} \equiv \ln(m_{jt}/(M - m_{jt})).$$

[30] See Scherer (1981) for a confirmation of these findings in a larger number of sectors and other indicators of demand pull.

[31] Specifically, cycles with periods between 8 and 50 years.

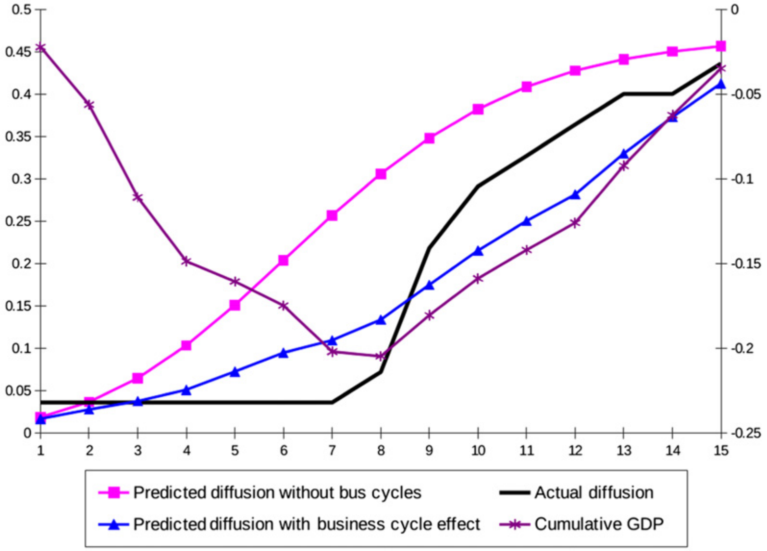


Figure 2.5 Diffusion of numerical control turning machines in the UK.

Note that the first difference of r_{jt} is the speed of diffusion of the technology. If the share of adopters (m_{jt}/M) follows a logistic curve, then the speed of diffusion (Δr_{j_t}) is constant. Comin explores the constancy of the speed of diffusion in a sample of 22 manufacturing processes in the UK.[32] He finds that the speed of diffusion is far from constant and it tends to decline as the technology diffuses. To capture this pattern and to explore the cyclicality of the speed of technology diffusion, Comin (2009) poses the following specification:

$$\Delta r_{j_{t+1}} = \beta_j + \alpha_1 t_j + \alpha_2 t_j^2 + \gamma y_{2200t} + \epsilon_{jt}, \tag{2.29}$$

where t_j is the number of years since the invention of technology j, and y_{2200t} is GDP detrended so that we keep fluctuations with periods between 2 and 200 quarters.

Comin (2009) obtains an estimate of the elasticity of the speed of diffusion with respect to detrended GDP of 5.12 with a 95% confidence interval of (1.9, 8.34). Figure 2.5 plots the implications of Equation (2.29) for the diffusion of numerical control turning machines. The black solid line represents the actual evolution of the share of adopters; the line with triangles plots the diffusion path predicted by Equation (2.29); the line with

[32] The data comes from Davies (1979) and spans from WWII to the late 1970s. The technologies include special presses, foils, wet suction boxes, gibberellic acid, automatic size boxes, accelerated drying hoods, electrical hygrometers, basic oxygen process, vacuum degassing, vacuum melting, continuous casting, tunnel kilns, process control by computer, tufted carpets, computer typesetting, photo-electrically controlled cutting, shuttleless looms, numerical control printing presses, numerical control turning machines, and numerical control turbines.

stars plots the evolution of cumulative detrended output; and the line with squares plots the diffusion path predicted by Equation (2.29) when the effect of the cycle on diffusion is ignored.

Ignoring the business cycle component in Equation (2.29) has important consequences. Eight years after the introduction of numerical control turning machines in the UK only 7% of potential adopters had adopted the technology. The model without the cycle component predicts that over 30% of potential adopters should be using the technology. One explanation for the slow diffusion of numerical control turning machines in the UK is that at that point it had been for eight consecutive years below trend. As a result, producers faced a low demand and had few incentives to invest in the new technology. Once this effect is taken into account, Equation (2.29) predicts that after 8 years the diffusion of numerical control turning machines in the UK should have been 13%, much closer to the actual 7%.

This evidence raises the question of why are the estimates for the income elasticity of technology are so different in Comin (2009) and Comin and Hobijn (2004). Is the difference due to differences in the measurement of technology, or to differences in the income frequencies considered?

Findings in Comin and Mestieri (2010) may start to shed some light on this question. Specifically, they estimate a version of Equation (2.25) where they decompose GDP between the cyclical and the trend component, and allow for different elasticities of technology with respect to each component. The cyclical component is captured by HP-filtered output and the non-cyclical component is the HP-trend. Comin and Mestieri use the same 15 technologies considered in Comin and Hobijn (2010), and estimate these elasticities for the US. Furthermore, they restrict the elasticities of technology with respect to each component of income to be the same across technologies. They obtain an income elasticity of 2.2 with respect to the HP-trend and of 6.6 with respect to the cyclical component of GDP. These estimates are consistent with the evidence presented above. In particular, they seem consistent with a much higher elasticity of technology diffusion with respect to cyclical measures of output than with respect to the trend in output with the former in the vicinity of 5 and the latter between 1 and 2. These estimates also confirm the importance of aggregate demand for technology diffusion.

2.4. EFFECTS OF TECHNOLOGY ADOPTION

In the final section of this chapter, we explore the macroeconomic consequences of technology. We organize our exposition according to the frequency of interest. First, we explore the roles played by technology in business cycle fluctuations. Then, we perform a development accounting exercise. We quantify the contribution of cross-country differences in technology diffusion to cross-country differences in income. Finally, we explore how changes in technology adoption patterns over time help us explain cross-country growth differences over protracted periods of time.

2.4.1 Business Cycles Fluctuations

The role conferred to technology in business cycle analysis has traditionally been reduced to an exogenous disturbance. This is a natural consequence of using the neoclassical growth model as the workhorse framework for business cycle analysis. Real business cycle models (e.g. Kydland and Prescott, 1982) intend to synthesize growth and business cycle fluctuations by introducing stochastic disturbances to the exogenous technology process. The propagation of the shocks is then governed by the dynamics of capital accumulation in the neoclassical model. In early versions of RBC models, total factor productivity (TFP) is interpreted as a measure of the technology in the economy. Growth in TFP drives growth in the long term. TFP has two components. A deterministic trend and exogenous, stochastic deviations from the trend that drive short-term fluctuations in the economy. Subsequent approaches to business cycle analysis have minimized the role of technology shocks as a source of business cycle fluctuations (Galí, 1999) and have re-interpreted the observed short-term fluctuations in TFP as reflecting cyclical variation in the intensity of use of the factors of production or in the degree of competition (Burnside et al. 1995; Basu and Fernald, 1997).

Next, we review two new lines of work on the role of technology on business cycle fluctuations. The first re-examines the classical question of the importance of technology shocks for business fluctuations by using more direct measures of technology in the identification of technology shocks. The second line of work proposes technology as a propagation mechanism based on the evidence discussed above on the cyclicality of R&D expenditures and the speed of diffusion of technology.

2.4.1.1 Shocks

The identification of technology shocks has been challenging in macroeconomics. Traditional approaches involve using indirect inference techniques based either on modified Solow residuals (e.g. Basu, 1996) or on restrictions on the responses from vector-autoregression (VAR) models (e.g. Galí, 1999). Alexopoulos (2011) revises the classical question of what is the short-term effect of a technology shock using her technology-books measure to identify technology shocks. To explore this issue, she detrends the measures of the number of technology books edited using a band-pass filter that permits her to consider two types of frequencies: the business cycle (periods between 2 and 8 years) and the medium-term cycle (periods between 2 and 30 years).[33]

Alexopoulos shows that both at high- and medium-term frequencies, her technology measures lead GDP fluctuations. The highest correlations are between the number of

[33] Note that she uses an upper bound that is slightly lower than the one used in Comin and Gertler (2006).

books published on computer software and hardware and GDP at $t+1$ which she finds to be 0.47. Technology books also lead investment and TFP by one or 2 years, specially when considering medium-term cycles.[34] Interestingly, the number of technology books edited has a sizable volatility both over the business cycle and the medium-term cycle. As a way of comparison, the standard deviation of the high-frequency measure of technology books edited with the lowest volatility—Alexopoulos has five different measures—has a standard deviation of 3.5% vs. 1.4% for GDP. Estimating a bivariate system where log GDP is ordered first and the technology measure second, she finds that, at a 3-year horizon, technology accounts for at least 9% of business fluctuations in log GDP and, at a 9-year horizon, this figure raises to 37%. These figures suggest that fluctuations in the technology available to firms (as proxied by the number of technology books edited) may be a significant source of business cycles fluctuations.

2.4.1.2 Propagation Mechanisms

As showed by Cogley and Nason (1995), capital accumulation dynamics are a weak propagation mechanism. The persistence of RBC models is basically the persistence of their shocks. As a result, these models do not explain the persistence of macro variables, they just assume it (in the form of persistent shocks). This approach is clearly unsatisfactory. On the one hand, macro shocks typically are not as persistent as macro series. On the other, it is of critical importance both for descriptive and prescriptive reasons to understand what mechanisms propagate short-term shocks and generate effects that last for several quarters and years.

Comin and Gertler (2006) explore this question by deviating from the standard macro framework. In particular, based on the evidence that R&D and adoption investments are pro-cyclical, they challenge the view that technology is exogenous. They do that by building a business cycle model where, as in endogenous growth models (e.g. Romer, 1990), technology is the result from purposeful investments by agents/firms in developing and adopting new technologies. The investments in upgrading technology (by innovation or adoption) affect the stock of technologies available for production. In this way, technology emerges naturally as a new state variable that may significantly affect the dynamic response of the model to business cycle shocks.

Next, we sketch a simplified version of the framework developed in Comin and Hobijn (2007) and review its implications.

A Business Cycle Framework with Endogenous Technology Our description focuses mostly on the production side of the economy. This version of the framework is

[34] With respect to hours, Alexopoulos finds that there is no significant contemporaneous effect of technology on hours work at the high frequency but she finds strong positive effects of current technology on hours at $t+1$ and at $t+2$. Over the medium term, she finds a negative and sizable contemporaneous association between hours and technology and a milder effect after one and two years.

a one-sector model with development and adoption of intermediate goods which affect the efficiency of production in the economy.

Final output—Final output is produced competitively according to:

$$Y_t = \left((U_t K_t)^{\alpha} L_t^{1-\alpha}\right)^{\gamma} M_t^{1-\gamma}, \tag{2.30}$$

where U_t denotes the intensity of utilization of capital, K_t, L_t denotes hours worked, and M_t denotes the materials used in production. Utilization comes at the cost of a faster depreciation of capital. Materials are produced competitively combining differentiated intermediate goods according to:

$$M_t = \left(\int_0^{A_t} x_{it}^{1/\theta}\right)^{\theta},$$

where $\theta > 1$ will also be the markup charged by the producer of intermediate good i. It takes one unit of final output to produce one unit of any intermediate good. A_t denotes the number of intermediate goods available for production at time t. Note that if producers use the same number of units of all intermediate goods available for production (e.g. x_t), then (2.30) can be expressed as:

$$Y_t = A_t^{(1-\gamma)(\theta-1)} \left((U_t K_t)^{\alpha} L_t^{1-\alpha}\right)^{\gamma} (A_t x_t)^{1-\gamma}. \tag{2.31}$$

The last term in (2.31) is the amount of output used to produce intermediate goods and the first term is the technological component of TFP.

Adoption—Intermediate goods are first invented and then are adopted for production. We first characterize the adoption process conditional on the available set of technologies, and then describe the research and development process that leads to new technologies.

Let Z_t denote the stock of invented technologies. The stock of not-yet-adopted technologies is $Z_t - A_t$. Each period, a fraction of the available new technologies become usable. Whether a technology becomes usable is a random draw with success probably λ_t. Once a technology is usable, all firms are able to employ it immediately. Note that under this setup there is slow diffusion of new technologies on average (as on average there is a lag between their invention and adoption dates). Furthermore, aggregation is simple as once a technology is in use, all firms have it.

Formally, the number of adopted technologies, A_t, is given by:

$$A_t = \lambda_t [Z_{t-1} - A_{t-1}] + \phi A_{t-1}, \tag{2.32}$$

with $0 < \phi < 1$ representing the probability that the technology has not become obsolete in one period, and $0 < \lambda_t < 1$. We assume that λ_t is given by the following function:

$$\lambda_t = \lambda(\Gamma_t h_t),$$

with $\lambda' > 0, \lambda'' < 0$, where Γ_t is the scaling factor that guarantees the existence of a balanced-growth path and exogenous to the adopter and h_t are the resources devoted to adopting a technology in time t.[35]

If ϕ is close to 1, it follows from (2.32) that transitory changes in λ_t will have transitory changes in the growth rate of A_t, but close to permanent changes in the level of A_t. This property is key to making endogenous technology a powerful propagation mechanism. However, it is not sufficient. In addition, it is necessary that λ_t is pro-cyclical. To explore the cyclicality of λ_t, we next endogenize it.

The value to the adopter of successfully bringing a new technology into use, v_t, is given by the present value of profits from operating the technology. Profits each period π_t arise from the fact that the producer of the new good is a monopolistic competitor. Accordingly, given that R_{t+1} is the one period discount rate between $t+1$ and t, we can express, v_t, as:

$$v_t = \pi_t + \phi E_t \left[\frac{v_{t+1}}{R_{t+1}} \right]. \tag{2.33}$$

If an adopter is unsuccessful in the current period, he may try again in the subsequent period to make the technology usable. Let w_t be the value of acquiring an innovation that has not yet been adopted. w_t is given by:

$$w_t = \max_{h_t} -h_t + \phi E_t \left[\frac{[\lambda(\Gamma_t h_t) v_{t+1} + (1 - \lambda(\Gamma_t h_t)) w_{t+1}]}{R_{t+1}} \right]. \tag{2.34}$$

At the margin, adopters determine how much to spend in adopting a technology by equalizing the marginal cost and the expected marginal benefit from adoption:

$$1 = E_t \left[R_{t+1}^{-1} \phi \Gamma_t \lambda'(\Gamma_t h_t)(v_{t+1} - w_{t+1}) \right]. \tag{2.35}$$

The expected marginal benefit has three terms. The first captures the discounting of the potential benefits from adoption (i.e. $R_{t+1}^{-1}\phi$); the second captures the marginal increase in the probability of succeeding in adopting when firms invest one extra unit of output (i.e. $\Gamma_t \lambda'(\Gamma_t h_t)$); the third captures the capital gain that occurs when an intermediate good becomes adopted (i.e. $v_{t+1} - w_{t+1}$). When the economy is booming, the expected capital gain increases because of the higher demand of adopted intermediate goods. Firms respond to this by investing more resources in adoption (h_t). This is why adoption expenditures and the speed of diffusion of technologies, λ_t, are pro-cyclical in the model.

Innovation—Now that we have solved for the adoption process given the number of available technologies, we need to derive the equations that determine the number of available innovations, Z_t.

[35] Comin and Gertler (2006) model Γ_t as inversely proportional to the (wholesale) value of capital. It is important that Γ_t is smooth.

New technologies are developed through R&D. Each innovator, indexed by p, faces the following technology to develop new intermediate goods:

$$Z_{t+1}\,(p) - Z_t\,(p) = \varphi_t S_t(p) - \phi Z_t(p), \tag{2.36}$$

where $Z_t(p)$ denotes the stock of (non-obsolete) intermediate goods she has developed up to time t, $S_t(p)$ is the number of units of output she devotes to R&D, and φ_t is the productivity of the R&D as perceived by the individual innovator. As in Romer (1990), the linear formulation permits a simple decentralization of the innovation process.[36]

We assume that φ_t depends on the aggregate values of the stock of innovations, Z_t, the scaling factor, Γ_t, and research and development S_t, and the stock of innovations as follows:

$$\varphi_t = \chi Z_t(S_t)^{\rho-1}(\Gamma_t)^{\rho}, \tag{2.37}$$

with $0 < \rho \leq 1$ and where χ is a fixed scale parameter. This formulation allows for aggregate congestion in R&D investments.[37]

The linearity of the R&D technology as perceived by the individual researchers together with a free entry assumption implies that each new product developer p must break even. As a result, the resources invested in R&D by the pth innovator satisfy the following arbitrage condition:

$$E_t\left[\frac{w_{t+1}}{R_{t+1}}\right] - 1/\varphi_t = 0,$$

where the first side is the discounted marginal benefit from an innovation and the left side is the marginal cost in units of final output. Note that the pro-cyclicality of w_{t+1} will tend to generate pro-cyclical R&D expenditures in equilibrium.

The resulting law of motion for the number of intermediate goods, Z_t, is:

$$Z_{t+1} - Z_t = Z_t\chi\left(S_t/P_t^I K_t\right)^{\rho} - (1-\phi)Z_t.$$

The model is closed by adding a law of motion for capital and the preferences of consumers which are standard.

Impulse Response Functions Comin and Gertler (2006) study a two-sector version of the model we just sketched and introduce shocks to the wage markup. These are non-technological shocks that capture frictions in labor markets and that are isomorphic to shocks to labor income tax and similar to money shocks. Figure 2.6 reproduces the impulse response of their model (solid) and compares it to a version of the model where the endogenous technology mechanisms have been shut down.

[36] We differ from Romer (1990), however, by having the innovation technology use as input a final good composite of capital and labor, as opposed to just labor. See also Barlevy (2007) for a discussion of the relevance of this choice.

[37] As with Romer, there is a positive spillover of the current stock of innovations on the creation of new products, i.e. φ_t increases linearly in Z_t.

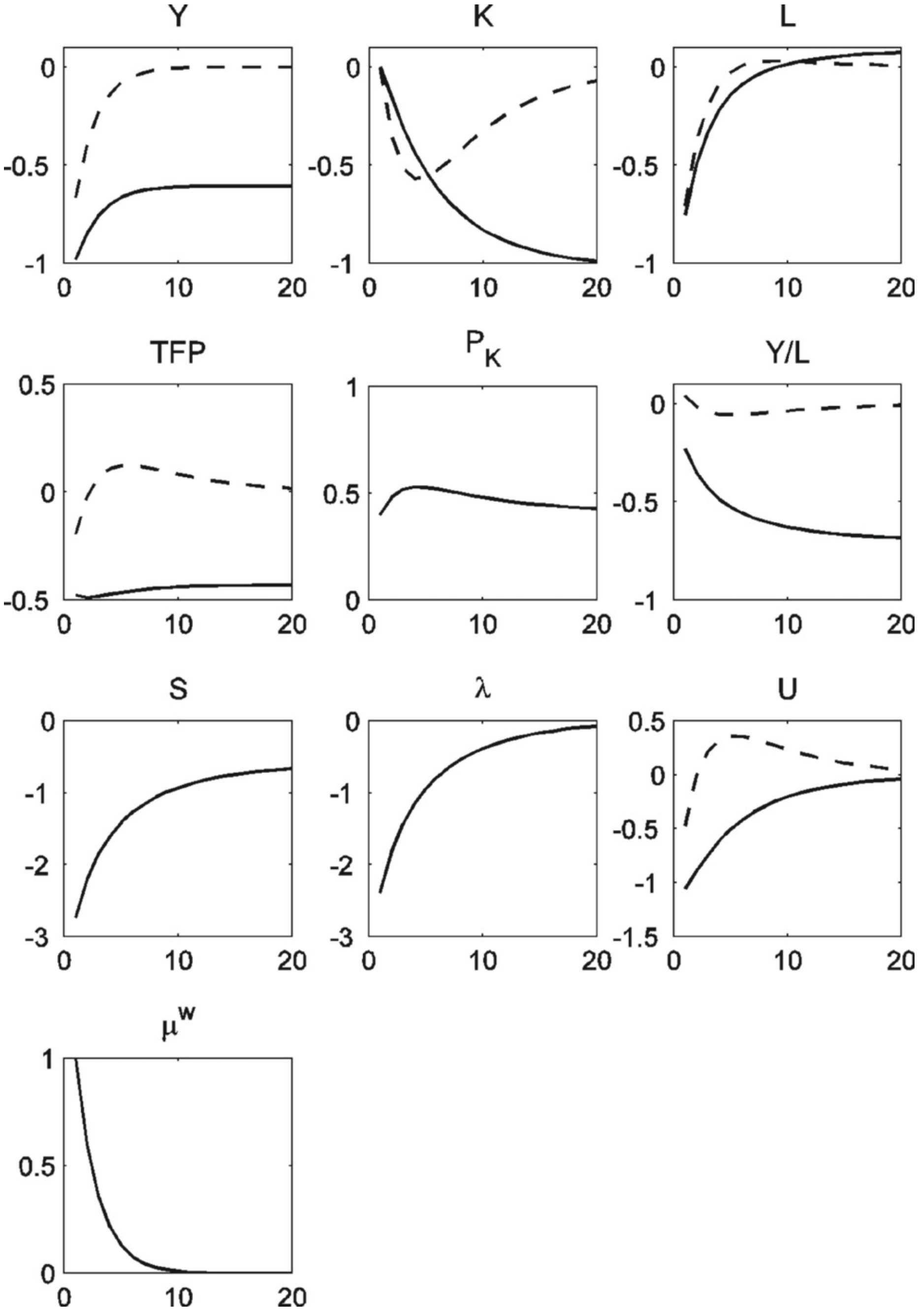


Figure 2.6 Impulse response functions.

The increase in the wage markup effectively raises the price of labor, reducing labor demand and output. Both the initial impact and the impact over time on output are larger in the model with endogenous technology. Over time, output climbs back to trend, but does not make it back all the way due to the endogenous permanent decline in productivity. This is captured by the evolution of TFP. The initial decline in measured TFP results mainly from variable factor utilization. Over time, there is a decline in true productivity relative to trend. In particular, the initial contraction in economic activity induces a drop in both R&D and the rate of technology adoption. The temporary drops in R&D and adoption slows down the rate at which new technologies are incorporated into production, ultimately leading to a permanent drop relative to trend in total factor productivity and labor productivity. In contrast, the model without endogenous productivity, output simply reverts back to its initial steady state. Hence, the propagation power of endogenous technology mechanisms.

Applications The framework outlined above has shed light into business cycle phenomena and historical episodes that are hard to rationalize by standard macro models.

One extreme case of a persistent economic downturn is Japan's so-called Lost Decade of the 1990s, when the country experienced very low growth for a whole decade despite the fact that the shocks that hit the economy lasted, at most, for 3 years. This is puzzling for standard macro models because they predict a quick recovery of the economy once the shocks are over. More elaborate theories based on realistic features of Japan's context, such as zombie companies (e.g. Caballero et al. 2008), or policy mistakes could increase the recession's duration. The endogenous technology framework can provide a natural complement for these theories. Consistent with the model's predictions, Japanese firms did, in fact, slow down their R&D intensity and slowed the adoption rate of new technologies during the 1990s, falling behind Korea in computer and internet-usage rates. Comin (2011) documents these facts and explores their consequence for the dynamics of output showing that they contribute significantly to the protractedness of the lost decade during the 1990s.

Standard macro models have had trouble reconciling the empirical pro-cyclicality of stock prices and the counter-cyclicality of the relative price of investment. The traditional strategy has been to use unrealistic adjustment costs to new investment. Two sector versions of the framework can rationalize the cyclical properties of stock prices and the relative price of capital. Intuitively, endogenous improvements in the productivity of the capital producing sector will lead to a counter-cyclical cost of production of investment goods and hence of the relative price of investment. Once technology is endogenous, their price is much more than the market value of installed capital. Their price includes the market value of the current and future technologies they will develop. These technological components of a company's value are pro-cyclical and fluctuate much more than the price of capital, hence dominate the value of capital in determining the cyclicality of stock prices.[38]

[38] See Comin et al. (2009a). Also, see Santos and Iraola (2010) for an exposition of the determinants of stock prices in the Comin and Gertler (2006) model.

Endogenous technology models of business fluctuations may also help us understand better international co-movement patterns. Shocks to developed economies have large and persistent effects in developing countries (see Comin et al. 2009b and the references therein). Standard international macro models struggle to explain the magnitude of these effects. In these models, domestic impulses only affect other economies by reducing demand for their exports. This channel is insufficient, both in the data and in the models, to account for the magnitude of the international propagation of shocks we seem to observe.

An additional international linkage arises when the technology available for production may be affected by disturbances to other economies. Comin et al. (2009b) document that the number of new technologies that diffuse from the United States or Japan to their main trade partners in the developing world is pro-cyclical, both with respect to the business cycle in the developed and in the developing economy. Building a two-country extension of the framework outlined above, they show that this mechanism is sufficient to generate an international propagation of fluctuations between developed and developing countries similar to what we observe in the data.

2.4.2 Development

The model presented in Section 2.2.3.3 can be used to explore the aggregate implications of technology diffusion for income and for income dynamics (Comin and Hobijn, 2010; Comin and Mestieri, 2010, 2013). To this end, we take advantage of the aggregate representation of the model. Normalizing aggregate labor to one, aggregate output is given by:

$$Y = AX^{\alpha}L^{1-\alpha} = AX^{\alpha} = A^{1/(1-\alpha)}(\alpha)^{\alpha/(1-\alpha)}, \tag{2.38}$$

with,

$$A = \left(\int_{-\infty}^{\bar{\tau}} Z_{\tau}^{\frac{1}{\theta-1}} d\tau \right)^{\theta-1}, \tag{2.39}$$

where $\bar{\tau}$ denotes the most advanced technology adopted in the economy.

These equations imply that output dynamics are completely determined by the dynamics of aggregate productivity, A. A sufficient condition to guarantee the existence of a balanced growth path is that D_{τ} and a_{τ} are constant across technologies—denoted by D and a.[39] Making the simplifying (and empirically relevant) assumption that $\theta = \mu$, aggregate productivity can be computed in closed form,[40]

$$A(t) = \left(\frac{(\theta - 1)^2}{(\gamma + \chi)\chi} \right)^{\theta-1} a\, e^{(\chi+\gamma)(t-D)}. \tag{2.40}$$

[39] Comin and Hobijn (2010) and Comin and Mestieri (2010) show in their microfounded models of adoption that this is a necessary and sufficient condition. Hence, this is a natural benchmark for us.

[40] This is what we observe in our estimation. We cannot reject the null hypothesis that $\theta = \mu$.

This expression shows that higher intensity of adoption, a, and shorter adoption lags, D, lead to higher aggregate productivity. Along this balanced growth path, productivity grows at rate $\chi+\gamma$ and output grows at rate $(\chi+\gamma)/(1-\alpha)$.[41]

These expressions can be used to explore the relevance of technology diffusion for cross-country differences in productivity. In particular, the model implies that the (log) gap in productivity between country c and the average of the Western countries is equal to:

$$y_c - y_{west} = \frac{a_c - a_{west}}{1-\alpha} + \frac{\chi+\gamma}{1-\alpha}(D_{west} - D_c). \tag{2.41}$$

where a_c and D_c, are respectively, the average intensive margin and adoption lag in country c. Using our estimates from Section (2.2.3.4), we compute a_c and D_c as follows:

$$a_c = \frac{1}{N_c}\sum_{\tau=1}^{N_c} a_{\tau c}, \qquad D_c = \frac{1}{N_c}\sum_{\tau=1}^{N_c} D_{\tau c}, \tag{2.42}$$

where N_c is the number of technologies for which we have precise estimates in country c.

Figure 2.7A and B plot the contribution to TFP of the extensive and intensive margins, which correspond to the first and second terms in Equation (2.41), against log per-capita income in 2000.[42] The thicker dashed line corresponds to the regression line:

$$\left(\frac{a_c - a_{west}}{1-\alpha}\right) = \alpha + \beta(y_c - y_{west}), \tag{2.43}$$

$$\frac{(\gamma+\chi)(D_{west} - D_c)}{1-\alpha} = \delta + \pi\,(y_c - y_{west})\,, \tag{2.44}$$

where we calibrate $\alpha = .3$ and $(\gamma+\chi)/(1-\alpha) = 2\%$ to compute the contribution to TFP of the intensive and the extensive margin. The light gray line in both figures is the 45° degree line. The slope of the regression lines (β, π) can be interpreted as contribution of each margin in accounting for differences in income per capita in 2000. If Equation (2.41) were to explain all the income variation today, we would expect the coefficient of (each) regression to be 1 and the data points to lie on the 45° degree line. We find that the slope for the extensive margin regression displayed in Figure 2.7A is 20%, while for the intensive margin it is almost 60%.[43] If these two margins were uncorrelated

[41] For utility to be bounded, this requires the parametric assumption that $(\chi+\gamma)/(1-\alpha) > \rho$.

[42] The income data is from the Penn World Tables 6.2. Results are very similar using Maddison (2004).

[43] The R^2 of these two regressions are .15 and .49, respectively. If we regress real income per capita in 2000 on the two margins, we find that the combined R^2 is .58. These regressions for the intensive margin are done as described in Comin and Mestieri (2010), filtering a common fixed effect across all estimates for a country, that are measured in terms of capital. This is motivated by a richer structural model that allows for an additional distortion on the price of capital. This correction reduces the estimated contribution of the intensive margin in this regression. Thus, this is our most conservative estimate. Should we not do this correction, we would find that it accounts for almost 70%.

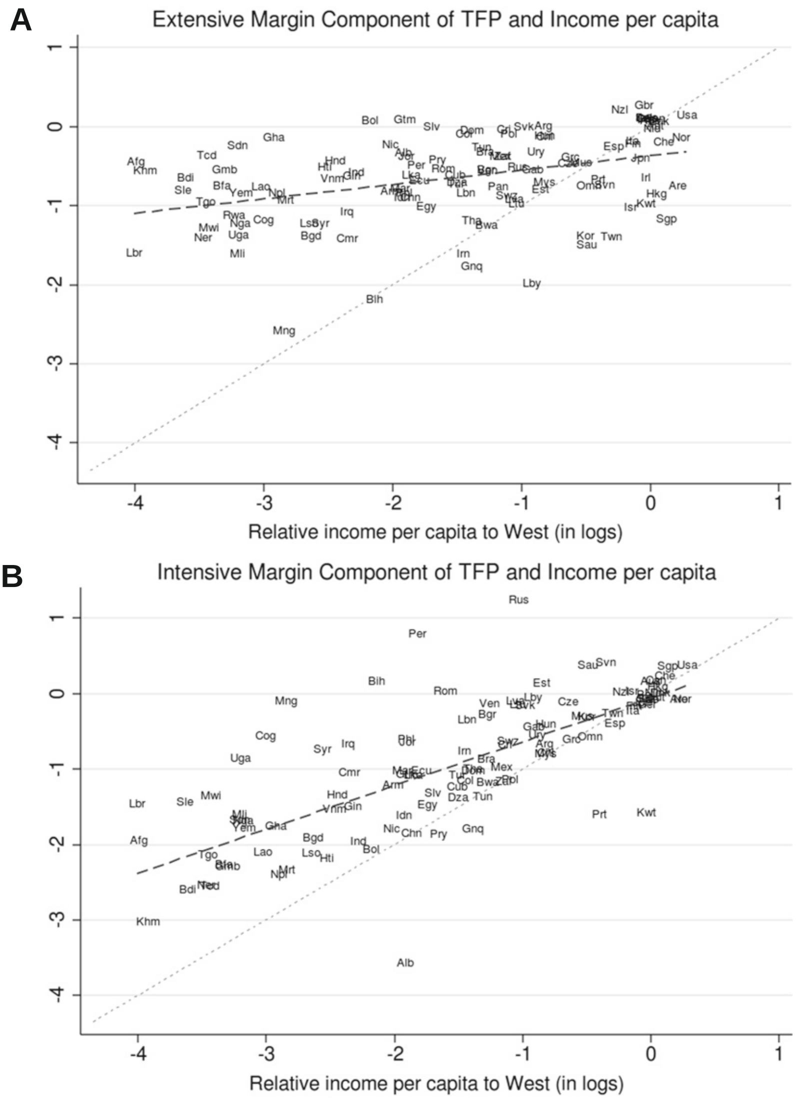


Figure 2.7 TFP components of the intensive and extensive margins.

we could infer that these two margins alone account for most of the variation (80%) of income per capita today. If we regress the extensive margin on the intensive margin and use the unpredicted part as the orthogonal component of the intensive margin not predicted by the extensive, we would find that the slope of Figure 2.7B goes down to 54%. Thus, most of the variation in income per capita today, over 70%, can be accounted for by cross-country differences in the two adoption margins. Among the two, the intensive margin takes the lion's share.

2.4.3 Growth

In this section, we study how the evolution of the two margins of adoption affects the evolution of aggregate productivity of the economy. We use the model from Section 2.2.3.2 to evaluate quantitatively its ability to generate the observed cross-country income growth dynamics over the last 200 years.

Specifically, we focus on three questions: (i) The significance of differences in early adoption to account for cross-country income differences in the pre-industrial balanced growth path; (ii) the protractedness of the model transitional dynamics; and (iii) the model's account of the Great Income Divergence. To this purpose, we keep Maddison (2004)'s division of countries between Western countries and the rest of the world.

Calibration To simulate the model we need to calibrate four parameters. First, we need to specify the path for the world technology frontier. Prior to year $T = 1765$ (the year in which James Watt developed his steam engine), we assume that the technology frontier grew at 0.2%. This is the growth rate of Western Europe according to Maddison (2004) from 1500 to 1800. After 1765, the frontier grows at $(1 - \alpha) \cdot 2\%$ per year. As shown in Equation (2.40), the growth rate along the balanced growth is equal to $(\gamma + \chi)/(1 - \alpha)$. Hence, the Modern balanced growth is 2%. The literature has not determined what fraction of frontier growth comes from each of these two sources. Therefore, we split evenly the sources of growth in the frontier between γ and χ. We take $\alpha = .3$ to match the labor income share.

Finally, we need to calibrate the elasticities of substitution between technologies, which we assume are the same and equal to $1/(\theta - 1)$. We back out the value of θ from the estimates of $\beta_{\tau 3}$. The average value we estimate for θ is 1.28, which is very similar to the values implied by the estimates of price markups from Basu and Fernald (1997) and Norbin (1993). Thus, we set $\theta = 1.28$.

Initial Income Differences It follows from expressions (2.38) and (2.40) that, in our model, differences in productivity in the pre-industrial balanced growth path are due to differences in adoption lags and in the intensive margin in the pre-Modern era. Given that we do not have data for pre-Modern technologies, we assume that pre-Modern levels of adoption were constant and coincide with the initial adoption levels that we estimate. Our estimates from Tables 2.5 and 2.6 imply that the difference between the average adoption lag in the sample of Western countries and in other countries is 49 years in 1820. The average gap in the (log) intensive margin is 0.39. With this assumption and using Maddison's estimates of pre-industrial growth in Western Europe (0.2%) to calibrate the pre-industrial growth rate of the world technology frontier, Equation (2.40) implies an income gap between Western countries and the rest of the world of 90%.[44] This is in line with the results from Maddison (2004), who reports an income gap of the same

[44] That is, $\exp(.2\% \cdot 49 + .39/(1 - \alpha)) = 1.9$.

magnitude. Hence, the pre-industrial income differences generated by our model account very well for those observed in the data.

Protracted Dynamics Next we explore the protractedness of the model transitional dynamics. To this end, we consider the average country in our sample. The average country is parameterized so that its adoption lag and its degree of penetration rate (a_τ) are constant and equal to the average adoption lag and intensive margin across countries over our sample of technologies. In particular, the resulting D is 44 years and the intensive margin is 54% of the US level.

We model the Industrial Revolution as a one time, permanent increase in the growth of the world technology frontier ($\gamma + \chi$), so that the balanced growth path increases from 0.2% to 2%. This view is consistent with Mokyr (1990) and Crafts (1997). Figure 2.8 plots the transition of the output gap in this representative economy. The output gap is defined as the ratio of output in the Modern balanced growth path relative to current output. In the figure, we can see that the model generates a very slow convergence to the new balanced growth path. The half-life of the output gap relative to the Modern balanced growth path is 117 years while for output growth it is 145 years. These half-lives are an order of magnitude higher than the typical half-life in neoclassical growth models (e.g. Barro and Sala-i-Martin, 2003).

There are three reasons why our model generates such protracted dynamics. First, the long adoption lags (44 years) imply that it takes this amount of time for the new technologies (which embody the higher productivity gains) to arrive at the economy. Until then, there is no effect whatsoever in output growth. Second, for a given growth

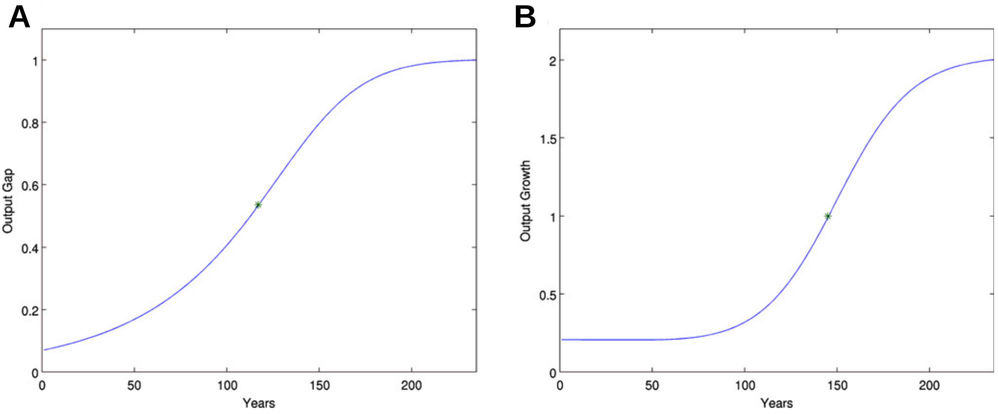


Figure 2.8 Slow transitional dynamics. (A) Consumption gap (relative to the modern BGP). (B) Growth path to modern BGP. This simulation corresponds to the transition to the new balanced growth path after an acceleration of the technological frontier from .2% to 2% for a country with a constant lag as the average lag in our sample (44 years) and average intensive margin (54% of the Western productivity level). The $*$ denotes the half-life.

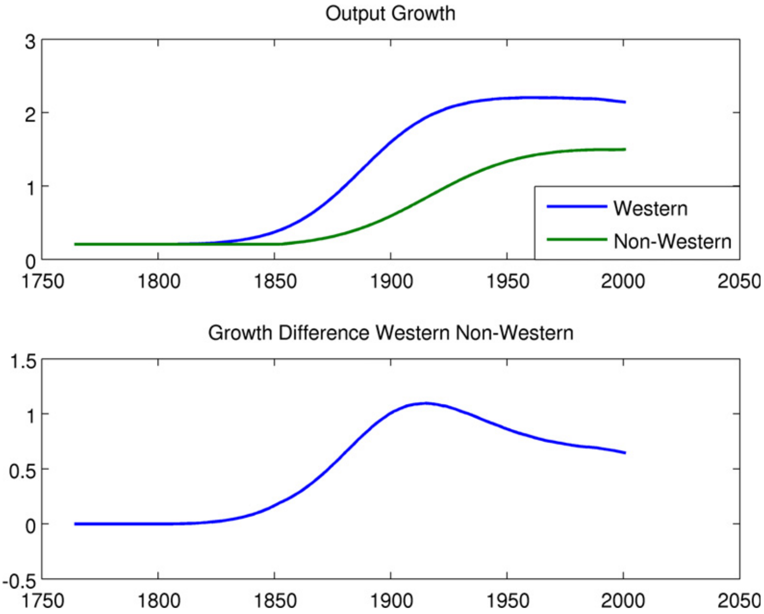


Figure 2.9 Simulated growth for Western and non-Western. Growth of income per capita in the last 200 years for Western and non-Western countries imputing the estimated evolution of the intensive and extensive margins to the baseline model.

in the Modern sector output, its impact in GDP depends on the share of the Modern sector. Since the Modern sector's share increases slowly, so does aggregate output. Third, the growth rate of the Modern sector is initially very small and grows progressively.[45]

Cross-Country Evolution of Income Growth To evaluate the model's power to account for the Great Divergence, we simulate the evolution of output for Western countries and the rest of the world after feeding in a (common) one time permanent increase in frontier growth *and* the estimated evolutions for adoption lags and the intensive margin for each group of countries reported in Tables 2.5 and 2.6.[46] The results from this exercise are reported in Figure 2.9 and Table 2.9.

The model generates sustained differences in the growth rates of Western and non-Western countries for long periods of time. Output growth starts to accelerate at the beginning of the 19th century in the Western economy, converging to the steady-state growth of 2% in the early 20th century. For the non-Western country, instead, growth

[45] Comin and Mestieri (2013) analyze the properties of the transitional dynamics of this model providing theoretical ground for this explanation.

[46] We assume that after the last technology invented in our sample (the Internet, in 1983), the estimated margins remain constant at their 1983 values. This ensures that both groups of countries exhibit the same long-run growth. This assumption is quantitatively inconsequential, as it only affects the dynamics of the last 10 years of our simulations. If anything, it tends to understate the effect of technology dynamics.

Table 2.9 Growth rates of GDP per capita

		Time period		
		1820–2000 (%)	**1820–1913 (%)**	**1913–2000 (%)**
Simulation	Western countries	1.47	.84	2.15
	Rest of the world	.82	.35	1.31
	Difference West-Rest	.65	.49	.84
Maddison	Western countries	1.61	1.21	1.95
	Rest of the World	.86	.63	1.02
	Difference West-Rest	.75	.58	.93

Notes: Simulation results and median growth rates from Maddison (2004). We use 1913 instead of 1900 to divide the sample because there are more country observations in Maddison (2004). The growth rates reported from Maddison for the period 1820–1913 for non-Western countries are computed imputing the median per-capita income in 1820 for those countries with income data in 1913 but missing observations in 1820. These represent 11 observations out of the total 50. We do the same imputation for computing the growth rate for non-Western countries for 1820–2000. This represents 106 observations out of 145. For the 1913–2000 growth rate of non-Western countries, we impute the median per-capita income in 1913 to those countries with income per-capita data in 2000 but missing observations in 1913. These represent 67 observations out of the total 145.

does not increase from the pre-industrial rate until the end of the 19th century. Growth in the non-Western country slowly accelerates, but it is still around 1.5% by year 2000. The gap in growth between both countries is considerable. Annual growth rates differ by more than 0.7% for over 100 years. The peak gap is reached around 1915 at 1.1%. From then, the gap declines monotonically until reaching around 0.6% by 2000. Table 2.9 reports the average growth and growth gaps of our simulation comparing it to Maddison (2004). The patterns and levels in our data trace quite well Maddison's.

The sustained cross-country gap in growth produced by the model leads to a substantial gap in income per capita. In particular, our model generates a 3.2 income gap between the Western countries and the rest of the world. Maddison (2004) reports an actual income widening by a factor of 3.9 between Western countries and the rest of the world since the Industrial Revolution. Hence, most of the variation (82%) in the income gap between Western and non-Western countries in the last two centuries is accounted for.

The simulation also does well in replicating the time-series income evolution of each country group separately. For Western countries, Maddison (2004) reports an 18.5-fold increase in income per capita between 1820 and 2000. Approximately 19% of this increase occurred prior to 1913. In our simulation, we generate a 14-fold increase over the same period, and 16% of this increase is generated prior to 1913. For non-Western countries, Maddison (2004) reports an almost 5-fold increase, with around 37% of the increase being generated prior to 1913. Our simulation generates a 4.3-fold increase in the 1820–2000 period with 32% of this increase occurring pre-1913. The fact that we under-predict the time-series increase in output, reflects, in our view, our omission of factor accumulation dynamics (e.g. human capital), which also contributed to income growth.

The Role of the Evolution of Adoption Margins After showing that the model does a remarkable job in reproducing the cross-country dynamics of income growth over the last two centuries, Comin and Mestieri (2013) dissect the mechanisms at work. In particular, we simulate the economy shutting down, sequentially, the dynamics in each adoption margin. From this exercise, we conclude that the large cross-country differences in adoption lags explain much of the income divergence during the 19th century. However, the convergence in adoption lags that we have documented would make non-Western countries grow faster than Western countries during the 20th century. The reason why we do not observe this catch-up is the divergence in the intensive margin that we have documented. The magnitude of the divergence in the intensive margin is sufficient to undo the catch-up induced by the converge in adoption lags. In fact, it sustains the gap in growth rates between Western countries and the rest of the world during the 20th century. Thus, in our simulation, the divergence in penetration rates accounts for the lack of convergence in income per capita between Western countries and the rest of the world during the 20th century.

2.5. CONCLUDING REMARKS

This chapter has explored three broad questions related to the diffusion of technology: (i) How can we measure technology diffusion and what are the broad patterns we have observed both across countries and over time on technology diffusion? (ii) What are the main drivers of technology diffusion? and (iii) What are the macroeconomic consequences of technology diffusion?

Given the strong linkages between these questions, addressing them in a unified way ensures internal consistency and exploits insights that have implications for more than one question. Despite the progress reported, we consider that still there is plenty of room to develop new studies that increase our understanding of these questions. Future work is likely to develop new comprehensive data sets that provide new measures of technology diffusion. Current and future data can be used to explore new implications of technology in new fields such as the development of institutions (e.g. political, educational, financial); the political consequences of technology; and the role of technology in wars and civil conflicts. Probably, where more and most obvious opportunities exist is in exploring the drivers of technology. The work summarized in this chapter, as well as other studies, have argued that technology is a key driver of cross-country income differences. However, it is still too early to conclusively assess what forces shape technology and how these forces operate over time. There is room both for exploration of new drivers and mechanisms of technology as well as for deeper investigations of mechanisms already discussed in the literature.

A. DESCRIPTION OF TECHNOLOGIES USED TO ESTIMATE DIFFUSION CURVES

The 25 particular technology measures, organized by broad category (transportation, communication, IT, industrial, agricultural, and medical), are described below.

Transportation

1. **Steam and motor ships:** Gross tonnage (above a minimum weight) of steam and motor ships in use at midyear. *Invention year:* 1788; the year the first (US) patent was issued for a steam boat design.
2. **Railways-Passengers:** Passenger journeys by railway in passenger-KM. *Invention year:* 1825; the year of the first regularly scheduled railroad service to carry both goods and passengers.
3. **Railways-Freight:** Metric tons of freight carried on railways (excluding livestock and passenger baggage). *Invention year:* 1825; same as passenger-railways.
4. **Cars:** Number of passenger cars (excluding tractors and similar vehicles) in use. *Invention year:* 1885; the year Gottlieb Daimler built the first vehicle powered by an internal combustion engine.
5. **Trucks:** Number of commercial vehicles, typically including buses and taxis (excluding tractors and similar vehicles), in use. *Invention year:* 1885; same as cars.
6. **Tractor:** Number of wheel and crawler tractors (excluding garden tractors) used in agriculture. *Invention year:* 1892; John Froelich invented and built the first gasoline/petrol-powered tractor.
7. **Aviation-Passengers:** Civil aviation passenger-KM traveled on scheduled services by companies registered in the country concerned. *Invention year:* 1903; the year the Wright brothers managed the first successful flight.
8. **Aviation-Freight:** Civil aviation ton-KM of cargo carried on scheduled services by companies registered in the country concerned. *Invention year:* 1903; same as aviation-passengers.

Communication and IT

1. **Telegraph:** Number of telegrams sent. *Invention year:* 1835; year of invention of telegraph by Samuel Morse at New York University.
2. **Mail:** Number of items mailed/received, with internal items counted once and cross-border items counted once for each country. *Invention year:* 1840; the first modern postage stamp, Penny Black, was released in Great Britain.
3. **Telephone:** Number of mainline telephone lines connecting a customer's equipment to the public switched telephone network. *Invention year:* 1876; year of invention of telephone by Alexander Graham Bell.
4. **Cellphone:** Number of users of portable cell phones. *Invention year:* 1973; first call from a portable cellphone.
5. **Personal computers:** Number of self-contained computers designed for use by one person. *Invention year:* 1973; first computer based on a microprocessor.

6. **Internet users:** Number of people with access to the worldwide network. *Invention year:* 1983; introduction of TCP/IP protocol.

Industrial

1. **Spindles:** Number of mule and ring spindles in place at year end. *Invention year:* 1779; spinning mule invented by Samuel Crompton.
2. **Synthetic Fiber:** Weight of synthetic (noncellulosic) fibers used in spindles. *Invention year:* 1924; invention of rayon.
3. **Steel:** Total tons of crude steel production (in metric tons). This measure includes steel produced using Bessemer and open hearth furnaces (OHF). *Invention year:* 1855; William Kelly receives the first patent for a steel-making process (pneumatic steel making).
4. **Electric Arc Furnaces:** Crude steel production (in metric tons) using electric arc furnaces. *Invention year:* 1907; invention of the electric arc furnace.
5. **Blast Oxygen Furnaces:** Crude steel production (in metric tons) in blast oxygen furnaces (a process that replaced Bessemer and OHF processes). *Invention year:* 1950; invention of blast oxygen furnace.
6. **Electricity:** Gross output of electric energy (inclusive of electricity consumed in power stations) in Kw-Hr. *Invention year:* 1882; first commercial power station on Pearl Street in New York City.

Agricultural

1. **Fertilizer:** Metric tons of fertilizer consumed. Aggregate of 25 individual types, corresponding to broadly ammonia and phosphates. *Invention year:* 1910; Haber-Bosch process to produce ammonia is patented in 1910.
2. **Harvester:** Number of self-propelled machines that reap and thresh in one operation. *Invention year:* 1912; the Holt Manufacturing Company of California produces a self-propelled harvester. Subsequently, a self-propelled machine that reaps and threshes in one operation appears.

Medical

1. **Kidney Transplant:** Number of kidney transplants performed. *Invention year:* 1954; Joseph E. Murray and his colleagues at Peter Bent Brigham Hospital in Boston performed the first successful kidney transplant.
2. **Liver Transplant:** Number of liver transplants performed. *Invention year:* 1963; Dr. Thomas Starzl performs the first successful liver transplant in the United States.
3. **Heart Transplant:** Number of heart transplants performed. *Invention year:* 1968; Adrian Kantrowitz performed the first pediatric heart transplant in the world on December 6, 1967 at Maimonides Hospital in New York.

B. ADDITIONAL TABLES

Table 2.10 Correlations and joint sample sizes

Technology	Invention	Year	Obs.	Mean	Median	st. dev.	IQR	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Real GDP		1950	138	149	128	93	128		0.93	0.82	0.75	0.77	0.75	0.75	0.77	0.71	0.71	0.75	0.61	0.7	0.63	0.71	0.63
Real GDP		1970	138	118	114	79	114	138		0.9	0.82	0.78	0.83	0.86	0.82	0.77	0.79	0.8	0.65	0.76	0.75	0.76	0.66
Real GDP		1990	164	113	94	82	119	138	138		0.94	0.74	0.8	0.89	0.86	0.8	0.81	0.75	0.66	0.77	0.73	0.68	0.61
Real GDP		2000	164	122	105	89	119	138	138	164		0.69	0.75	0.83	0.87	0.8	0.82	0.7	0.65	0.75	0.68	0.65	0.58
Electricity	1882	1950	99	46	48	20	29	96	96	97	97		0.94	0.89	0.81	0.82	0.72	0.9	0.66	0.75	0.8	0.88	0.81
Electricity	1882	1970	124	52	59	25	37	119	119	121	121	99		0.94	0.84	0.85	0.8	0.89	0.67	0.77	0.79	0.87	0.77
Electricity	1882	1990	127	58	57	28	46	120	120	126	126	97	121		0.86	0.86	0.82	0.85	0.68	0.8	0.82	0.8	0.7
Internet	1983	2002	123	9	8	5	6	101	101	120	120	81	99	103		0.89	0.86	0.81	0.74	0.81	0.74	0.78	0.69
PCs	1973	2002	127	18	20	7	6	108	108	124	124	86	107	111	110		0.8	0.89	0.76	0.84	0.75	0.83	0.74
Cellphones	1973	2002	145	8	8	6	11	118	118	142	142	93	116	120	122	127		0.76	0.77	0.79	0.69	0.72	0.58
Telephones	1876	1970	108	63	70	23	15	104	104	106	106	90	106	105	90	95	102		0.7	0.79	0.83	0.93	0.84
Avi. cargo	1903	1990	96	27	30	16	23	92	92	95	95	76	93	95	79	88	92	83		0.89	0.68	0.69	0.51
Avi. pass.	1903	1990	102	34	38	14	21	98	98	101	101	81	99	101	83	92	97	88	96		0.83	0.75	0.6
Trucks	1885	1990	98	59	67	18	26	97	97	98	98	78	96	96	81	84	94	86	78	83		0.82	0.66
Cars	1885	1990	127	66	74	20	14	106	106	127	127	81	100	105	102	103	119	92	81	86	97		0.84
Tractors	1894	1970	130	47	52	16	11	124	124	128	128	97	121	123	105	113	123	105	95	101	96	107	

Note: St.dev. is standard deviation. IQR is the interquartile range. In the correlations and joint samples sizes part, the numbers above the diagonal are the correlations and the numbers below the diagonal are the sample size on which the correlations are based.

Table 2.11 Quality of the estimates

							Detrended R^2		
Technology	**Invention Year**	**Total**	**Implausible**	**Imprecise**	**Precise**	**% Precise**	**$R^2 > 0$**	**Mean**	**SD**
Spindles	1779	34	3	0	31	91	26	0.57	0.25
Steam and motor ships	1788	61	16	0	45	74	45	0.84	0.10
Railways-freight	1825	85	39	0	46	54	46	0.78	0.15
Railways-passengers	1825	80	41	0	39	49	39	0.79	0.14
Telegraph	1835	62	19	0	43	69	30	0.58	0.25
Mail	1840	67	20	0	47	70	47	0.87	0.10
Steel (Bessemer, open hearth)	1855	52	11	0	41	79	41	0.74	0.17
Telephone	1876	139	84	0	55	40	54	0.88	0.15
Electricity	1882	134	52	0	82	61	82	0.91	0.12
Cars	1885	124	54	0	70	56	68	0.76	0.22
Trucks	1885	108	46	0	62	57	62	0.78	0.20
Tractor	1892	135	45	2	88	65	81	0.74	0.20
Aviation-freight	1903	93	50	0	43	46	43	0.88	0.10
Aviation-passengers	1903	96	52	0	44	46	44	0.90	0.07
Electric arc furnace	1907	75	22	0	53	71	46	0.62	0.24
Fertilizer	1910	132	39	4	89	67	76	0.62	0.25
Harvester	1912	104	32	2	70	67	59	0.68	0.24
Synthetic fiber	1924	49	1	0	48	98	45	0.69	0.24
Blast oxygen furnace	1950	49	10	0	39	80	30	0.62	0.29
Kidney transplant	1954	27	3	0	24	89	24	0.82	0.17
Liver transplant	1963	21	0	0	21	100	20	0.81	0.16
Heart surgery	1968	18	0	0	18	100	17	0.63	0.20
Cellphones	1973	84	2	0	82	98	82	0.91	0.07
PCs	1973	69	1	0	68	99	68	0.93	0.07
Internet	1983	59	1	0	58	98	58	0.96	0.04
All technologies		1957	643	8	1306	67	1233	0.79	0.21

REFERENCES

Acemoglu, D., Robinson, J.A., 2000. Why did the west extend the franchise? Democracy, inequality, and growth in historical perspective. The Quarterly Journal of Economics 115 (4), 1167–1199.

Acemoglu, D., Johnson, S., Robinson, J.A., 2002. Reversal of fortune: geography and institutions in the making of the modern world income distribution. The Quarterly Journal of Economics 117 (4), 1231–1294.

Acemoglu, D., Johnson, S., Robinson, J.A., 2005. Institutions as a fundamental cause of long-run growth. In: Aghion, P., Durlauf, S. (Eds.), Handbook of Economic Growth, vol. 1, 6. Elsevier, pp. 385–472.

Aghion, P., Howitt, P., 1992. A model of growth through creative destruction. Econometrica 60 (2), 323–351.

Alexopoulos, M., 2011. Read all about it!! What happens following a technology shock? American Economic Review 101 (4), 1144–1179.

Arellano, M., Bond, S., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Review of Economic Studies 58 (2), 277–297.

Arrow, K.J., 1962. The economic implications of learning by doing. The Review of Economic Studies 29 (3), 155–173.

Arthur, W.B., 1989. Competing technologies, increasing returns, and lock-in by historical events. Economic Journal 99, 116–131.

Ashraf, Q., Galor, O., 2013. The "out of africa" hypothesis, human genetic diversity, and comparative economic development. American Economic Review 103 (1), 1–46.

Bandiera, O., Rasul, I., 2006. Social networks and technology adoption in northern mozambique. Economic Journal 116 (514), 869–902.

Banerjee, A., 1992. A simple model of herd behavior. Quarterly Journal of Economics 107 (3), 797–817.

Barlevy, G., 2007. On the cyclicality of research and development. American Economic Review 97 (4), 1131–1164.

Barro, R.J., Sala-i-Martin, X., 2003. Economic Growth, second ed. The MIT Press.

Basu, S., 1996. Procyclical productivity: increasing returns or cyclical utilization? The Quarterly Journal of Economics 111 (3), 719–751.

Basu, S., Fernald, J.G., 1997. Returns to scale in U.S. production: estimates and implications. Journal of Political Economy 105 (2), 249–283.

Benhabib, J., Spiegel, M.M., 1994. The role of human capital in economic development evidence from aggregate cross-country data. Journal of Monetary Economics 34 (2), 143–173.

Benhabib, J., Spiegel, M.M., 2005. Human capital and technology diffusion. In: Aghion, P., Durlauf, S. (Eds.), Handbook of Economic Growth, vol. 1, 13. Elsevier, pp. 935–966.

Besley, T., Case, A., 1995. Does electoral accountability affect economic policy choices? Evidence from gubernatorial term limits. The Quarterly Journal of Economics 110 (3), 769–798.

Bockstette, V., Chanda, A., Putterman, L., 2002. States and markets: the advantage of an early start. Journal of Economic Growth 7 (4), 347–369.

Burnside, C., Eichenbaum, M., Rebelo, S., 1995. Capital utilization and returns to scale. In: NBER Macroeconomics Annual 1995, vol. 10, NBER Chapters. National Bureau of Economic Research, Inc., pp. 67–124.

Caballero, R.J., Hoshi, T., Kashyap, A.K., 2008. Zombie lending and depressed restructuring in Japan. American Economic Review 98 (5), 1943–1977.

Caselli, F., 1999. Technological revolutions. American Economic Review 89 (1), 78–102.

Caselli, F., Coleman, W.J., 2001. Cross-country technology diffusion: the case of computers. American Economic Review 91 (2), 328–335.

Chari, V.V., Hopenhayn, H., 1991. Vintage human capital, growth, and the diffusion of new technology. Journal of Political Economy 99 (6), 1142–1165.

Clark, G., 1987. Why isn't the whole world developed: lessons from the cotton mills. The Journal of Economic History 47 (1), 141–173.

Coe, D.T., Helpman, E., 1995. International r&d spillovers. European Economic Review 39 (5), 859–887.

Cogley, T., Nason, J.M., 1995. Output dynamics in real-business-cycle models. American Economic Review 85 (3), 492–511.

Comin, D., 2009. On the integration of growth and business cycles. Empirica 36 (2).

Comin, D., 2011. An exploration of the japanese slowdown during the 1990s. In: Koichi Hamada, A.K.K., Weinstein, D.E. (Eds.), Japans Bubble, Deflation, and Long-term Stagnation, vol. 1. MIT Press, pp. 375–398.

Comin, D., Gertler, M., 2006. Medium-term business cycles. The American Economic Review 96 (3), 523–551.

Comin, D., Hobijn, B., 2004. Cross-country technology adoption: making theory face the facts. Journal of Monetary Economics 51, 39–83.

Comin, D., Hobijn, B., 2007. Implementing Technology. Technical Report.

Comin, D., Hobijn, B., 2009a. The CHAT Dataset. Working Paper 15319, National Bureau of Economic Research.

Comin, D., Hobijn, B., 2009b. Lobbies and technology diffusion. Review of Economics and Statistics 91 (2), 229–244.

Comin, D., Hobijn, B., 2010. An exploration of technology diffusion. American Economic Review 100 (5), 2031–2059.

Comin, D., Mestieri, M. 2010. An Intensive Exploration of Technology Adoption. NBER Working Paper 16379.

Comin, D., Mestieri, M., 2013. If Technology Has Arrived Everywhere, why has Income Diverged? Working Paper 19010. National Bureau of Economic Research.

Comin, D., Hobijn, B., Rovito, E., 2006. Five Facts You Need to Know About Technology Diffusion. NBER Working Papers 11928, National Bureau of Economic Research, Inc.

Comin, D., Hobijn, B., Rovito, E., 2008a. A new approach to measuring technology with an application to the shape of the diffusion curves. The Journal of Technology Transfer 33 (2), 187–207.

Comin, D., Hobijn, B., Rovito, E., 2008b. Technology usage lags. Journal of Economic Growth 13 (4), 237–256.

Comin, D., Gertler, M., Santacreu, A.M., 2009a. Technology Innovation and Diffusion as Sources of Output and Asset Price Fluctuations. NBER Working Papers 15029, National Bureau of Economic Research, Inc.

Comin, D., Loayza, N., Pasha, F., Serven, L., 2009b. Medium Term Business Cycles in Developing Countries. NBER Working Papers 15428, National Bureau of Economic Research, Inc.

Comin, D., Easterly, W., Gong, E., 2010. Was the wealth of nations determined in 1000 BC? American Economic Journal: Macroeconomics 2 (3), 65–97.

Comin, D., Trumbull, J.G., Yang, K., 2012. Fraunhofer: Innovation in Germany. Technical Report 711–02, Harvard Business School Case.

Comin, D., Dmitriev, M., Rossi-Hansberg, E., 2013. The Spatial Diffusion of Technology. Technical Report.

Conley, T.G., Udry, C.R., 2010. Learning about a new technology: Pineapple in Ghana. American Economic Review 100 (1), 35–69.

Cox, W.M., Alm, R., 1996. The economy at light speed: technology and growth in the information age and beyond. Annual Report, pp. 2–17.

Crafts, N. 1997. Endogenous growth: lessons for and from economic history. In: Kreps, D., Wallis, K.F. (Eds.), Advances in Economics and Econometrics: Theory and Applications, vol. 2. CUP, pp. 38–78.

Davies, S., 1979. The Diffusion of Process Innovations. Cambridge University Press.

Diamond, J. 1999. Guns, Germs, and Steel: The Fates of Human Societies. Norton paperback, W W Norton & Company Incorporated.

Dixon, R., 1980. Hybrid corn revisited. Econometrica 48, 145–146.

Dupas, P., 2009. What matters (and what does not) in households' decision to invest in malaria prevention? American Economic Review 99 (2), 224–230.

Erosa, A., Koreshkova, T., Restuccia, D., 2010. How important is human capital? A quantitative theory assessment of world income inequality. Review of Economic Studies 77 (4), 1421–1449.

Evenson, R.E., Gollin, D., 2003. Assessing the impact of the green revolution, 1960 to 2000. Science 300 (5620), 758–762.

Fatas, A., 2000. Endogenous growth and stochastic trends. Journal of Monetary Economics 45 (1), 107–128.

Ferejohn, J., 1986. Incumbent performance and electoral control. Public Choice 50 (1), 5–25.

Feyrer, J. 2009a. Distance, Trade, and Income The 1967 to 1975 Closing of the Suez Canal as a Natural Experiment. NBER Working Papers 15557, National Bureau of Economic Research, Inc.

Feyrer, J. 2009b. Trade and Income – Exploiting Time Series in Geography. NBER Working Papers 14910, National Bureau of Economic Research, Inc.

Foster, A.D., Rosenzweig, M.R., 1995. Learning by doing and learning from others: human capital and technical change in agriculture. Journal of Political Economy 103 (6), 1176–1209.

Foster, A.D., Rosenzweig, M.R., 2010. Microeconomics of technology adoption. Annual Review of Economics 2 (1), 395–424.

Frankel, J.A., Romer, D.H., 1999. Does trade cause growth? American Economic Review 89 (3), 379–399.

Galí, J., 1999. Technology, employment, and the business cycle: do technology shocks explain aggregate fluctuations? American Economic Review 89 (1), 249–271.

Geroski, P., 2000. Models of technology diffusion. Research Policy 29, 603–625.

Gort, M., Klepper, S., 1982. Time paths in the diffusion of product innovations. Economic Journal 92 (367), 630–653.

Griliches, Z., 1957. Hybrid corn: an exploration in the economics of technological change. Econometrica 25 (4), 501–522.

Griliches, Z., 1990. Patent statistics as economic indicators: a survey. Journal of Economic Literature 28 (4), 1661–1707.

Guiso, L., Sapienza, P., Zingales, L., 2008. Alfred marshall lecture social capital as good culture. Journal of the European Economic Association 6 (2–3), 295–320.

Hannan, M.T., Freeman, J., 1989. Organizations and social structure. Organizational Ecology. Harvard University Press, Cambridge, pp. 3–27.

James S. Coleman, E.K., Menzel, H., 1966. Medical Innovation: A Diffusion Study. Bobbs-Merrill Co.

Johansen, L. 1959. Substitution versus fixed production coefficients in the theory of economic growth: a synthesis. Econometrica 27 (2), 157–176.

Karshenas, M., Stoneman, P.L., 1993. Rank, stock, order, and epidemic effects in the diffusion of new process technologies: an empirical model. The RAND Journal of Economics, pp. 503–528.

Keller, W., 2004. International technology diffusion. Journal of Economic Literature 42 (3), 752–782.

Kunicová, J., Rose-Ackerman, S., 2005. Electoral rules and constitutional structures as constraints on corruption. British Journal of Political Science 35, 573–606.

Kydland, F.E., Prescott, E.C., 1982. Time to build and aggregate fluctuations. Econometrica 50 (6), 1345–1370.

Levin, S., Levin, S., Meisel, J., 1987. A dynamic analysis of the adoption of a new technology: the case of optical scanners. Review of Economics and Statistics 69, 12–17.

Lucas, R.E., Buera, F.J., Alvarez, F., 2011. Trade and Idea Flows. 2011 Meeting Papers 984, Society for Economic Dynamics.

Maddison, A., 2004. Contours of the World Economy and the Art of Macro-Measurement 1500–2001. Ruggles Lecture, Cork, Ireland.

Mansfield, E., 1961. Technical change and the rate of imitation. Econometrica 29 (4), 741–766.

Mansfield, E., 1963. The speed of response of firms to new technologies. Quarterly Journal of Economics 29 (77), 290–311.

Mansfield, E., 1968. Industrial Research and Technological Innovation. W.W, Norton, New York.

Metcalfe, J.S., 1981. Impulse and diffusion in the study of technical change. Futures 13 (5), 347–359.

Metcalfe, J.S., 1998. Evolutionary Economics and Creative Destruction. Routledge.

Mokyr, J., 1990. The Lever of Riches: Technological Creativity and Economic Progress. Oxford paperbacks, Oxford University Press, USA.

Myerson, R.B., 1993. Effectiveness of electoral systems for reducing government corruption: a game-theoretic analysis. Games and Economic Behavior 5 (1), 118–132.

Nelson, R.R., Phelps, E.S., 1966. Investment in humans, technological diffusion and economic growth. American Economic Review 56 (1/2), 69–75.

Norbin, S., 1993. The relation between price and marginal cost in U.S. industry: a contradiction. Journal of Political Economy 101 (6), 1149–1164.

Olson, M., 1982. The Rise and Decline of Nations: Economic Growth, Stagflation and Social Rigidities. Yale University.

Peregrine, P., 2003. Atlas of cultural evolution. In: Gray J.P. (Ed.), World Cultures 14 (1), 2–88.

Perla, J., Tonetti, C., Waugh, M.E., 2012. Equilibrium Technology Diffusion, Trade and Growth. Technical Report, Mimeo.

Persson, T., Roland, G., Tabellini, G., 2000. Comparative politics and public finance. Journal of Political Economy 108 (6), 1121–1161.

Persson, T., Tabellini, G., Trebbi, F., 2003. Electoral rules and corruption. Journal of the European Economic Association 1 (4), 958–989.

Porter, M.E., 1998. Clusters and the new economics of competition. Harvard Business Review 76 (6), 77–90.

Putterman, L., Weil, D.N., 2010. Post-1500 population flows and the long-run determinants of economic growth and inequality. The Quarterly Journal of Economics 125 (4), 1627–1682.

Riddell, W.C., Song, X., 2012. The Role of Education in Technology Use and Adoption: Evidence from the Canadian Workplace and Employee Survey. IZA Discussion Papers 6377, Institute for the Study of Labor (IZA).

Romeo, A., 1975. Inter-industry and inter-firm differences in the rate of diffusion of an innovation. Review of Economics and Statistics 57, 311–316.

Romer, P.M., 1990. Endogenous technological change. Journal of Political Economy 98 (5), S71–102.

Rose, N., Joskow, P., 1990. The diffusion of new technologies: evidence from the electric utility industry. The Rand Journal of Economics 21, 354–373.

Sachs, J., Warner, A., 1995. Economic Reform and the Progress of Global Integration. Harvard Institute of Economic Research Working Papers 1733, Harvard - Institute of Economic Research.

Santos, M.S., Iraola, M.A., 2010. Long-Term Asset Price Volatility and Macroeconomic Fluctuations. 2010 Meeting Papers 374, Society for Economic Dynamics.

Scherer, F., 1981. Using linked patent and R&D data to measure inter-industry technology flows. In: Griliches, Z. (Ed.), R&D, Patents, and Productivity. University of Chicago Press, pp. 417–464.

Schmookler, J., 1966. Invention and Economic Growth. Harvard University Press.

Seshadri, A., Manuelli, R., 2005. Human Capital and the Wealth of Nations. 2005 Meeting Papers 56, Society for Economic Dynamics.

Skinner, J., Staiger, D., 2007. Technology adoption from hybrid corn to beta-blockers. In: Hard-to-Measure Goods and Services: Essays in Honor of Zvi Griliches, NBER Chapters, National Bureau of Economic Research, Inc., pp. 545–570.

Solow, R.M., 1956. A contribution to the theory of economic growth. The Quarterly Journal of Economics 70 (1), 65–94.

Solow, R.M., 1960. Investment and technical progress. In: Arrow, S.K.K., Suppes, P. (Eds.), Mathematical Methods in the Social Sciences, vol. 1. Stanford University Press, pp. 89–104.

Spolaore, E., Wacziarg, R., 2009. The diffusion of development. The Quarterly Journal of Economics 124 (2), 469–529.

Stoneman, P., 1981. Intra-firm diffusion, bayesian learning and profitability. Economic Journal 91 (362), 375–388.

Stoneman, P., 1983. The Economic Analysis of Technological Change. Oxford University Press Oxford.

Stoneman, P., 1987. The Economic Analysis of Technology Policy. Clarendon Press Oxford.

Stoneman, P. et al., 1995. Handbook of the Economics of Innovation and Technological Change. Blackwell Oxford.

Tabellini, G., 2007. Institutions and Culture. Working Papers 330, IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University.

Thirtle, C., Ruttan, V., 1987. The Role of Demand and Supply in the Generation and Diffusion of Technical Change. Fundamentals of Pure and Applied Economics Series, Harwood Academic Publications.

Vickery, G., Northcott, J., 1995. Diffusion of microelectronics and advanced manufacturing technology: a review of national surveys. Economics of Innovation and New Technology 3 (3–4), 253–276.

CHAPTER THREE

Health and Economic Growth

David N. Weil
Brown University and NBER, USA

Abstract

This chapter examines the relationship between health and economic growth. Across countries, income per capita is highly correlated with health, as measured by life expectancy or a number of other indicators. Within countries, there is also a correlation between people's health and income. Finally, over time, the historical evolution of cross-country health differences has largely paralleled the evolution of income differences, with the exception that in the last half century the convergence of health has been much faster than the convergence of income. How are health and income related? Theoretically, there is good reason to believe that causality runs in both directions. Healthier individuals are more productive, learn more in school, and, because they live longer, face enhanced incentives to accumulate human capital. Similarly, higher income for individuals or countries improves health in a variety of ways, ranging from better nutrition to construction of public health infrastructure. Empirically, there is evidence for both of these causal channels being operative, but the magnitude of the effects is limited, at least as they apply to cross-sectional differences among countries or individuals. Apparently, other factors that simultaneously raise income and improve health outcomes, such as institutional quality (for countries) and human capital (for individuals), are responsible for a good deal of the observed health–income correlation. The final section of the chapter discusses measures of aggregate welfare that combine consumption and life expectancy.

Keywords

Economic Growth, Health, Mortality, Productivity, Disease

JEL Classification Codes

I10, J17, N30, O11, O40

3.1. INTRODUCTION

The largest part of this literature, and the part on which I focus most extensively, examines the effect of health on economic growth. Does making a population healthier make it richer? Over what time horizon and through which channels? What is the magnitude of health's impact on income, and how much of income variation among countries is explained by variation in health?

The second topic on which I focus is causality running in the other direction, from income to health. Humanity has experienced great improvements in health over the last two centuries, roughly contemporaneously with the period of steady income growth

Handbook of Economic Growth, Volume 2B
ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53540-5.00003-3

that followed the Industrial Revolution. But the causes of this health improvement are not transparent, particularly the extent to which better health is attributable to income growth per se, to changes in health technology, and to changes in the institutions that deliver health services. Most notably, over the past century, the cross-sectional relationship between income and health has changed significantly, indicating that the "technology" of health, and perhaps the price of health, have changed. I discuss the nature of this health technology.

The final large topic I address is how to comprehend health improvements in a growth framework focused on utility, rather than income. An important difference between health and many of the other determinants of income that are considered in the growth literature is that health is primarily valued in its own right, rather than for its effects on output. This has led to a certain politicization of the health–growth literature, in which the view that health is an important determinant of economic growth sometimes seems to be embraced in part because the widespread acceptance of such a view would lead to greater spending on health, which is viewed as a good thing in and of itself. For example, the WHO Commission on Macroeconomics and Health (Sachs, 2001) writes

> *Improving the health and longevity of the poor is an end in itself, a fundamental goal of economic development. But it is also a means to achieving the other development goals relating to poverty reduction. The linkages of health to poverty reduction and to long-term economic growth are powerful, much stronger than is generally understood. The burden of disease in some low-income regions, especially sub-Saharan Africa, stands as a stark barrier to economic growth and therefore must be addressed frontally and centrally in any comprehensive development strategy.*

The rest of this article is organized as follows. Section 3.2 presents the facts regarding the relationship between income and health, both between and within countries. Section 3.3 presents a very simple theoretical framework for thinking about the simultaneous determination of health and income, and then uses this framework to highlight some of the key issues that will inform the rest of the article. Section 3.4 looks at the role of income and other factors in explaining improvements in health, taking both a historical approach (focused on the currently wealthy countries), and looking at differences between rich and poor countries today. Section 3.5 focuses on causality running from health improvements to income growth. I lay out several channels that theoretically could lead to such causality, discuss available evidence, and address the problem of quantifying the overall effect. In this section, I also discuss empirical work that has assessed the overall effect of particular episodes of health improvement historically. Section 3.6 presents a framework in which one can assess health as an aspect of economic growth, in practice producing an income-equivalent measure of the value of health improvements. I also discuss how this framework can be parameterized using data on the revealed value of living vs. dying, and some of the problems this approach raises. Section 3.7 concludes.

3.2. FACTS

I start by laying out the facts regarding the relationship between income and health, cross-sectionally among countries, cross-sectionally within countries, and over time. I use a number of indicators of health, because health is by its nature a multidimensional concept. One natural and widely used measure of health is the probability of death, as captured by life expectancy or the infant mortality rate. But variations in death probabilities are far from fully informative about the health status of the living. Some conditions that cause premature death may leave little health impact on those who survive, and may even raise the health of the living via selection. Other conditions that cause high mortality (for example, smallpox) also leave a great deal of physical damage among survivors. Similarly, "improvements in health" can take the form of reduced probabilities of dying, better health among those who are alive, or both. Even within the category of health of the living, there are many different dimensions. Some conditions may impact a person's physical but not mental functioning, or vice versa. Similarly, some conditions may have a larger relative effect on quality of life or utility on the one hand, compared to economic productivity, on the other. And of course, the economic impact of a specific condition will vary with the structure of the economy: the relative wage of brawn relative to brains has declined as countries have developed, meaning that the relative productivity decrement from physical vs. mental disability has declined as well (Galor and Weil, 1996).

3.2.1 Cross-Section

3.2.1.1 Cross-Country Data

Life Expectancy

Life expectancy at birth is the number of years that a newborn baby would be expected to live, using current age-specific survival rates. Life expectancy is thus a scalar summary measure of the underlying vector of age-specific survival rates, which demographers call the life table. (Age-specific survival is not actually measured in many instances, and the full set of life table values is imputed from observation of only a few elements, such as infant mortality). In principle, a given life expectancy at birth is consistent with many different possible shapes of the survival function; in practice, there are empirical regularities regarding how the survival function changes shape as life expectancy rises. Demographers construct model life tables that embody these regularities (sometimes with adjustments for the constellations of diseases found in different locations or historical eras). Figure 3.1 shows the probability of survival to different ages for a family of model life tables for a variety of life expectancies.[1] The figure shows that infant and child survival is the most important component of increased survival associated with increases in life expectancy

[1] Li and Gerland (2011). This is the general table. Data are for females.

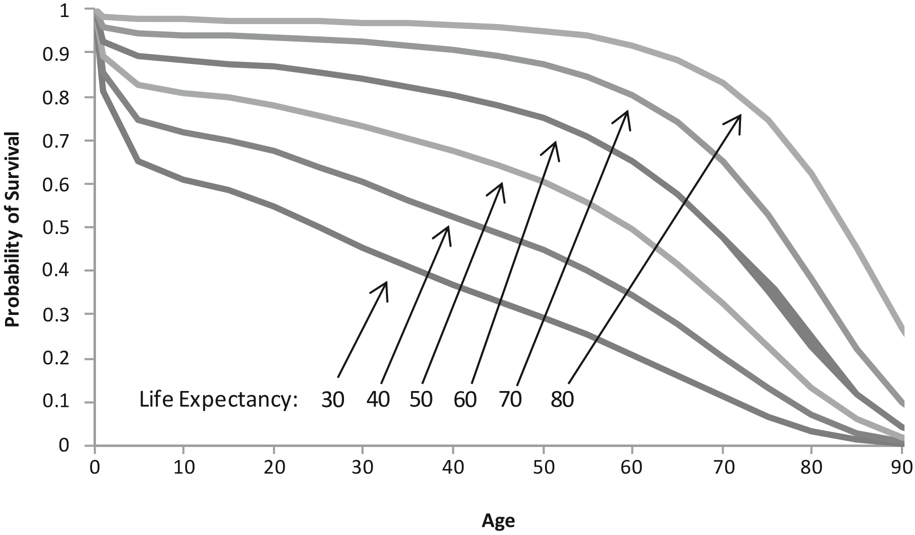


Figure 3.1 Model life tables.

(from a low level). This pattern is close to universal in examining both cross-sectional differences and time trends in life expectancy (with the effect of HIV in Africa today being an exception). An implication of this regularity is that differences in life expectancy at ages other than birth tend to be far smaller than differences in life expectancy at birth. (Another implication of the typical pattern of change as pointed out by Peltzman (2009), is that as life expectancy rises, inequality in experienced lifetimes declines. In the US, the Gini coefficient for lifetimes declined from roughly 0.50 to 0.12 between 1850 and 2000.)

Figure 3.2 shows the cross-sectional relationship between the log of income per capita and life expectancy, using data from 2009. There is obviously a very strong relationship between the two. The R-squared from a simple regression of life expectancy on the log of income per capita is 0.67. There are no major outliers lying above the regression line, and those lying below are characterized by high rates of HIV (South Africa and Botswana), war (Afghanistan), or are oil producers that only recently experienced enormous increases in income (Gabon and Equatorial Guinea).

Years Lost to Disability

Life expectancy is often used as a measure of the health impact of the disease environment because premature death is the most significant (and certainly the most observable) impact of disease. But death is not the only impact of disease. In the scheme of the World Health Organization, Disability Adjusted Life Years (DALYs) lost as a result of a disease or injury are the sum of years lost to premature death (Years of Life Lost, YLLs), and healthy life

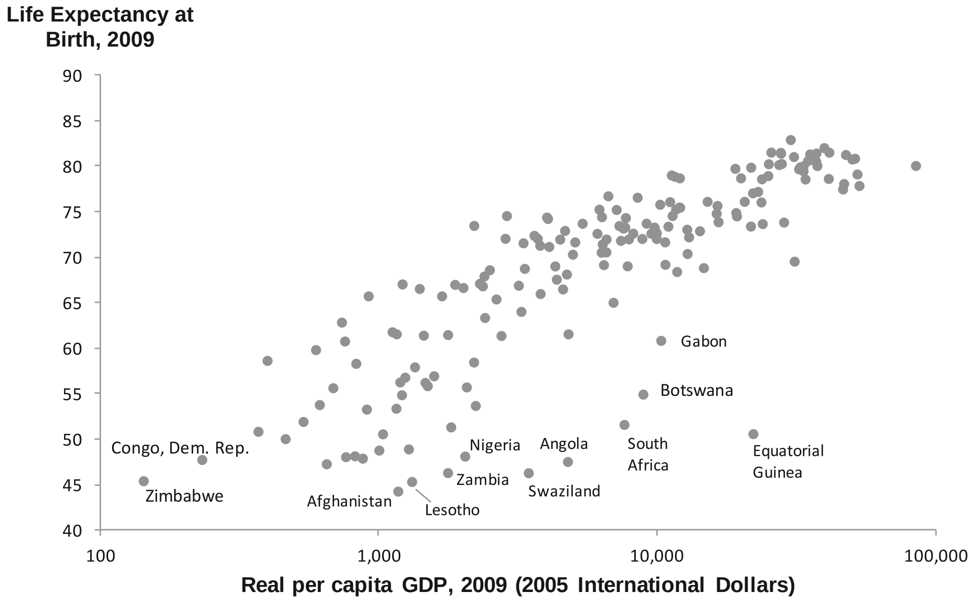


Figure 3.2 Income and life expectancy across countries.

year equivalents lost as a result of being in a state of poor health or disability.[2] The latter are called Years lost to Disability (YLDs) and can thus be thought of as a measure of the non-death costs of disease. According to Mathers et al. (2008), 60% of DALYs lost in 2004 were due to premature mortality, with the other 40% due to non-fatal health outcomes.

Figure 3.3 shows the cross-sectional relationship between life expectancy at birth and YLDs, looking across WHO country groupings. There is clearly a very tight fit, establishing that health as measured by deaths and health as measured by sickness among the living, vary in tandem. However, it is worth noting that, at least in this data, the gap in life expectancy understates the gap in the health of the living: the poorest regions in the world have roughly twice the rate of YLDs as the richest, while the gap in life expectancy at birth is closer to a factor of 1.7.

Other Health Measures

Beyond summary measures such as life expectancy and years lost to disability, one can look at individual indicators of health. Figure 3.4 shows data from Shastry and Weil (2003) on the cross-country relationship between income and the fraction of women

[2] Equivalence between healthy life years and years under different states of poor health or disability is established using a person trade-off in which experts compare the utility of living with different conditions to the utility of living fewer years disability-free.

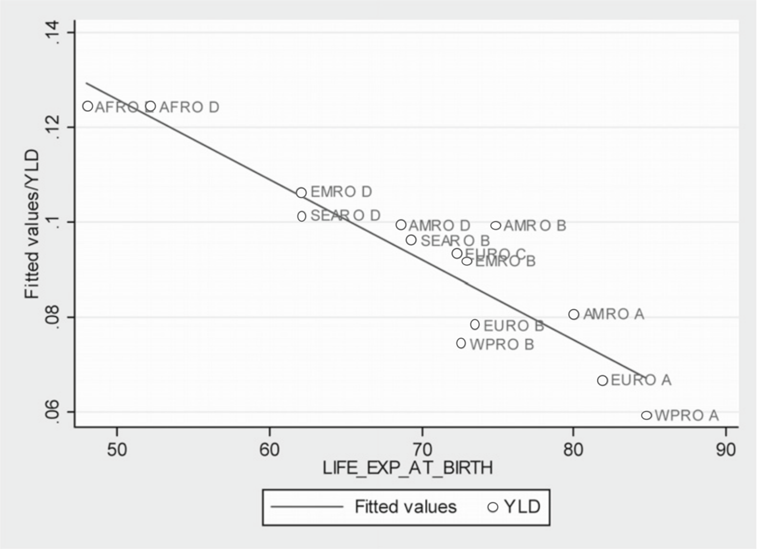


Figure 3.3 Life expectancy and years lost to disability.

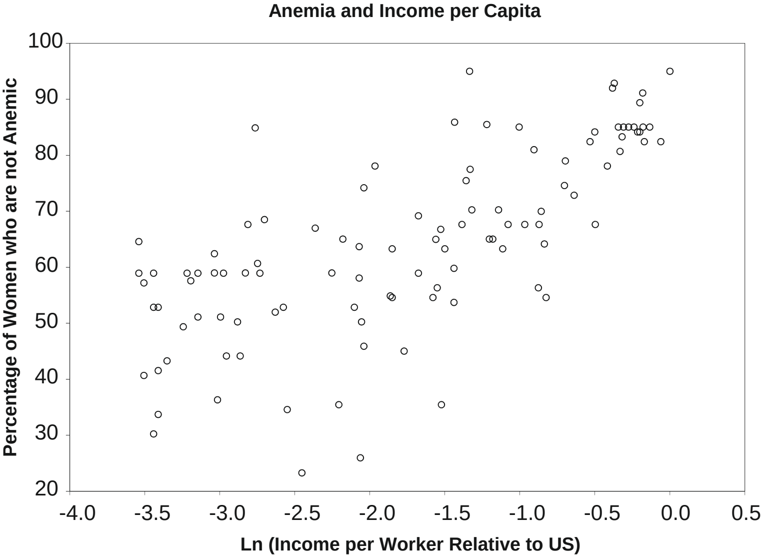


Figure 3.4 Income per capita and anemia.

who are not anemic. Anemia is defined to be a low level of hemoglobin in the blood, resulting in reduced transportation of oxygen to the tissues in the body. Iron deficiency anemia, the most common form of this health condition, results from either insufficient dietary intake of iron and/or presence of diseases such as malaria (which attacks red

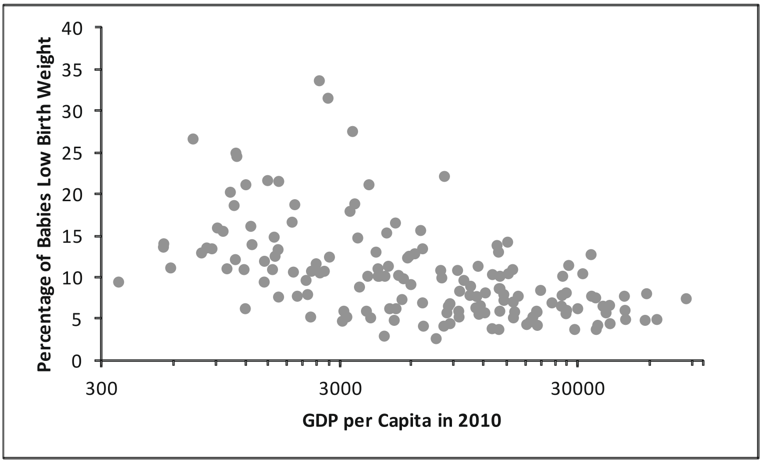


Figure 3.5 GDP and low birth weight.

blood cells) and helminth infections (which lead to intestinal bleeding). Anemia has negative effects on fetal and child growth as well as cognitive function of students, and increases morbidity and mortality among people of all ages. Anemia also affects a person's stamina, making him or her tire more easily, thus causing workers to be less productive (Thomas and Frankenberg, 2002). Although anemia is clearly only one dimension of health (and is far more prevalent among women than among men), it is of particular interest because there exist direct measures of its effect on productivity, which are discussed in Section 3.5.

Figure 3.5 shows the fraction of babies that are classified as low birth weight for a cross-section of countries.[3] As discussed further below, birth weight is a useful summary measure of health and nutritional insults in utero, a crucial period for human development. Low birth weight is correlated with high blood pressure and many other health conditions, as well as reduced cognitive development.[4] Behrman and Rosenzweig (2004) and Black et al. (2007) show that differences in birth weight among identical twins translate into differences in education and wages among adults.

Other indicators of health that one can look at in cross-section include age of menarche (the onset of menstruation, Weil (2007)), height (Subramanian et al. 2011), and body mass index (BMI).

[3] Data on GDP and low birth weight are both from the WDI database. Low birth weight is for the most recent year available in the range 2000–2010.

[4] Almond et al. (2005) point out that it may not be low birth weight per se that causes poor health outcomes, but rather other inputs to health that cause both low birth weight and poor health outcomes. Thus, policies that directly target a reduction in low birth weight will not necessarily have the impact on other health measures that would be predicted by the correlation between health outcomes and low birth weight.

3.2.1.2 *Within-Country Covariation of Health with Income*

The relationship between income and health or life expectancy that is observed across countries is echoed in within-country data. Deaton (2003), using US data, calculates that the probability of a 50-year-old man dying within the next 9 years was more than twice as high for men in families with income below $10,000 as in households with income above $60,000 (1980 dollars). Income has more effect on health outcomes at the lower end of the income distribution. Deaton and Paxson (2001) find that in general, higher income and education both reduce mortality within the US, although there is evidence that short-run increases in income may raise mortality for males. More specifically, using data from the National Longitudinal Mortality Study (Table 4.4), they find that the elasticity of mortality risk with respect to income per adult equivalent is −0.35 for men and −0.26 for women (when education is not controlled for), and the semi-elasticity of mortality with respect to years of education is −0.037 for men and −0.038 for women (when income is not controlled for). The effects when both income and education are controlled for are inconsistent between men and women, and more generally Deaton and Paxson argue that it is hard to see in their data whether the effect of education operates solely through income or has an independent effect as well. Case et al. (2002) show that there is a significant gradient of child health with respect to income in the United States, and that the gradient grows steeper as children age, reflecting the accumulation of adverse health impacts over children's lives.

Gwatkin et al. (2007) present data on a large number of health indicators broken down by quintile of wealth (rather than income) for 56 developing countries. The underlying data come from the Demographic and Health Surveys (DHS). Table 3.1 shows several representative indicators.

Turning to measures of adult health beyond mortality, Floud et al. (2011) show a strong relationship between economic outcomes, on the one hand, and markers of nutritional status, on the other. Earnings increase with height for both Union Army veterans in the 19th century and for modern American males. Similarly, in both time periods, risk of poverty and non-labor force participation rise as body mass index falls below a cutoff of approximately 24. In developing countries, the relationship between height and income is more steeply sloped than in rich countries. In five different samples from the United States and United Kingdom, Case and Paxson (2010) estimate semi-elasticities of wages

Table 3.1 Health indicators by wealth quintile in developing countries

Wealth quintile	**1**	**2**	**3**	**4**	**5**
Infant Mortality	85.0	80.1	75.6	65.1	50.1
Under-5 Mortality	135.4	129.0	120.2	102.5	73.5
Prevalence of Diarrhea in Children	19.0	18.2	17.4	16.5	13.9
Moderate Stunting in Children	21.8	19.6	18.3	16.2	12.1
Moderate Underweight in Children	20.5	18.9	17.0	14.8	11.1

with respect to adult height (controlling only for ethnicity) of between 0.48 and 1.1% per centimeter for men and 0.26 and 1.1% for women. In Mexican data, Vogl (2012) finds a semi-elasticity of wage with respect to height of 2.5% per centimeter. In the Indonesian data for 1997, Thomas and Frankenberg (2002) find that a 1% increase in height is associated with a 5% increase in wages (implying a semi-elasticity of roughly 3.1% per centimeter). In the same data, the elasticity of wage with respect to BMI, not conditioning on other factors, is 2. Unlike the US data, the relationship between BMI and log wage is linear throughout the range of observed BMI.[5]

Conditioning on other determinants of wages does not eliminate the effect of height. Schultz (2002) regresses log wages on height, controlling for education, experience, and rural residence. An extra centimeter raises wages by 1.5% for men and 1.7% for women in Ghana; 1.4% for men and 1.7% for women in Brazil; but only 0.45% for men and 0.31% for women in the United States. Height is believed to be related to economic outcomes through one or more channels: because taller people are healthier and stronger, and these characteristics are rewarded in the labor market; because adult height is affected by childhood inputs that also contribute to cognitive ability, which is rewarded; or because height affects self-esteem or social status, which in turn affect wages (Currie and Vogl, 2013).[6]

3.2.2 Historical

Economists studying long-run growth have established a set of rough facts describing the historical evolution of income per capita and the world distribution of income. See Lucas (2000) for a summary. Although health data are just as imprecise as those for income, and health is itself, as mentioned above, a fundamentally multidimensional concept, it is nonetheless the case that in rough terms the evolution of health looks very similar to that for income. In particular:

1. In the period prior to the Industrial Revolution, there was little or no long-run change in countries' levels of health, though with considerable short-run variations.
2. Prior to the Industrial Revolution, cross-sectional differences among countries were relatively small.
3. The 19th century saw a takeoff of health status in Europe and its offshoots, with little change elsewhere, leading to growing health inequality among countries.
4. Starting in roughly the middle of the 20th century, health improvements in trailing countries began to exceed those in the leaders, resulting in a narrowing of the cross-sectional health variance.

[5] Thomas and Frankenberg find that a good deal of the predictive power of BMI for wages goes away once they control for height and education. This shows that much of the predictive effect of BMI on wages results from the endogeneity of BMI, rather than a direct effect of health on income.

[6] Baten et al. (forthcoming) show a negative correlation between height and innumeracy, as measured by age-heaping in survey data, for a variety of historical data sets.

It is with respect to the last of these four points that the analogy between the evolution of health and that of income breaks down most significantly. While growth economists question why the convergence of income per capita has been so slow (to the extent that it has happened at all), convergence in health measures has been quite rapid.

3.2.2.1 Life Expectancy

At the end of the 18th century, Malthus wrote "With regard to the duration of human life, there does not appear to have existed from the earliest ages of the world to the present moment the smallest permanent symptom or indication of increasing prolongation." He was basically right. Prior to the 19th century, data on life expectancy come from diverse and sometimes inconsistent sources, including family reconstructions, census records, and temple and parish records. Though imprecise, the picture painted by these data is of little or no improvement in life expectancy over a span of millenia, as well as some cross-sectional variation, with Europe (and Japan) being slightly healthier than the rest of the world. Maddison (2001) reports life expectancy in Roman Egypt at 24 years, the same as the value for England in the 14th century. By the middle of the 18th century, life expectancy was 35 years in England, but still 25 years in France. In Japan at the same time, life expectancy was in the early 30s.[7]

De la Croix and Licandro (2012) conduct an examination of long-run mortality trends looking at biographies of 300,000 famous individuals born starting in the 24th century BCE. By construction, their data focus only on adults (who lived long enough to become prominent) and on regions that were sufficiently developed that written biographical records survive, primarily Europe. De la Croix and Licandro date the beginning of mortality improvements to the cohort born 1640–1649, more than a century earlier than most other sources. The mean lifespan of famous people was 60 years in the four millenia prior to that; by the time of the cohort born in 1869, it had risen to 69.

Around 1800, life expectancy started to increase, first in Europe and its offshoots, spreading to the rest of the world by the middle of the 20th century. Average life

[7] Historical data for regions outside of Europe are extremely sparse. Acemoglu and Johnson (2007, Appendix C) provide an extensive and well-documented compilation of estimates for developing countries in the first half of the 20th century (these data underlie Figures 3.11 and 3.12 below). Riley (2005) estimates that prior to the "health transition" (he uses a different definition than Acemoglu and Johnson) that began in Africa in the 1920s, life expectancy at birth averaged 26.4 years. In Asia, life expectancy prior to the health transition, which started there between 1870 and 1890, was 27.4. In Europe, the transition started in the 1770s, and prior to it life expectancy was 34.3. Riley comments that available estimates of African mortality prior to the health transition all come from European colonies in Africa. There is a reasonable basis for thinking that life expectancy may have been higher prior to colonization, the arrival of Arabic speaking merchants, and the dislocations produced by the slave trade. Unfortunately, almost no information for this period is available. Steyn (2003) examines mortality in the pre-colonial period in northern South Africa through an examination of skeletal remains. She estimates life expectancy in the period 1000–1300 AD at 23.2. Remains for the post-1830 period show a slight decline in life expectancy after the expansion of European influence.

expectancy in Western Europe rose from 36 in 1820 to 47 in 1900, 67 in 1950, and 78 in 1999. In the analysis of Oeppen and Vaupel (2002), life expectancy in the "best practice" countries (those with the highest life expectancy in the world) has increased linearly since 1840 at a pace of 3 months per annum, with no sign of a slowdown.

In the last half of the 20th century there were rapid gains in life expectancy associated with the international epidemiological transition in which modern health technologies were quickly diffused to the developing world (Acemoglu and Johnson, 2007). Between 1950 and 1999, life expectancy rose by 22 years in both Brazil and Mexico, 28 years in India, and 30 years in China. The pattern of widening and subsequent narrowing of the world health distribution can be seen in the difference between life expectancy in the United States vs. the world average. In 1820, this gap stood at 13 years (39 vs. 26). By 1900 it had risen to 16 years (47 vs. 31), and by 1950, 19 years (68 vs. 49). By 1999, however, the gap had narrowed to only 11 years (77 vs. 66) (Maddison, 2001).

Another way to see this same phenomenon is to look at the speed with which different countries traversed a particular set of life expectancies. For example, in India, life expectancy increased from 26.9 years in 1930 to 55.6 years in 1980. In France, a roughly comparable change took more than three times as long: Life expectancy at birth was 27.9 years in 1755 and reached 56.7 years only in 1930 (Livi-Bacci, 1997; Kalemli-Ozcan, 2002).

Since 1960, the cross-country standard deviation of the infant mortality rate has fallen by almost 40%. However, the cross-sectional standard deviation of life expectancy fell from 1960 to 1990 before turning upward due to the effects of HIV. By 2004, it had returned to its 1960 level. Similarly, Soares (2007) shows that there was "β convergence" in cross-country life expectancy (lower life expectancy predicting faster growth in life expectancy) from 1960 to 1990, but not thereafter.

3.2.2.2 Other Health Indicators

Data on other health indicators show improvements that parallel the increase in life expectancy. Figure 3.6 shows data from Weil (2007) on the adult height and the adult survival rate (the probability of living from 15 to 60 years of age, using the current life table) for 10 countries covering different time intervals, up to 180 years. In a regression with country and year fixed effects, an increase in 10 percentage points in the adult survival rate is associated with a rise in adult height of 1.6 cm. Over the period 1775–1995, average height in Great Britain rose by 9.1 cm.[8,9]

[8] Although height is a useful measure of long-run growth within countries, it does not perform well in cross-section as a measure of the standard of living. Deaton (2007) examining data for women in Demographic and Health Surveys for 43 developing countries, finds no consistent relationship between adult height on the one hand and mortality rates or living standards from the period when those women were children, on the other.

[9] Fogel (1994).

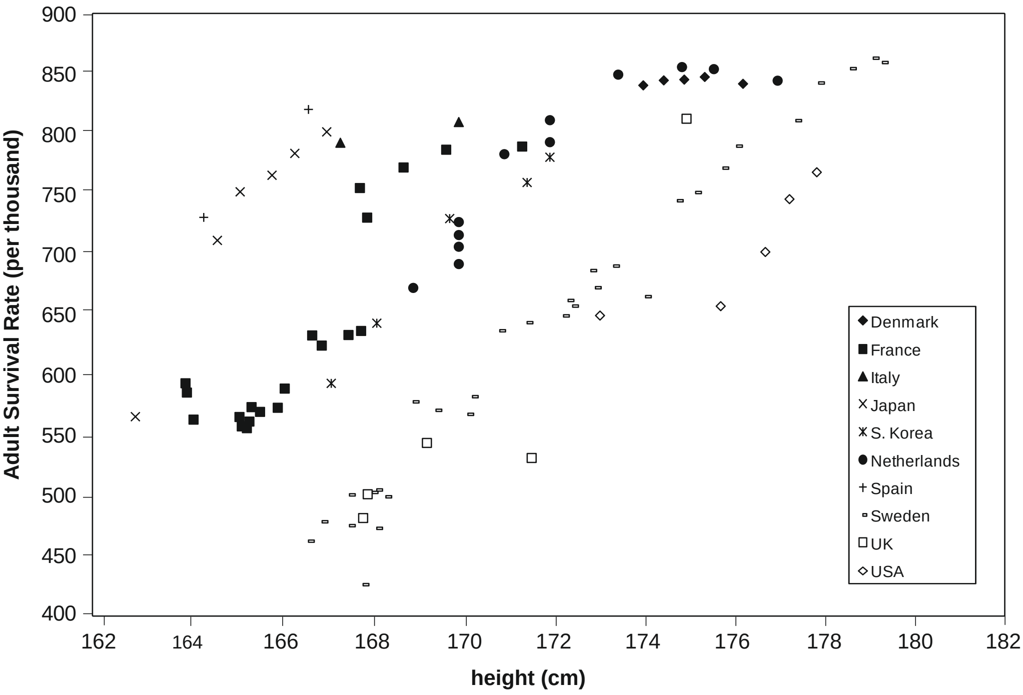


Figure 3.6 Height and adult survival.

The pattern of rapid catch-up during the second half of the 20th century that is observed in the case of life expectancy is repeated for other health measures. Figure 3.6 shows that the relationship between height and adult survival is roughly linear. But what one cannot see in that figure is that the time scale over which these measures grew is not the same for all countries. In Sweden, whose experience is typical for Europe, height increased by 5.5 cm between 1820 and 1900 and a further 6.8 cm between 1900 and 1965. By contrast, in South Korea, the height of adult males rose by 4.8 cm over the 33-year period, 1962–1995, and in Indonesia, adult height attainment as a function of birth year rose by 1.5 cm per decade between 1925 and 1955.[10] Schultz (2010) reports differences in adult female height for birth cohorts separated by 30 years (25–29 years old vs. 55–59 year olds, as measured in roughly 1990) of 3.10 cm in Brazil and 3.43 cm in Vietnam, but smaller jumps of 1.60 in Ghana and 1.54 in Côte d'Ivoire.[11] Similarly, among industrialized countries in Europe, there was a roughly linear decline in age at

[10] Sohn (2000) and Thomas and Frankenberg (2002).

[11] Currie and Vogl (2013) suggest that the slow rate of increase in height in some developing countries may be explained by decreased selection into mortality of unhealthy children.

menarche of 0.2–0.3 years per decade over the period 1860–1980. By contrast, in South Korea, age of menarche fell from 16.8 to 12.7 between 1958 and 1998, a decline of more than 1 year per decade.[12]

A final measure of health is intelligence. Intelligence is a combination of biological aspects of human development and education, whether formal or informal. Thus it falls on the border of health and more conventionally measured human capital from schooling. Nonetheless, there is a good deal of evidence that the aspect of intelligence that is related to biological health has risen over time in developed countries. The so-called Flynn effect refers to the rise in measured IQ that has been observed in many developed countries over the last half century or more (Flynn, 1987). Test scores have been rising at a rate in the neighborhood of three points (out of 100) per decade. The specific IQ tests on which researchers focus (predominantly the Raven's progressive matrices) are designed to measure fluid intelligence, which in theory should not reflect skills acquired in schooling. A number of the health insults and nutritional deficiencies both in utero and very early in life are known to affect intelligence. Thus, the Flynn effect is often taken as resulting from improved nutrition and health over time (Lynn, 1998). Martorell (1998) reports on a number of studies that estimate the impact of low birth weight in currently developed countries as six IQ points among early school age children, and speculates that in more impoverished environments the effect is larger. He also reports that severe, clinical malnutrition is associated with an IQ deficit of 15 points. Within populations, the correlation between body size and IQ tends to be higher in environments where food intake is limited (Sigman and Whalley, 1998). Eppigg et al. (2010) find a very strong statistical relationship between the prevalence of intestinal parasites and average IQ, looking across countries, even when controlling for GDP per capita and average years of education. They theorize that parasites compete for nutrients that are needed for proper development of the brain. They suggest that the Flynn effect may be due to reduced pressure from infectious diseases.[13] While most studies of the Flynn Effect use data from developed countries, Daley et al. (2003) examine data on children in rural Kenya roughly 4 months after school entry at two points in time (1984 and 1998). They find improvements in IQ commensurate with trend growth observed in industrial countries. The fraction of children with insufficient nutrition in their samples fell from 56% in the first cohort to 36% in the second, and the fraction with hookworm fell from 36% to 18%.

[12] Hwang et al. (2003) and Eveleth and Tanner (1990).

[13] There is also a strong relationship between IQ and income per capita, looking across countries. Jones (2011) reports the correlation as 0.7, and that the increase in GDP per capita associated with a one point increase in IQ is far higher than the within-country increase in individual wages associated with the same change in IQ (6–7% for the former vs. roughly 1% for the latter). He argues that there is an effect of IQ on national income that goes through channels outside individual productivity, such as higher patience and ability to solve public goods problems.

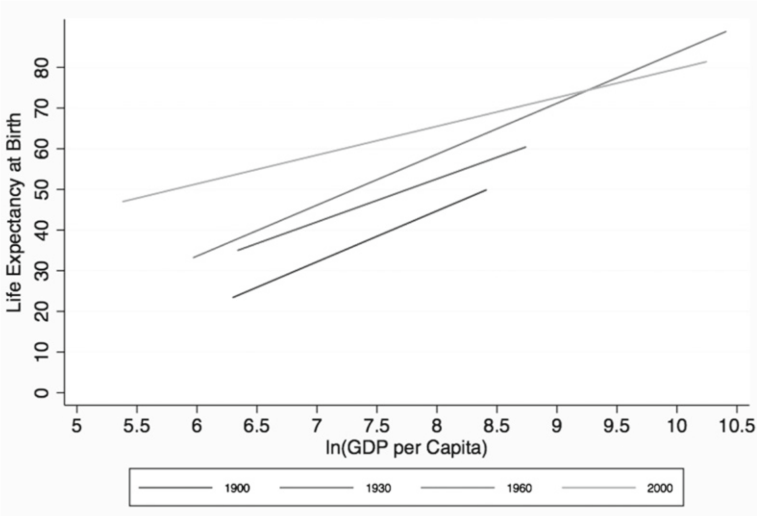


Figure 3.7 Preston curves by year.

3.2.3 Changing Relationship Between Income and Health

Changes over time in the relationship between income and health were most famously pointed out by Preston (1975). Preston noted that the curve representing the relationship between income and life expectancy had shifted upward over time. He estimated that at most, 1/3 of the increase in life expectancy observed between 1930 and 1960 could have been a result of increasing income, with the rest due to the shift in the curve. Figure 3.7 shows a family of estimated Preston curves for cross-country data for the years 1900, 1930, 1960, and 2000.[14]

Another way to look at the relationship between health and income is to examine short-run changes in the two measures. From the cross-sectional relationship, it is clear that over very long periods of time (for example, since the beginning of the 19th century, at which time both income and health differences among countries were small) there must be a positive correlation between the growth of income and the growth of life expectancy. However, as discussed below, there are many reasons why such a relationship might not hold at high frequencies. Figure 3.8 shows the relationship between 40-year changes in the two measures. Weighting country observations by population, the R-squared of a regression of change in life expectancy on change in ln(GDP/capita) is 0.50, and the coefficient implies that a change of 1% in GDP per capita is associated with an increase in life expectancy of 0.13 years.

[14] Data are from Acemoglu and Johnson (2007). Each curve is graphed for the range of income values found in the data. Observations are weighted by country population.

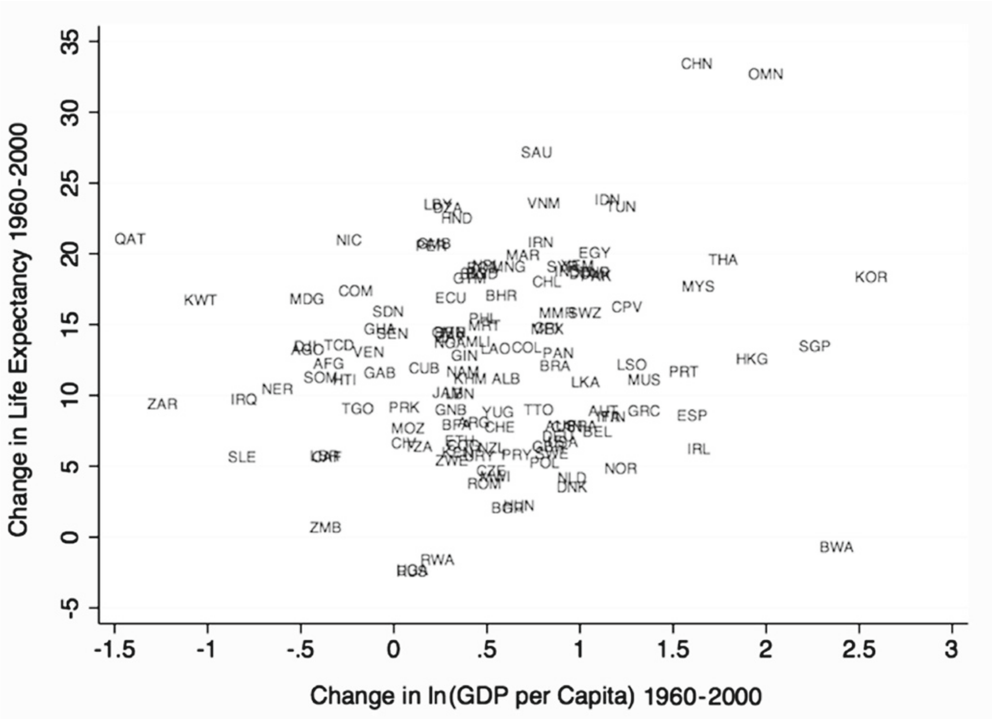


Figure 3.8 Changes in GDP and life expectancy.

3.3. INTERACTION OF HEALTH AND INCOME: A THEORETICAL FRAMEWORK

The various pieces of evidence presented above establish that, in a statistical sense, income and health are strongly related. The exact nature of the correlation varies with the setting (cross-section, time series, country vs. individual), but it is clearly strong. As is usual in economics, the real debate is over what structural relationships underlie these observed data. What causes what, how much variance does this causality explain, and at what time horizon?

As a starting point, one can think of health and income being determined simultaneously. Figure 3.9 presents a simplified framework in which y represents income per capita and v (for vitality) represents health. The effect of higher income in improving health is represented by the $v(y)$ curve. The effect of better health in raising income is represented by the $y(v)$ curve. Equilibrium health and income are given by the intersection of the two curves. In this abstract form, the model can be thought of as applying equally well to either individuals or countries.

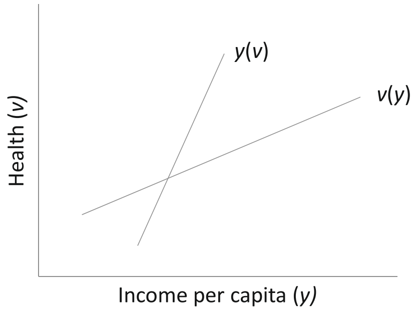


Figure 3.9 Interaction of health and income.

In this simple model, a positive correlation between health and income (looking across countries or individuals or over time) can be induced by three forces:

1. Variation in the $y(v)$ curve, holding the $v(y)$ curve fixed. This would be due to factors other than health that affect income. Examples in cross-country data could be availability of natural resources, differences in institutions or productive technology, etc. Among individuals, shifts in the $y(v)$ curve could be caused by variation in non-health aspects of human capital. Such variation would trace out the $v(y)$ curve, and so in order to match the observed positive correlation between v and y in the data, it would have to be the case that the $v(y)$ curve was upward sloping. In other words, it would have to be the case that raising income improved health.
2. Variation in the $v(y)$ curve, holding the $y(v)$ curve fixed. This would be due to factors other than income that affected health, such as variation in the "disease environment" across countries, or variation in idiosyncratic health outcomes across individuals. Such variations would trace out the $y(v)$ curve, and so for the observed data to fall on a upward sloping line it would have to be the case that the $y(v)$ curve had a non-zero slope (when viewed in a rotated fashion). In other words, it would have to be the case that improving health actually did raise income.
3. Correlated shifts in both curves. This would be the case if some factor shifted both health and income. Looking over time, a natural candidate to produce such correlated shifts is technology, which allows for higher output (given a set of factor inputs) and for better health, holding income constant. Looking across countries, one might think that differences in institutional quality would produce correlated shifts of the two curves. Finally, looking among individuals, a natural candidate for producing such correlated shifts would be education, which raises wages and imparts knowledge that improves health at any given wage level. Correlated shifts in the $v(y)$ and $y(v)$ curves would produce an upward sloping relationship between y and v even if both of these curves had zero slope (in other words, even if there were no causal link from health to income or vice versa).

The empirically observed pattern of health and growth in any particular setting will depend on the slopes of these curves, the relative variances of shocks to them, and the covariance of such shocks.

As in any model where the two curves describing structural effects slope in the same direction, there will be multiplier effects in this simple setup. For example, some exogenous change that affects the level of income, holding health constant, will shift the $y(v)$ curve to the right, raising income directly, but also improving health and resulting in a second round of health-induced increases in income. There will be similar multiplier effects to exogenous shocks to health. These multiplier effects will be larger, the larger are the responses of income-health and health-income. Similarly, it is not hard to introduce nonlinearities in one or both of these relationships that could produce multiple equilibria.

To a large extent, debates about the importance of health in economic growth can be boiled down to claims about the slopes of, as well as the relative variances and correlations of shocks to, the $y(v)$ and $v(y)$ curves. Sachs (2001) stresses the variability of the underlying health environment across countries, arguing that even if tropical countries were rich, they would still be unhealthy. Implicitly, he views the variance of the $y(v)$ curve to be small, and so the observed data on y and v will trace out the $y(v)$ curve—and thus we learn from the data that the $y(v)$ curve is steeply sloped—health has a big effect on income. By contrast, Acemoglu and Johnson (2007) interpret their results (discussed below) as showing that the $y(v)$ curve is flat, and so the correlation between health and income observed in the data results from a combination of correlated shocks to the two and causality running from income to health. Pritchett and Summers (1996) use an instrumental variables approach to argue for a positive effect of income on health—that is, that the $v(y)$ curve is not flat.

Looking at the within-country covariation of health and income, Cutler et al. (2006) argue that relatively little is due to causation running from income to health—in other words, that the $v(y)$ curve is relatively flat. Rather, they view the two most important sources of the observed correlation to be causation from health to income (in particular, the effect of disability on wages) and the effect of education in producing correlated shocks to both curves.

An important observation is that the degree to which different causal channels shape the relationship between income and health need not be the same in all contexts, that is, across countries, historically within individual countries, or cross-sectionally within a country. Indeed, as discussed below, there is good reason to think that this is not the case.

3.3.1 Timing

The framework presented above abstracts from the dynamics of adjustment in health and income. Such an approach is reasonable if one is thinking about differences among countries that vary greatly in their levels of health and income, or alternatively, if one is thinking about changes over very long time spans, like centuries. In considering shorter timespans—for example, in thinking about a country undergoing rapid economic growth or a rapid improvement in health—the dynamics of the process become important.

3.3.1.1 Delays in Health Improvements Due to Human Physiology

One important source of dynamics in the health-growth relationship results from delays inherent in the process of human development. Adult health, and thus the labor input of adult workers, is strongly affected by health conditions early in life, or even prior to birth. Thus, the health of adults does not immediately respond to changes in the health environment.

The most obvious manifestation of this phenomenon is height. In many countries that have experienced rapid income growth, there is a striking age gradient in height among adults: young adults tower over their elderly grandparents. The crucial periods for determining adult height are in utero, during childhood up to the age of four, and then during the adolescent growth spurt. After the early 20s, at the latest, nutrition does not affect adult height. Fogel and Engerman (1992) find that slaves who were fed abundantly only after they entered the labor force (after age 10) remained stunted by at least four inches in adulthood.

The 9 months of gestation are a particularly important period for determining adult health outcomes. Nutritional deficiencies, either in terms of total calories or in terms of specific micronutrients, as well as other health insults, can all have major impacts on adult health. For example, Bleichrodt and Born (1994) estimate that iodine deficiency in utero causes reductions in adult IQ averaging 13.5 points. The Barker Hypothesis holds that fetal malnutrition is associated with ill health, particularly in the form of chronic diseases such as diabetes and coronary artery disease, in the adult years that follow reproductive age (see Almond and Currie (2011) for a review). The hypothesized channel is “fetal programming” via the mechanisms of epigenetics. Nutritional deficiency in utero is often reflected in low birth weight, but this need not be the case. Fetuses exposed to malnutrition only early in gestation may attain normal birth weight but still suffer long-term damage to health.

Beyond malnutrition, there are also disease insults. Almond (2006) shows that in the United States, cohorts exposed to Spanish Influenza in utero had lower educational attainment and higher rates of disability than surrounding birth cohorts. Wages of men were 5–9% lower because of in utero exposure to the infection. Other examples of health insults in utero that produce lifelong health impairments include congenital rubella syndrome, fetal alcohol syndrome, and the effects of maternal smoking. Case et al. (2002) find that maternal smoking affects health outcomes in middle age, even controlling for an individual’s education as well as health and social status earlier in adult life.

Nutrition and disease in childhood can also have lifelong effects. For example, in areas where malaria is endemic, most people have developed substantial immunity by age five. However, cases of cerebral malaria in young children leave lifelong scars. In sub-Saharan

Africa, the prevalence of malaria episodes among adults (that is, the fraction suffering at any point in time) is only half of the prevalence among adults of neurological sequelae from childhood episodes of cerebral malaria (Ashraf et al. 2009). In addition to its direct effects on adult health (and thus adult productivity), ill health in childhood can also impact adult outcomes by reducing human capital accumulation. For example, Bleakley (2007) shows lifelong effects on education and wages from exposure to hookworm during childhood. Bleakley (2010), Cutler et al. (2010), and Lucas (2010) all use national anti-malaria campaigns in the middle of the 20th century as quasi-experiments in order to study the effect of childhood exposure to the disease on human capital accumulation and adult economic outcomes, though with varying findings. Cutler et al. find no effect of malaria eradication on schooling or literacy, no effect on adult female income, and a modest effect on adult male income. At the other end of the spectrum, Bleakley's estimate is that persistent childhood malaria reduces adult income (through a combination of adult health and human capital accumulation) by 50%. Lucas finds that a 10% reduction in malaria incidence raises completed schooling by 0.1 years.

Several studies have examined the long-run effects of childhood nutrition, using a variety of exogenous sources of variation in nutrition, including randomized controlled trials of nutritional supplementation as well as shocks to income during childhood, such as rainfall, war, and famine. These studies generally find that better nutrition leads to improvements in school completion, IQ, height, and wages.[15] Case et al. (2005) suggest that the positive effect of parental income on child health, along with the positive effect of child health on adult economic status, may be an important pathway leading to the observed inter-generational correlation of economic outcomes.

While much of the literature looking at the relationship between child health and adult outcomes focuses on developed countries, where the data are better, Currie and Vogl (2013) suggest that the effect is probably larger in poor countries, where health insults are more frequent and are likely to positively reinforce each other. In order to achieve clean identification, most of the studies economists have conducted on the long-term effects of childhood health have examined one particular health insult (a nutritional deprivation or disease exposure) at a time. (Although Currie and Vogl also point out that in developing countries, the negative effects of early-life health insults on adult outcomes may be blunted by positive selection into survival.)

The extent to which adult health is a function of the adult health environment vs. the health environment that prevailed when current adults were young will be one of the factors determining the speed with which improvements in the overall health environment are translated into improvement in adult health (which is the aspect of health that matters for output). Ashraf et al. (2009), in their simulation of the output effects of health improvements, introduce a parameter that measures the relative importance of

[15] See Currie and Vogl (2013) for an extensive review.

child vs. adult health inputs and show that the dynamic economic effects of health improvements are very sensitive to it.

3.3.1.2 Negative Short-Run Effects of Income Growth on Health

A second important dimension of timing has to do with the impact of income on health via inputs at the national level. As discussed below, an important contributor to health gains is improvements in public health infrastructure, most notably clean water and sanitation. Public health expenditures rise with national income, but in many cases there are appreciable lags in the implementation. Further, many of the short-term effects of economic growth can be deleterious to health.

Inter-regional or international migration, often associated with economic growth, can have negative health consequences, particularly in the short run. The most stark example of this phenomenon is the spread of old-world diseases to the Americas that followed the voyage of Columbus, which resulted in tens of millions of deaths. Smaller expansions of settlement have also produced similar results. In the early 20th century, the spread of Chinese settlement into Manchuria brought a new population into contact with rodents that harbored the bacteria causing plague, leading to a local outbreak that almost became a worldwide pandemic (McNeill, 1998). As described in McGuire and Coelho (2011), increases in the speed of transport in the centuries following Columbus allowed for ever more pathogens to make the leap between continents.

Another important source of increased disease exposure from economic growth is urbanization, both because it brings people into contact with new infectious agents, and because the collections of food and waste in cities make the spread of disease far more likely. Until the early 20th century, even in the most developed countries, cities were far less healthy than the countryside.

Costa and Steckel (1997) find that the average height of native-born residents of the United States declined by 4 cm between the cohort born in 1830 and that born in 1880. Life expectancy at age 10 also fell from the cohort born in 1790 to the cohort born in 1850.[16] They suggest that the decline in health was due to greater exposure to infection resulting from inter-regional trade and migration, as well as to less healthy working conditions that accompanied the move away from farming and home manufacturing. (They do not view urbanization itself as a major cause of the health decline; as of 1860, only 10% of the population lived in cities with a population greater than 50,000.)

3.3.1.3 Health Improvements and Population Growth

A final dimension in which timing is important in considering the relationship between health and growth is in the entanglement of health with population growth. Health is related to population growth via infant and child mortality. One of the reasons that fertility was high in undeveloped countries was to compensate for the fact that so many

[16] Floud et al. (2011), Figure 6.1.

newborns would not reach adulthood. A standard idea in demographic transition theory is that when mortality falls, there is a delay in the response of fertility, and that as a result of this delay there is a spurt of population growth. Acemoglu and Johnson (2007) attribute their finding (discussed below) that mortality reductions in developing countries led to a decline in income per capita to exactly this channel. Their IV estimate of the effect of a change in log life expectancy on the change in the log population size between 1940 and 1980 is 1.67. This implies that an increase in life expectancy from 40 to 60 years would raise population size by a factor of 1.97 over this 40-year period, which is an increase in the annual growth rate of slightly less than 2%. Acemoglu and Johnson claim that the negative economic effects of rapid population growth more than compensated for direct economic benefits from better health, and so income per capita fell.

While the approach of Acemoglu and Johnson is purely econometric, Ashraf et al. (2009) pursue the question of how much a reduction in mortality would be expected to affect population growth, and in turn economic growth, using a simulation model. The demographic side of the model is set up to roughly match the international epidemiological transition studied by Acemoglu and Johnson. Ashraf et al. consider a stylized economy in which age-specific mortality and fertility rates have been constant for sufficiently long that the age structure of the population is unchanging—what demographers call a stable population. The stable population is constructed with life expectancy at birth of 40 years and a total fertility rate of 5.2, yielding population growth of 1.5% per year. The authors then consider an instantaneous shock to health that raises life expectancy at birth to 60 years. To represent the effect of mortality reduction on fertility, they allow age-specific fertility to fall linearly so that after a fixed number of years the net rate of reproduction has returned to its pre-shock level. They trace through the effect of this change on population growth. They find that if fertility adjusts over a period of 50 years, the maximum increase in the population growth rate is 1 percentage point, and that at a horizon of 40 years, population is 1.36 times as large as it would have been without the reduction in mortality. This is significantly smaller than Acemoglu and Johnson's estimate of 1.97. Ashraf et al. also find that in their economic model (discussed below), the rise in life expectancy from 40 to 60 produces an increase in income of 2%. By contrast, the coefficient estimated by Acemoglu and Johnson, applied to this change in life expectancy, implies a decline in income per capita of 41%.

Ashraf et al. then experiment with altering their model so that it produces population dynamics similar to those estimated by Acemoglu and Johnson. This requires having fertility rise in response to a decline in mortality. In this case, they find that at a horizon of 40 years, income per capita would fall by 20% in response to the rise in life expectancy from 40 to 60. In other words, more than half of the gap between the findings of the two studies can be explained by the divergent conclusions regarding the response of fertility, and thus population growth, to a decline in mortality.

3.4. IMPACT OF INCOME ON HEALTH

The improvement in health and extension of life that has taken place around the world in the last two centuries, as described in Section 3.2 of this chapter, is one of humanity's greatest accomplishments. As such, it has been the subject of a voluminous literature. In this section, I take a very selective tour through this literature, focusing in particular on the question of how much of the improvement in health can be attributed, either directly or indirectly, to increasing income. In assessing this question, of course, it is inevitable that one has to address the question of what else, if not rising income, is responsible for health improvements over time and health differences among countries.

The improvement in health that has taken place over the last two centuries resulted from three sets of forces: first, improvements in the standard of living, in particular, better nutrition; second, changes in the public health environment, including sanitation and the supply of clean water; and finally, improvements in medical technology, including antibiotics and other medial treatments. The extent to which credit for improved health should be divided among these sources is a matter of debate. Further, there are not only interactions among the different forces, but also cases where a particular health problem could be remedied by more than one channel (for example, both sanitation improvements and treatment with antibiotics will reduce mortality from infectious diseases).

In the countries that developed first, and in which health improvements began earliest, the three channels just mentioned had their primary effects on health in the order just discussed. In countries that experienced health improvements later, the time pattern has been more heterogeneous. Thus, for example, in some developing countries, the state of medical treatment today exceeds what was available in rich countries in the middle of the 20th century, while nutrition and public health lag further behind.

Improvements in health, as measured by mortality, can be linked to specific changes in both the ages at which people die and to the diseases that they die from. Cutler and Meara (2004) estimate that in the United States, 80% of life expectancy improvements in the first four decades of the 20th century were due to reductions in death before age 45, with two-thirds of that coming before age 15. In the last four decades of the century, by contrast, two-thirds of life expectancy gains came from mortality reductions at ages greater than 45. This change in the distribution of mortality improvement reflected in part the distribution of mortality itself: by the latter part of the 20th century, the infant mortality rate was low enough that even though the rate of mortality decline in this age group was the same as earlier in the century, the contribution to increased life expectancy was only a quarter as large. But it also reflected changes in the rate of progress at different ages. Mortality among the 65+ age group declined at a rate of 0.3% per year in the first 40 years of the century, vs. 1.1% per year in the last 40 years.

Reduced death from infectious diseases accounted for three quarters of the reduction in mortality in the first four decades of the 20th century. The rate at which mortality from infectious diseases declined sped up appreciably in the period 1940–1960 with the deployment of antibiotics, but because mortality from these conditions was already lower than it had been in 1900, the contribution of infectious diseases to overall mortality decline was only half of what it had been in the century's first four decades. By 1960, infectious diseases accounted for only 5% of mortality. On an age-adjusted basis, death rates from both cardiovascular disease and cancer increased over the first 60 years of the century. Cardiovascular disease accounted for 22% of mortality in 1900 but 59% in 1960, while deaths from cancer rose from 5% to 15% over the same period. Between 1960 and 1990, declining death from cardiovascular disease was equal to 98% of entire decline in death rates, and it was for this reason that the decline in mortality was concentrated in ages above 45.[17]

3.4.1 The Standard of Living and Health Improvements

3.4.1.1 Positive Effects of Economic Growth on Health

McKeown (1976) famously argued that much of the reduction in mortality that took place over the last centuries was due to improvements in nutrition, rather than explicit interventions, either public health or medicine. Most significantly, McKeown showed that declines in mortality from a number of infectious diseases took place prior to any such intervention. For example, the death rate from tuberculosis declined by 80% from when his data begin in 1848 to the advent of effective treatment in the mid-1940s. Similarly, Cutler and Meara (2004) show that great reductions in death from infectious diseases that took place in the United States in the first four decades of the 20th century occurred before the availability of medical treatments such as sulfa drugs (invented in 1935) and widespread vaccination.[18] McKeown's argument has been carried forward by Robert Fogel and co-authors in a series of articles and books. Fogel cites both direct evidence on caloric intake as well as data on the resulting changes in body sizes.

Caloric intake is the simplest measure of an input into health. As described in Floud et al. (2011), economic historians have put enormous effort into measuring this input. The majority of work has focused on Britain and France over the last two centuries. Data sources include estimates of total food production and imports; household surveys; and institutional records. In assessing average caloric intake, it is important to adjust for the demographic structure of the population, since children eat less than adults. Their estimate for France in 1785 is 2413 calories per standardized consuming unit (male age 20–39). In England in 1800, the equivalent was 3271 calories. Floud et al.'s estimate is that calories per consuming unit in England rose 20% between 1800 and 1913 and by a further 10% by 1960.[19] In France, calories rose by 65% between 1800 and 1960. The rise in calorie consumption somewhat understates the degree to which nutrition improved,

[17] Cutler and Meara (2004), Table 9.3.

[18] See Deaton (2006) and Cutler et al. (2006) for extensive discussion of this argument.

[19] Tables 4.13 and 5.5.

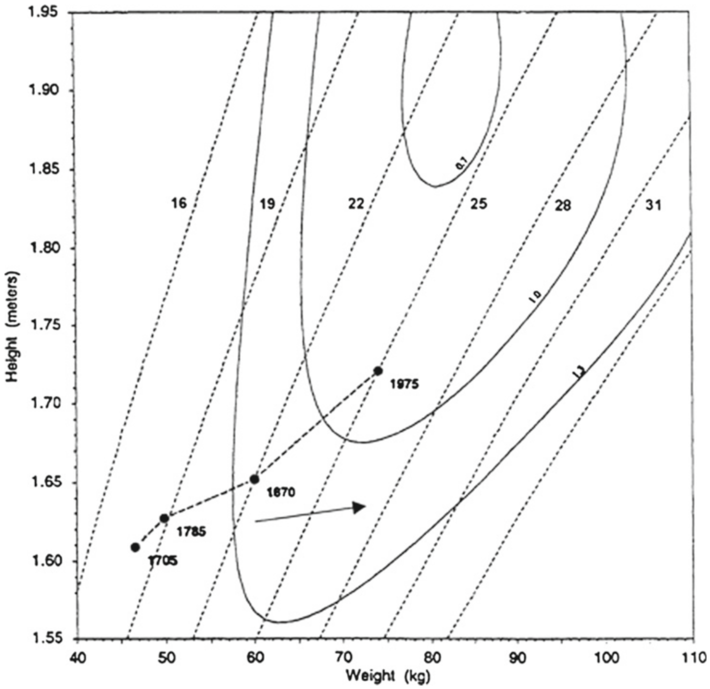


Figure 3.10 A Waaler surface.

in that the caloric demands of labor done by most adults have declined over time, so that more calories are left over for bodily maintenance.

Better nutrition translated into bigger bodies. As discussed above, both adult height and body mass index increased in leading countries over the last two centuries. The final piece of Fogel's argument that increases in living standards have been a major source of health improvement is the observed relationship between anthropomorphic measures, on the one hand, and health outcomes, on the other. Figure 3.10 is an example of a Waaler surface, which shows the relationship between weight, height, and mortality risk. The oval-shaped curves are iso-mortality risk contours for men aged 50–64, labeled to show relative mortality hazards, based on data from Norwegian men. A man with weight/height on the outermost curve had almost twice the mortality risk of a man with weight/height on the innermost curve. The upward sloping lines are iso-BMI curves. Finally, the figure shows estimates of average weight and height for French men at four points in time. Assuming that the relationship between body size and mortality embodied in the Waaler surface has remained stable over time, the change in height and weight shown in this data would have contributed to a significant reduction in mortality. Fogel (1997) shows that changes in height and weight alone explain 90% of the reduction in French crude death rates between 1785 and 1870 and a further 50% of the reduction between 1870 and 1975.

The Fogel/McKeown view that living standards played a major role in improving health has attracted a good deal of criticism. One important argument against the view that nutrition is of paramount importance is the shifting upward of the Preston curve, as discussed above. In the period during which this phenomenon is observed, the vast majority of improvement in life expectancy is due to such shifts, rather than to movements of countries along a fixed curve as income rises. Going further, Soares (2007) shows that between 1940 and 1970, life expectancy rose, holding not only income but also average daily calorie consumption constant. However, the evidence on the Preston curve comes only from the 20th century (particularly after 1930), while much of Fogel's argument applies to an earlier period. Looking at this earlier period, Smith (2013) argues that changes in mortality in the 18th and 19th centuries were well synchronized among countries at different levels of economic development within Europe and North America, and similarly synchronized among different parts of the social spectrum within England. He interprets this finding as evidence that income cannot have played an important role, instead attributing the mortality changes to variation over time in the epidemiological environment.

Another line of argument against the Fogel/McKeown view is that the bigger bodies we observe (and thus the movement over the Waaler surface) do not result solely from a better standard of living (i.e. nutrition). Infection both increases the body's need for nutrition and interferes with the absorption of nutrients from food consumed. Thus, the increase in height and BMI observed historically is not necessarily due solely to more food intake, but may also have resulted from decreased rates of infection (due in the first instance to improved public health, and later to antibiotics).

3.4.1.2 Negative Effects of Economic Growth on Health

Although there is debate about the fraction of increased health that can be attributed to better nutrition, there is little doubt that until recently, with the rise of obesity, diabetes, and other diseases of overconsumption, better nutrition has been a net contributor to better health. However, other behaviors associated with economic growth worked in the opposite direction. Most significant among these has been urbanization. Historically, cities were notably unhealthy, both because they put people into contact with many other potential disease carriers, and because in a large population the risk of contamination of food and water by human waste is greatly increased. For example, in 1900 in the United States, the rural-urban gap in life expectancy among white males was 10 years (54 vs. 44 years). Cities were particularly hard on children. Mortality at ages 1–4 was twice as high in cities as in the countryside. The urban penalty in US mortality disappeared in the early decades of the 20th century (Haines, 2001). In currently developing countries, the dynamics of the rural-urban mortality gap have been different, evidently because superior access to medical care more than made up for the inherent unhealthiness of cities. In South Asia (with the exception of Sri Lanka), infant mortality has been lower in cities than rural areas since the 1970s, with the gap narrowing over time. For example, India's infant mortality rate over the period 1978–1983 was 68 for urban and 111 for rural areas while during the period 1994–1999 the values were 47 and 73, respectively. In sub-Saharan Africa the picture is more mixed, although on average cities also have lower mortality[20]

A second channel through which growth can negatively affect health is pollution. In the presence of an environmental Kuznets curve, growth in a poor country will worsen pollution.The experience of China over the last several decades is a case in point.

Finally, economic growth may not automatically produce health improvements because the extra consumption spending afforded by rising incomes is not always directed in a manner that improves health outcomes. Consumption of tobacco is an obvious example. Further, the translation from higher income to better nutrition is not automatic. Deaton and Dreze (2009) note that although the Indian economy has been growing rapidly since the 1980s, average intake of calories, protein, and other nutrients has declined. Although in cross-section there is a strong, positive relationship between household expenditure and household calorie intake, the intercept of this Engel curve has been shifting down over time. Part of the decline may have been due to reduced calorie demands from infectious disease, physical labor, and fertility, all of which were reduced over this period. However, indicators of nutrition such as children's weight-for-age and height have improved only modestly over this period, and undernutrition remains widespread. Deaton and Dreze suggest that one explanatory factor may have been changes in preferences away from coarse grains and in the direction of more processed foods, due to advertising or emulation of the more affluent classes. See Easterly (1999) for a more general intellectual history of the idea that income growth does not translate into health improvements.

3.4.1.3 Econometric Evidence

Cutler et al. (2006) point out that almost all of China's remarkable improvement in infant mortality took place before economic growth took off in 1980, and similarly that the acceleration in economic growth in India following economic reforms in the early 1990s was accompanied by a slowdown in the rate of decline in infant mortality. Similarly, in Bolivia, Honduras, and Nicaragua, gains in life expectancy on the order of 20 years took place during periods of modest or even negative income growth (Soares, 2007).

Caldwell (1986), looking cross-sectionally at the relationship between income and health outcomes, focuses on the outliers, that is, countries with unusually good or bad health outcomes relative to their income levels. Among the poor health achievers he notes that most have large Muslim populations (leading to limited female autonomy and schooling). He attributes health differences beyond those explained by income to schooling, local health service provision, and possibly family planning (as well as being

[20] Data from Demographic and Health Survey summary reports (various issues) as well as Sahn and Stifel (2003).

a former British colony). Studying episodes of "mortality breakthroughs" such as Sri Lanka, where life expectancy rose by 12 years between 1946 and 1953, his conclusion is that such episodes are more a matter of the political and social will to address health issues than the availability of economic resources.

In terms of the framework described above, such examples of large changes in health in the absence of shifts in income are evidence of there being a large variance of shocks to the $v(y)$ curve. Further, to the extent that these mechanisms are at work, there is no need for the $v(y)$ curve to have a positive slope in order to explain the observed correlation between health and income. At the same time, there is nothing inconsistent with viewing the $v(y)$ curve as having both a lot of variance and a positive slope.

Attempts to estimate the structural effect of income on health, that is, the slope of the $v(y)$ curve, run into obvious identification issues. They also suffer from the difficulty that feasibly identified estimates may only pick up a short-run effect. Easterly (1999) uses cross-country data for the period 1960–1990 on income per capita and a number of health indicators. In decadal data, income growth is linked to *lower* life expectancy, while the relationship between income growth and infant mortality has the expected sign. Income growth is also positively related, though only sometimes statistically significant, with observable inputs into health such as calorie intake, physicians per capita, and access to clean drinking water. Of course, all of these correlations are not well identified. In an attempt to achieve identification, Easterly estimates IV regressions (in changes), using twice-lagged income as well as "policy" measures (black market premium, financial depth, and inflation) as instruments. Here, he finds mixed results with income growth significantly reducing infant mortality but having no significant effect on life expectancy. He concludes that there are "long and variable lags" in the translation of higher income growth into better health. This is also consistent with the observation made above that many of the outliers in the cross-sectional income-growth relationship are countries where income has recently grown very quickly but health has not improved.

Unfortunately, the identification in both Easterly (1999) and Pritchett and Summers (1996) papers is far from perfect. The policy measures used in both papers may easily be correlated with the types of effective institutions that affect health through channels other than income. Combined with the fact that such approaches are really only suited to looking at short-run effects of income changes on health, one is left with little hope of learning much about the slope of the $v(y)$ curve through this approach.

3.4.2 Public Health, Medicine, and Economic Growth

As mentioned above, to the extent that the improvements in health are not direct results of economic growth, via changes in nutrition and other aspects of the standard of living, then such improvements are due to two other channels: improved public health and direct application of medicine. A natural question is to what extent these forces are, in turn, linked to economic growth.

Cutler et al. (2006) organize their narrative of the sources of reduced mortality around the themes of knowledge, science, and technology. Knowledge of the causes of ill health, most importantly the germ theory of disease (empirically validated in the 1880s and beginning to displace previous theories around the turn of the century), allowed for the introduction of effective public health infrastructure, particularly clean water. Accumulation and dissemination of knowledge also allowed for improvements in private health behaviors, ranging from washing hands and boiling water to the reduction in smoking in the United States over the last half century. And of course, new science and technology have been driving forces in medical improvements since the middle of the 20th century. This focus on the role of knowledge has the implication that the explanation for the time series relationship between income and health, on the one hand, is different than the explanation for the cross-country or within-country relationships between these same variables, on the other. The reason that the explanations differ is that at any point in time, at least in the world today, gaps between or within countries in applicable health knowledge must be very small. Cutler et al. (2006) seem to view the cross-country relationship between health and income as resulting from correlated shocks to the $v(y)$ and $y(v)$ curves, particularly in the form of differences among countries in the quality of institutions that impact both income and health. For an explanation of the within-country correlation of health and income, they focus on both causality running from health to income, in particular the effects of disability on earnings, and on the role of education in raising wages and allowing for better application of existing health knowledge.

Soares (2007) similarly stresses the diffusion of ideas (new technologies, personal health practices, and public goods) from rich to poor countries as the driving force shifting the Preston curve upward in the post-war period. However, unlike the diffusion of ideas used in producing output more generally, the ideas that are relevant for health often have significant dimensions of public goods (such as sanitation and clean water), externalities (quarantine, vaccination), or principal-agent problems. Even private health practices that are not reliant on public infrastructure, such as hand-washing, often require public information campaigns to put in place the relevant information. For these reasons, there is a strong complementarity between health ideas and the strength of institutions, particularly government. There is similarly a strong complementarity between health ideas and human capital of those in a position to apply them. Preston and Haines (1991) find that in the late 19th century, prior to the widespread acceptance of the germ theory of disease, the children of doctors and teachers had only slightly lower mortality rates than average. By 1925, such children had mortality rates that were one-third below average. Similarly, at the time of the Surgeon General's report on the health hazards of smoking, there was little variation in smoking rates by educational group. By 1987, smoking among male college graduates had fallen to 17%, vs. 41% for high-school dropouts (Preston, 1996).[21]

This focus on the role of knowledge, science, and technology, rather than economic growth per se, leaves open two questions. First, whether it is possible to really separate the growth of knowledge from the process of economic growth more generally. And second, whether, even with adequate knowledge, there remains a role for income in determining the application of this knowledge.

The ideas about health on which Cutler et al. focus have to be put into practice in order to impact health. Some actions can be taken at the individual level. In the historical context these have included hand-washing, boiling milk, appropriate food storage, and breastfeeding. In currently developing countries, other examples of actions that are taken at the individual level and rely on individuals understanding and valuing their health benefits are the use of chlorine drops in water, sleeping under a bed net, and use of condoms to prevent transmission of HIV. However, much of the benefit from improved knowledge required public action, ranging from quarantine of the sick to food inspection and regulation to construction of public works.

Public health measures, broadly viewed, were probably responsible for most of the improvement in health in the later 19th and early 20th centuries. In the United States, the infant mortality rate declined from 229 per thousand births in 1850 to 69 per thousand in 1930. Cutler and Miller (2005) use a difference-in-difference approach to estimate the effects of water filtration and chlorination on mortality in a sample of 13 large US cities. Their finding is that these water quality improvements reduced mortality by 13% over the period 1900–1936, which was 43% of the total decline in mortality over these years. In addition to clean water, the introduction of refrigeration (especially for milk) played an important role.

Cutler et al. (2006) are certainly correct that from the perspective of a single developing country, economic growth can be considered to be divorced from the advancement of medical knowledge. This same argument is less germane when one considers the *application* of knowledge, either in the form of medical treatments or public health measures. Trained medical personnel, equipment for treatment, and public health infrastructure all cost money. It is certainly true that there are cases where efficiently organized medical establishments have produced great leaps in health using relatively few resources. Cuba has long served as a model of medical care, even as its economy has collapsed. Similarly, in China, massive advances in life expectancy took place prior to economic liberalization (and for this reason, data for 1980 shows China as a notable outlier above the Preston Curve). But these examples do not disprove the claim that economic growth, by allowing

[21] The United States in the late 19th and early 20th centuries is often the subject of studies on the causes of health improvements. However, it is worth noting that the McKeown/Fogel theory is at a severe disadvantage in this context, because the country had unusually good nutrition. Consumption per adult equivalent was 3700 calories in 1900 (Preston (1996)).

for increased spending on public health and medical treatment, will usually pay at least some dividend in terms of health outcomes.

Writing of long-term improvements in health, Cutler et al. (2006) say "Perhaps controversially, we tend to downplay the role of income. Over the broad sweep of history, improvements in health and income are both the consequence of new ideas and new technology, and one might or might not cause the other." Such a view only makes sense if we imagine that scientific and technological progress of the type that has taken place since the Industrial Revolution as being possible in a context in which economic output did not grow—or did not grow at the speed actually observed. In other words, we are to imagine that science could have advanced from its level of 1860 (when the germ theory of disease was developed) until 1960s (beta blockers) in the absence of the massive increases in both income per capita and total output observed in the richest countries during this period. Such a scenario is hard to swallow, for several reasons. First, in standard models of technological progress such as the Schumpterian model of Aghion and Howitt (1992), effort devoted to R&D is a function of the size of market. In the absence of income growth, incentives for R&D would have been much smaller. Second, as stressed by Jones (1995), maintaining a relatively constant pace of technological progress over the last century has required vastly increased resources to be devoted to R&D. The science that produced the germ theory was low budget. More recent medical advances have required enormous spending. The channeling of such resources (through both the public and private sectors) would have been inconceivable in the absence of robust income growth.

The statement of Cutler et al. (2006) that economic growth was not essential for observed health improvements in the most advanced countries is probably meant to be more rhetorical than serious. Underlying it, however, are two serious observations. The first is that, over periods shorter than the "very long run," there indeed may be very little relationship between income growth and health improvement. Even in the most advanced countries, the stock of usable but non-applied health knowledge is so large that many decades of health improvement could take place without any new discoveries being made. Second, when one considers developing countries, the assumption that income growth will automatically lead to health improvement is unwarranted; and the assumption that income growth is the best way to achieve health improvement is even more unwarranted. As Deaton (2006) writes, "Economic growth frequently needs help to guarantee an improvement in population health."

3.5. IMPACT OF HEALTH ON ECONOMIC GROWTH

3.5.1 Direct Productivity Effects

The simplest channel of causality running from health to economic growth is via the productivity of workers. Individuals who are healthier are able to work more effectively, both physically and mentally. Further, adults who were healthier as children will have acquired more human capital in the form of education. Weil (2007) refers to this as the "proximate effect" of health on the level of income.

To pursue this issue, I start with a simple production framework in which health is explicitly incorporated. Assume that aggregate output is given by a Cobb-Douglas production function taking as its arguments physical capital and a composite labor input,

$$Y_i = A_i K_i^{\alpha} H_i^{1-\alpha}, \tag{3.1}$$

where Y is output, K is physical capital, A is a measure of productivity, and i indexes countries. The labor composite is in turn composed of raw labor L, average human capital in the form of education h, and average human capital in the form of health v:

$$H_i = h_i v_i L_i \tag{3.2}$$

A setup like this has been used in the development accounting literature to assess the contributions of productivity, physical capital, and human capital in the form of education to variation in income among countries (see Caselli (2005) for a review). To implement such a calculation, one has to be able to measure the average level of human capital in the form of education at the country level. The approach taken in the literature has been to combine data on the average number of years of education among adults with an estimate of the return to education (Mincer coefficient) that converts years of education into a measure of human capital. The rate of return estimate plays a key role here, because the units in which the data are measured (years of schooling) are not proportional to the amount of human capital. For example, using a standard estimate of 10% per year of schooling, a person with 4 years of schooling does not have twice as much human capital as someone with 2 years, but rather only 1.21 times as much.

To proceed analogously in the case of health, we need a consistent measure of health across countries and a measure of "return to health" that can be used to convert such a measure into units of human capital in the form of health. Compared to the case of human capital in the form of education, there are two additional complications. First, unlike human capital in the form of education, where years of schooling seems like a reasonable summary measure, health has many different dimensions that might be relevant for productivity. Second, in the case of health there is not as long a tradition of measuring rates of return as there is in the case of education.

3.5.1.1 Estimates of the Return to Health Characteristics

Define w_i as the wage of the labor composite in country i (this could be its marginal product, although this is not necessary). The wage of worker j will depend on his own health and education, as well as this aggregate wage:

$$ln(w_{i,j}) = ln(w_i) + ln(h_{i,j}) + ln(v_{i,j}) + \eta_{i,j}, \tag{3.3}$$

where the last term is an individual-specific error.

As a simplification, I take the approach that the many different aspects of health that we observe (height, survival, etc.) are all functions of a single, underlying (latent) health status which is scalar. This is obviously an extreme approach—Weil (2007) discusses some of the biases that it introduces. Modeling underlying health as being a scalar does not mean that individual aspects of health will all move together, however. Instead, I allow for outcome-specific shocks at the individual level that can reflect luck, genetics, and so on. For example, consider the relationship between underlying health (z_j) and height:

$$height_j = constant + \gamma_{height} z_j + \epsilon_{height,j}. \tag{3.4}$$

Latent health is never observed directly, so the coefficient that relates health to height is not observable either. However, the assumption of latent health being scalar allows for the calculation of a useful measure of health's impact. Assume that the relationship between latent health and v (the aspect of health that determines wages) is determined by a similar equation (the use of log here follows the existing literature).

$$ln(v_j) = constant + \gamma_v z_j + \epsilon_{v,j}. \tag{3.5}$$

The ratio of the coefficients γ_v/γ_{height} is defined as the "return to height." It tells by how much a change in underlying health that raises height by one unit will raise log wages. The return to height is not the same as the observed relationship between height and wages, both because observed height contains a component that is unrelated to underlying health ($\epsilon_{height,j}$), and because in general latent health (and thus height) will be correlated with other factors that affect wages.

Calculations of the return to health characteristics like this can be done for any health outcome. They will be most informative, however, to the extent that the health outcome is representative of the totality of health—in other words, in cases where the assumption of latent health being scalar does the least violence to reality. It is for this reason that I focus on height, which is often viewed as a useful summary measure of nutrition and health insults through early adulthood.

The question then becomes, how to estimate the return to height? Simply regressing log wages on height is clearly problematic. People with higher incomes can afford better health inputs, and unobserved characteristics (such as being from a wealthy family) may affect both income and height. A series of papers has estimated the return to height using a IV approach (see Schultz (2002), Ribero and Nunez (2000)). Data are from Ghana, Brazil, and Mexico, and the instruments are inputs into health in childhood such as distance to local health facilities and the relative price of food in the worker's area of origin. Education is included as a control. The estimated return to height ranges from 7.8% to 9.4% per centimeter.[22] Unfortunately, the instruments used in these analyses

[22] Knaul (2000) does a similar analysis using age at Menarche as a measure of health, while Schultz (2005) presents IV regressions in which health is measured by both height and BMI.

have potential problems. To the extent that good inputs into child health reflect family characteristics that also lead to high wages, these estimates of the return to health will be upwardly biased.

As an alternative, I identify the return to height using exogenous variation in uterine nutrition among monozygotic twins, taking advantage of estimates from Behrman and Rosenzweig (2004). Within pairs of monozygotic twins there are significant variations in birth weight, reflecting differences in intrauterine nutrition due to the location of fetuses within the womb. In their sample, which covers female monozygotic twins from the United States, the average absolute gap in birth weight is 10.5 oz, compared to a mean birth weight of 90.2 oz. Behrman and Rosenzweig regress within-pair difference in log wages, adult height, and education on the difference in fetal growth (measured in ounces per week of gestation). Their estimate is that a one-unit difference in fetal growth leads to a gap of 0.190 (standard error of 0.077) in log wages, 3.76 (0.43) centimeters in adult height, and 0.657 (0.211) years of schooling. Dividing the estimated effect of fetal growth on log wages by the estimated effect of fetal growth on height gives a TSLS estimate of the return to height of 5.1% per centimeter. This return includes the effect of improved health in raising education. Making an adjustment to eliminate this channel (see Weil (2007) for details) yields an estimated effect of health as proxied by height on wages, holding education constant, of 3.3% per centimeter. A similar calculation using data on Norwegian twins from Black et al. (2007) yields an estimate of the same effect of 3.5% per centimeter. In the calculations that follow, I use the average of these two estimates—3.4% per centimeter.

This estimate of the return to height can be applied to the historical data discussed above. In the typical developed country, the rise in adult height over the last 200 years has been roughly 10 cm. My estimate of the return to height thus implies that labor input per worker went up by a factor of 1.4 (in the steady state of a standard growth model, this will also be the effect on output per worker). Thus, while higher labor productivity due to health has been an important factor, it is certainly not nearly the dominant factor in income growth. To put some quantitative flesh on this statement, consider a country in which income has risen by a factor of 15 over this period. The fraction of this rise due to improved labor productivity from better health can be calculated as $ln(1.4)/ln(15)$, which comes to 12.4%.

One important benchmark against which to compare my estimate of the increase in labor input over time comes from Fogel (1997). Looking at data on the distributions of caloric intake and basal metabolic needs in the UK over the period 1780–1980, he calculates that improved nutrition raised labor input per working age adult by a factor of 1.96.

3.5.1.2 Health's Overall Contribution of Cross-Country Income Variance

In addition to asking how health has contributed to growth over time, we would like to ask how much of cross-country variation in income it explains. However, there are two obstacles to using the estimate of the return to height just derived for this purpose.

First, there do not exist consistent cross-country data on adult height. Second, one might worry that genetic factors affecting height but not health vary across countries. For these reasons, Weil (2007) creates a mapping from changes over time in height to changes over time in the adult survival rate (ASR), using the data on both variables presented in Figure 3.6. His estimate is that a change in the ASR of 0.1 is associated with a change in height of 1.92 cm. This implies a "return to ASR" of 0.653, which in turn says that an increase in ASR by 0.1 will raise labor input per worker by 6.7%. In cross-country data, ASR ranges from 0.214 (Botswana) to 0.904 (Iceland). The implied increase in labor input per worker moving over this range would be a factor of 1.59.

Using this estimate, Weil then asks how much of the variance in cross-country income can be explained by health. Following Caselli (2005), variance in log output per worker is decomposed into pieces attributable to physical capital, human capital in the form of education, human capital in the form of health, and a productivity residual. The variance in log output per worker is equal to the sum of the variances of these component terms, along with a full set of covariances. One can then calculate the reduction in the variance of the log of output per worker that would result from eliminating health gaps among countries; this is just the variance in $ln(v)$ plus all of the covariance terms that involve v. Setting these to zero reduces the variance of log output per worker by 9.9%. As an additional measure, Weil calculates the contribution of health to the 90/10 income ratio. In the raw data, the ratio is 20.5. Eliminating health gaps, the ratio would fall to 17.9, with the large majority of that reduction coming from a fall in the 50/10 income ratio.

These results say that health is a significant contributor to cross-country income differences, but that it is not of overwhelming importance. For comparison, the effect of health estimated here is a little more than one-third as large as the contribution of human capital in the form of education to cross-country income variance. It is also of interest to note that the fraction of cross-sectional income variance explained by health (9.9%) is quite similar to the back-the-envelope calculation in the last section of the fraction of long-term income growth explained by health (12.4%).

3.5.2 Other Channels

The analysis in Section 3.5.1 focuses on worker productivity. However, there are several other channels by which changes in health may impact economic growth.

3.5.2.1 Longevity and Human Capital Accumulation

The idea that reducing mortality will raise the return on human capital investments, and thus the level of schooling, has a long pedigree in economics. Discussions of the literature can be found in Kalemli-Ozcan et al. (2000) and Hazan (2009); the latter traces the mechanism to Ben-Porath (1967).

To assess the potential size of this effect, I consider a simple model in which individual earnings are proportional to human capital h, which is in turn a function of years of schooling: $h = f(s)$. I abstract for trend growth in wages and the method by which schooling is financed as well as the risk associated with mortality, and simply assume that schooling is chosen to maximize the expected present discounted value of lifetime earnings. Further, for simplicity, I assume that the only cost of schooling is the opportunity cost of foregone wages. The value of s is chosen to maximize:

$$\int_{s}^{\infty} S(a)f(s)e^{-ra}da, \tag{3.6}$$

where $S(a)$ is the probability of survival to age a. For the $f(s)$ function, I use the specification from Bils and Klenow (2000):

$$f(s) = \frac{\Theta}{1-\Psi}s^{1-\Psi}.$$

Based on cross-country data on the Mincerian return to schooling, they estimate values of $\Theta = 0.32$ and $\Psi = 0.58$. I take age zero (the first age at which schooling is possible) to be five. To match the example studied by Hazan, I start by using data on survival (the $S(a)$ function) from age five for the cohort of males born in the United States in 1850, when life expectancy at age 5 was 52.5 years.[23] Hazan reports that this cohort received an average of 8.7 years of schooling. I choose the real interest rate so that optimal schooling matches this value.[24]

To assess the effect of declining mortality, I hold the other parameters constant and change the $S(a)$ function to match that of the cohort born in 1930, for which life expectancy at age five was 66.7 years. Optimal schooling rises to 9.6 years. In fact, average years of schooling for this cohort was 13.3. Thus, the pure mortality effect explains roughly one-fifth of the actual increase in schooling that took place over this period. This exercise shows that reduced mortality over the range found in cross-country or historical data should have some effect on schooling, but that we would not expect it to be the dominant explanation.

Some empirical evidence also supports the model of decreasing mortality raising schooling. Of course, estimation of the effect is made difficult by the fact that mortality is correlated with other determinants of schooling, and is itself endogenous. The solution is to look for cases in which there is plausibly exogenous and sharp variation in mortality. Oster et al. (forthcoming) examine the effect in US data on individuals at risk for Huntington Disease, which onsets during adulthood, reducing life expectancy by roughly 20 years and healthy life expectancy by 35 years. Individuals with one parent who suffered from the disease have a 50% chance of developing it themselves. They can find out their fate either by taking a genetic test or with the appearance of early symptoms.

[23] I am grateful to Moshe Hazan for sharing this data.

[24] The implied value of r is 8.7%, which might be viewed as high. However, given both that human capital investment carries risk, and that the discount rate applied may reflect credit market imperfections, I don't think of this value as unreasonable.

Oster et al. find that the information that he/she will suffer from Huntington Disease lowers the probability of an individual's completing college by 30–33% points. Their estimate, extrapolated to cross-country data, implies that differences in mortality explain about 10% of the observed variation in college enrollment. Consistent with the calculation above, they conclude that the time-horizon effect exists as predicted by economic theory, but that it is not the major determinant of schooling variation. Jayachandran and Lleras-Muney (2009) examine the effects of a rapid reduction in maternal mortality in Sri Lanka over the period 1946–1953, which raised female life expectancy at age 15 (censored at 65) by 1.5 years, or 4%. Their estimates, based on regional variation in maternal mortality as well as male-female differences, are that every extra year of life expectancy raised literacy by 0.7% points and education by 0.11 years. Once again, the 0.17 years of increased female schooling due to mortality reduction is small compared to the total increase of 1.5 years comparing women in the treated and untreated cohorts.

The most trenchant critique of the view that time horizon influences schooling has come from Hazan (2009). He argues that the essence of the Ben-Porath mechanism is that an increase in survival that induces a rise in schooling must also induce a rise in lifetime labor supply. In his paper, Hazan measures expected total working hours (ETWH) over the lifetime for cohorts of American men born between 1850 and 1970. In addition to mortality, ETWH is affected by labor supply along both the extensive margin (working or not working) and the intensive margin (hours per week). He shows that declines in weekly hours, along with earlier retirement, have more than offset the decline in mortality. For example, ETWH at age 20 fell from 112,199 for men born in 1850 to 81,411 for men born in 1930.

Hazan's observation that ETWH has fallen over time is indeed well taken, but it is worth noting that the paper does not actually show that changing mortality did not affect schooling. Rather, it shows that even though falling mortality worked to increase ETWH, other factors more than undid this effect. We can still believe that the Ben-Porath mechanism works, which is to say that ETWH positively affects schooling. In that case, some other factors must have raised schooling even though the effect of ETWH would be to reduce it. Thus, if mortality had not fallen, ETWH would have fallen more than what we observe, and schooling would have risen less. Even if one knew with certainty that the effect of mortality on schooling took the form described above, it would have to be the case that other factors also affected schooling. Hazan shows that the net effect of lifetime hours on schooling should have been negative, because the working week and retirement age have fallen. However, if mortality had not fallen as well, the decline in ETWH would have been larger, and so schooling would have risen less.[25] Another critique of Hazan is that, as pointed out by Cervellati and Sunde (forthcoming), reduced

[25] Lonstrop (2013) points out that even in the framework of Hazan, there is an important interaction by which increased longevity raises the impact of other factors, such as the return to human capital, on optimal years of schooling.

labor supply on the intensive margin (i.e. fewer hours worked per year) decreases the opportunity cost of schooling as well as the benefits to additional years of education.

3.5.2.2 Mortality, Fertility, and Human Capital Investment

In the model of Soares (2005) reductions in both child and adult mortality lead parents to increase investment in their children's human capital (beyond a zero level that holds in Malthusian equilibrium) and to lower fertility. The reduction in fertility goes beyond the amount that would be induced by lower mortality, if parents were aiming to hold the expected number of survivors fixed. Somewhat similarly, in the model of Kalemli-Ozcan (2002), reduced mortality, by reducing variance of realized survival outcomes, allows parents to reduce their precautionary child-bearing, and thus the average number of surviving children, and this in turn allows for higher human capital investment.[26] However, Hazan and Zoabi (2006) argue that the effect of longevity on human capital investment is not clear in the presence of quality-quantity tradeoffs because increased longevity positively affects quantity as well as quality.

3.5.2.3 Other Theoretical Channels

Lorentzen et al. (2008) stress a set of effects of short time horizons due to high mortality that go beyond investment in human capital. Specifically, they see high mortality as negatively affecting physical capital accumulation (because individuals see lower probability of using their savings) as well as raising fertility. "The prospect of early death," they write "brings shortsighted behavior." This can include not only failure to put aside resources for the future, but risky activities such as unprotected sex and smoking that further raise the hazard of mortality. (Consistent with this view, Oster (2012) finds that reductions in risky sexual behavior in response to the HIV epidemic in Africa were smallest in areas with high non-HIV mortality.)

Bloom et al. (2003) show in the context of a life cycle model that increased longevity will raise saving rates at every age, even allowing for endogenous changes in retirement age. This in turn will raise capital accumulation and output. In their empirical work they find that higher life expectancy raises national saving rates, controlling for the age structure of the population.

Change in health is also related to population aging. Although the largest contribution to aging in developed countries is the decline in fertility that has taken place over the last half-century or more, a secondary contributor has been the decline in mortality at older ages. Population aging, in turn, puts great strain on government transfer schemes, potentially leading to tax increases that will sharply reduce growth (see Weil, 2008 for a review). Of course, lower old-age mortality has been accompanied by lower morbidity,

[26] Lorentzen et al. also stress the importance of adult mortality for the net rate of reproduction: since deaths beyond childhood are nearly impossible for a parent to "replace," mortality in this period should lead to precautionary childbearing and a higher expected number of survivors.

and so the economic problems due to this aspect of population aging are relatively easy to fix by increases in the retirement age (the same cannot be said for population aging due to lower fertility in the past). For political reasons, retirement ages do not seem to rise as quickly as would be warranted by better health of the elderly. Kalemli-Ozcan and Weil (2010) present a model in which falling uncertainty about mortality in old age can actually lower retirement ages.

3.5.3 Econometric Analyses of Health's Effect on Economic Growth

Given the numerous theoretical channels by which health improvements can affect economic growth, one strategy is to look at the reduced-form effect of actual health differences (or improvements over time) on growth. Of course, the endogeneity of health, and the possibility of omitted factors that affect both health and growth mean that any econometric approach must carefully deal with the issue of identification.

Gallup and Sachs (2001) set the pattern for much of the literature that was to follow, putting disease in the framework of a standard growth regression. In their framework, the steady-state level of income per capita in a country is determined by the level of disease as well as some other covariates X:

$$ln(y_{ss,i}) = \beta_0 + \beta_1 disease_i + \sum_j \beta_j X_{j,i}. \tag{3.7}$$

Following Mankiw et al. (1992), the growth rate of income per capita is taken to be a function of the gap between the current level of income and the steady state:

$$\frac{\dot{y}}{y} = \lambda(ln(y_{ss,i}) - ln(y_i)). \tag{3.8}$$

Substituting (3.7) into (3.8) gives an equation relating growth with the current level of income and the determinants of the steady state. This allows for the interpretation of the parameters in a standard growth regression of the form:

$$growth_i = \gamma_0 + \gamma_1 disease_i + \gamma_2 ln(y_i) + \sum_j \gamma_j X_{j,i} + \epsilon_i. \tag{3.9}$$

Specifically, $\gamma_1 = \lambda\beta_1$ and $\gamma_2 = -\lambda$.

The disease measure that Gallup and Sachs (2001) use is the fraction of the population at risk for falciparum malaria. The dependent variable is income growth between 1965 and 1990, while the controls include measures of geography, institutions, schooling, and life expectancy (to control for other diseases). Their estimated value of γ_1 is -1.3, leading to their widely cited conclusion that eliminating malaria in a country where it was endemic (index of 1.0) would raise growth by 1.3% per year. Another way to interpret the Gallup-Sachs finding is to look at the implied effect of malaria on the steady-state level of income per capita. Their estimate of γ_2 is -2.6. Dividing γ_1 by γ_2 and reversing the sign gives an estimate β_1, the effect of the disease on steady-state output per capita. In this case, the value is -0.5, implying that going from a malaria index of one to an index of zero would raise steady-state output by 65%.

Threats to identification in the Gallup-Sachs approach arise if malaria is endogenous and/or if the disease environment that generates malaria is correlated with geographic factors that independently affect output (and are not properly controlled for). The former problem can be dealt with by instrumenting for malaria prevalence with a measure of "malaria ecology" created by Kiszewski et al. (2004). This measure is based on the biological characteristics of vector mosquitoes as well as climate data on a 0.5-degree grid.[27] Sachs (2003) runs regressions in which the dependent variable is the log of GDP per capita. He controls for the quality of institutions but not for schooling or life expectancy (in order to measure the overall effect of malaria). The coefficient on malaria, instrumented with malaria ecology, is close to one, implying that in high prevalence regions, eliminating the disease would raise output per capita by a factor of 2.7.

While the above papers focus on malaria, Bloom et al. (2004) use a more general measure of health, specifically life expectancy. Using lagged values of the endogenous variables as instruments, they estimate that an increase in life expectancy by one year raises steady-state output per capita by 4%. (The paper controls for accumulation of physical capital and human capital in the form of schooling, and so any health effect that runs through these channels is not included in the estimated effect.) Bloom et al. (2004) also summarize the results of 13 other studies that run similar regressions of GDP per capita or productivity on life expectancy, which get broadly similar results. Unfortunately, the use of lagged dependent variables as instruments is very questionable in this case.

Lorentzen et al. (2008) follow a IV approach similar to Sachs (2003), but using health measures that go beyond malaria. They include both infant mortality and adult mortality on the right-hand side of a regression in which the growth rate of income per capita from 1960 to 2000 is the dependent variable, also including a relatively standard set of controls for institutions. The instruments are measures of climate and geography, along with the malaria ecology measure. Their IV estimates of the effect of mortality are extremely large—even larger than the OLS relationship between mortality and income.[28] For example, moving from the values from India (infant mortality of 0.108, adult mortality of 0.294) to the values for the United States (0.015 and 0.197) implies that steady-state income would rise by a factor of 13.1—an enormous amount. In terms of the framework discussed above, they view the slope of the $y(v)$ curve as large. While

[27] Alsan (2012) creates a similar index of suitability for TseTse fly, and finds that this predicts patterns of pre-modern development in Africa.

[28] I focus on the coefficients in column (1) of Table 7 of their paper. These are -6.699 on adult mortality, -20.299 on infant mortality, and -0.985 on the log of initial income per capita.

the specification "passes" an overidentifying test, there is good reason to be suspicious that climate and geography cannot in fact be excluded from the second stage of their regression.

3.5.3.1 Using Changes in Health as Instruments

The studies discussed above rely on cross-sectional variation in the health environment in order to identify the effect of health on income. This approach suffers from the inevitable problem that determinants of health are correlated with aspects of geography and climate that may have independent effects on income, and which are difficult to properly control for. The other way to achieve identification of the effect of health is to look in the time domain, in particular to examine rapid changes in health status that differentially affect different groups or regions. The papers by Bleakley (2010), Cutler et al. (2010), and Lucas (2010) discussed above all used this approach in studying the effect of malaria on human capital accumulation. In order to estimate the effect on GDP at the national level, rather than in data on individuals, the health shocks examined must be large enough to produce an effect that can be distinguished from background noise. Two papers have taken this approach.

Ahuja et al. (2007) examine the economic impact of HIV/AIDS in sub-Saharan Africa. The disease has reduced life expectancy by up to two decades in a number of countries. Because it is primarily productive adults who are dying, one would expect the economic impact of the disease to be particuarly strong. As an instrument for the spread of HIV, Ahuja et al. use variation in the male circumcision rate, which differs significantly among countries for cultural reasons. The authors show that a low circumcision rate is a good predictor of the extent to which HIV spread in the population, and that it is uncorrelated with other factors likely to have affected growth. In the second stage of their analysis, they show that HIV, as predicted by circumcision, has no effect on the level of GDP per capita, although it is correlated with a slowdown in educational gains and an increase in poverty as measured by malnutrition.

Acemoglu and Johnson (2007) (AJ) look at cross-country variation in health improvements during the international epidemiological transition, starting in the 1940s, in which modern health technologies were rapidly transferred to the developing world. Their analysis proceeds in three steps. They begin by looking at cross-country data on disease-specific death rates prior to the transition. They combine these data with information on the rates at which death rates from different diseases declined, based either on the dates of discovery of disease-specific treatments or worldwide declines in disease-specific mortality, in order to construct a measure of predicted mortality change for every country in their sample. This measure of predicted mortality change should not be related to the component of actual mortality change in each country that results from economic growth or institutional improvements. In the second stage, they show that the predicted change in mortality that they construct is a very good predictor of actual change in life

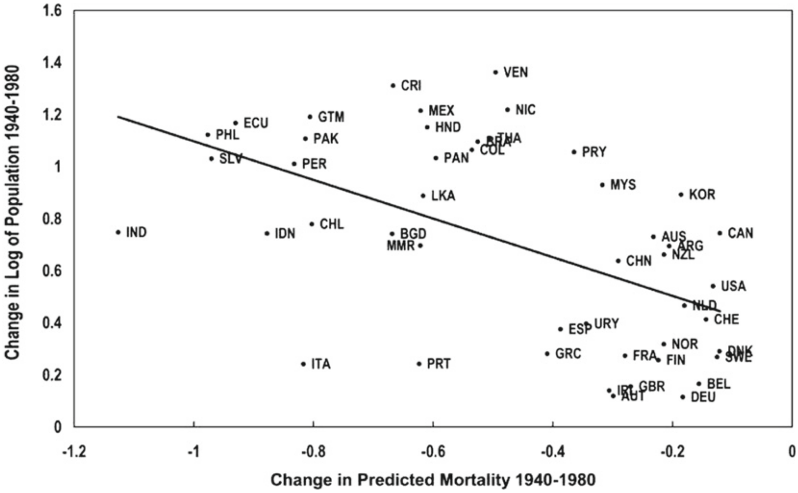


Figure 3.11 Effect of health improvements on population.

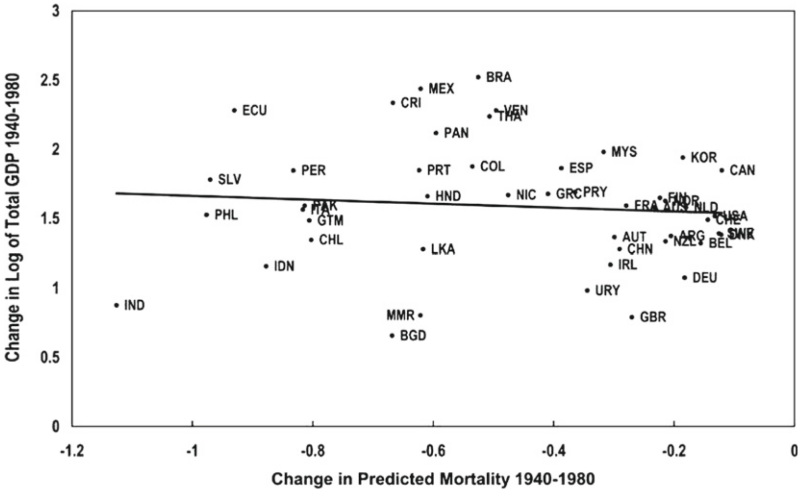


Figure 3.12 Effect of health improvements on total GDP.

expectancy over the period 1940–1980. Finally, they regress a series of outcome variables (population size, total GDP, GDP per capita, GDP per working age adult) on the change in life expectancy, instrumented with changes in predicted mortality.

Figures 3.11 and 3.12 show the key result in the paper. Figure 3.11 shows that reductions in mortality led to higher population growth. Their point estimate from a regression of log population change from 1940 to 1980 on the (instrumented) change in log life expectancy over the same period is 1.67. By contrast, Figure 3.12 shows no statistically significant relationship between the change in predicted mortality and the change in total GDP. In other words, countries with larger reductions in predicted mortality did not see total income rise faster. When AJ regress the change in log GDP per capita on log life expectancy, instrumented with predicted mortality, the coefficient is −1.32 and is statically significant. The coefficient implies that a country in which life expectancy rose from 40 to 60 would experience a 41% decline in income per capita, holding other factors constant. AJ attribute the negative effect they find to the entanglement of health with population growth discussed above: rapid declines in mortality unleashed a population explosion which through the mechanisms of Solow and Malthus reduced income per capita.

The findings of AJ are not completely comparable to those of the cross-sectional studies discussed above, since the cross-sectional studies are implicitly looking at the very long-run effects of health, while AJ are looking only over a period of 4–6 decades. Nonetheless, it is clear that AJ results are contrary to the drift in much of the cross-country literature, which finds a large, positive effect of health.

Bloom et al. (forthcoming) note that in the data studied by AJ, declines in mortality were not randomly distributed among countries. Rather, as a consequence of the narrowing of cross-country health gaps discussed above, the largest gains in life expectancy were in the countries where life expectancy was lowest. The correlation between initial life expectancy and the subsequent change life expectancy is −0.97. Initial life expectancy is also correlated with subsequent growth of income per capita (correlation of 0.50). The latter correlation, say Bloom et al., is to be expected: In a model of conditional convergence such as that presented in Section 3.5.3, any factor that affects steady-state income per capita will also affect growth, conditioning on initial income. They argue that life expectancy falls into this category, since there is abundant evidence that health raises individual productivity.

Bloom et al. argue that for these reasons, the initial level of health cannot be excluded from a regression in which income growth is the dependent variable. When they re-run the AJ analysis, including both initial life expectancy and initial income on the right-hand side (the latter to control for convergence dynamics), the AJ result goes away. In their reply to this critique (Acemoglu and Johnson, 2013), AJ start by pointing out that simply controlling for initial income does not make their result go away. Regarding the effect of controlling for initial life expectancy, they concur with Bloom et al. that this is very highly correlated with the change in life expectancy, and so in a mechanical sense there is no surprise that putting both on the right-hand side of a regression kills the statistical significance of the change in life expectancy. However, they argue that theory imposes limits on how large the effect of life expectancy on subsequent growth should be. When they impose these limits (either using the approach of Ashraf et al. (2009), discussed in the next section), they find that their result survives. They find the same thing when they control for the potential effect of initial life expectancy using a panel data approach.

My own work in this area (Ashraf et al. 2009) focuses on a different potential problem with the AJ result. Ashraf et al. question AJ's finding regarding the effect of reduced mortality on population growth. Although falling mortality should indeed increase population growth, Ashraf et al. in their simulation model, are unable to match the size of the increase that AJ find. This suggests that there is a negative correlation between life expectancy in 1940 and some unobserved factor(s) (that is, something other than the decline in mortality) that affected population growth over the period 1940–1980. As discussed above, this unexplained population growth explains about half of the difference between AJ's finding and the conclusion of Ashraf et al. (2009) that increases in life expectancy should have a modestly positive effect on income growth. A potential explanation for the remainder of the gap is that the same unobserved factor(s) that predicted rapid population growth in countries with low life expectancy in 1940 also predicted slow income growth (for reasons unrelated to population or health) in such countries. Countries that had low life expectancy in 1940 differed in a systematic fashion from those that had high life expectancy: they had different environments, colonial history, levels of institutional development, and demographic histories. It would not be surprising if some element in that set of characteristics had a direct effect on subsequent population or income growth.

3.5.4 Simulation Models

The reduced form estimates discussed in the previous section are one attempt to encompass all the different channels by which health affects economic growth in a single analysis. This reduced form approach is attractive precisely because there are so many different channels through which health could matter, each with its own long and variable lags. However, the difficulty of achieving identification in this context is severe.

The alternative to such reduced form regressions is to create a simulation model in which the different channels can be individually specified, based on credible microeconomic evidence. Crucially, it is easier to achieve identification of individual channels than it is to identify the reduced form effect of health. A further benefit of the simulation approach is that it allows the researcher to exploit a good deal of quantitative macroeconomic theory developed in the context of growth.[29]

The progenitor of this type of analysis is Young (2005), who used a simulation model to examine the effects of HIV/AIDS on the development of the South African economy. His model combines a relatively standard aggregate production framework with a model of household optimization over fertility, labor supply, and children's education. AIDS is incorporated into the model via estimates of the fraction of the population that is HIV positive as well as transition times into illness (at which time labor input ceases) and

[29] Ashraf et al. (2013) discuss the history of the use of simulation models to address the somewhat related question of how fertility decline affects economic growth. Such models have been around for more than half a century, but they fell out of favor in the 1980s, being viewed as ad hoc and opaque.

death. The underlying parameters describing the fertility response to HIV and to female wages, returns to education, determinants of parental investment in children, and the elasticity of labor supply are estimated by Young using South African household data. His results are dominated by effects of reduced fertility, both as a collateral result of measures to prevent HIV transmission, and because of rising wages for women. Lower fertility, combined with adult mortality from AIDS, slows labor force growth, and through the Solow and Malthus channels (rising capital/labor, and land/labor ratios, respectively) raises income per capita. These effects outweigh the reduction in education that results from the high number of AIDS orphans. Young finds that income per capita in the HIV scenario is roughly 10% higher than in the non-HIV scenario in 2010 (15 years after the start of the simulation), and remains higher than the non-HIV scenario for the first 50 years of the simulation.

Ashraf et al. (2009) take a somewhat similar approach, simulating the effect of a generalized health improvement. The specific health intervention (a rise in life expectancy from 40 to 60 years) and its demographic consequences are discussed in Section 3.3.1.3 above. Unlike Young, Ashraf et al. look to existing literature for estimates of the different channels relating health and income, rather than producing their own. Their model allows for several channels by which health affects output. First, there is the demographic effect of increased child survival that is discussed above. Second, there is the direct effect of improved health on worker productivity. This is calibrated in two different ways: first, using the methodology of Weil (2007), and second under the assumption that the Years Lost to Disability, as discussed in Section 3.2.1.1, can also be used to measure the decrement to labor productivity associated with poor health. (The latter methodology yields productivity effects that are about half as large as the former). Third, the authors allow for effects of improved health on education. As in Young's paper, the model has an aggregate production function in which quality-adjusted labor is combined with physical capital (accumulated with a fixed saving rate) and land.

Ashraf et al. find a short-run effect that is consistent with Young's (as well as the findings of Acemoglu and Johnson discussed above): improving health lowers income per capita, primarily through the demographic channel of raising the ratio of dependent children to working age adults. Fifteen years into the simulation, income per capita is 5% lower than it would have been absent the health improvement. In the long run, the demographic effect is undone by endogenously falling fertility, while better health and higher education raise worker productivity, and so the effect on income reverses. Income per capita returns to its baseline level after 30 years, and in the long run is 15% higher thanks to the health improvement. While this long-run finding is in the same direction as empirical papers such as Bloom et al. (2004), the magnitude is far smaller. In addition to using their model to consider a general improvement in health, Ashraf et al. examine reductions in two particular diseases: malaria and tuberculosis. Again, the effects that they find are small. In the case of malaria, calibrating their model to the prevalence of the disease in Zambia, they find that complete elimination would raise income per capita by only 2% in the long run—far below the estimates of, for example, Gallup and Sachs (2001).[30]

3.6. HEALTH AS A COMPONENT OF ECONOMIC GROWTH

The above section discusses how improvements in health lead to increases in conventionally measured GDP. However, as mentioned in Section 3.1, health plays an additional role of being, in itself, a measure of a country's development. This quality is not unique to health, of course. One can make a good argument that education, political freedom, gender equality, and many other social attributes are both themselves aspects of development and contributors to increases in conventionally measured income. But while health is not unique in this sense, it stands out as likely being the most important non-income component that one would want to include in a measure of economic development, for two reasons: first, the fact that individuals clearly assign very high value to a long and healthy life, and second, the large extent to which achievement of this aim varies among countries as well as historically.

The best known metric for combining measures of health and income (and education as well) into a single metric is the Human Development Index (HDI), created by Mahbub ul Haq and Amartya Sen in 1990 with the explicit goal of shifting analysis of economic development away from a focus on income per capita. The HDI is the geometric mean of three "dimension indices," which in turn cover income, life expectancy, and education:[31]

$$HDI_i = I_{Income,i}^{1/3} \times I_{Life,i}^{1/3} \times I_{Education,i}^{1/3}.$$

Each dimension index is in turn defined as:

$$Dimension\ Index_i = \frac{actual\ value_i - minimum\ value}{maximum\ value - minimum\ value}.$$

Income is measured as the log of gross national income (GNI) and life expectancy in years. The minimum values used in both the numerator and denominator (ln(100) and 20 years) are conceived of as subsistence levels, while the maximum in each case is the highest value observed in the sample ($87,478 and 83.6 years in 2012).

[30] Gollin and Zimmermann (2007) also construct a simulation model to study the effects of malaria. They pay particular attention to the behavioral responses of people living in malaria-endemic regions, such as sleeping under bed nets, that may limit the impact of the disease. The endogeneity of malaria prevalence leads to the possibility of multiple history-dependent steady states. Uncontrolled malaria, in the most extreme case, can reduce income per capita by up to half. A large part of the effect in their model is via asset holdings, which in turn determine the capital stock: malaria shortens lifespans and so reduces both the incentive to save and the time over which assets can build up. By contrast, the direct effect of malaria on labor productivity is small, with infected individuals losing only 10% of their labor input.

[31] Malik (2013).

The HDI establishes an equivalence scale relating increases in life expectancy given by the formula to changes in income. Specifically, a rise in income by 1% (one log point) has the same effect on a country's HDI as a rise in life expectancy given by the formula:

$$\frac{(83.6-20)}{(ln(87,478)-ln(100))}\times\frac{I_{Life,i}}{I_{Income,i}}\times 0.01.$$

For example, in the case of Ghana where life expectancy in 2012 was 64.6 years and GNI was $1684, a 1% rise in GNI would have an impact on HDI equivalent to raising life expectancy by 0.16 years. The country with the lowest implied gain in life expectancy equivalent to a 1% rise in income (0.097 years) is the United States (life expectancy of 78.8, GNI of $43,480). At the other end of the spectrum, the country with the highest value (0.24 years) is Eritrea (life expectancy of 62.0, GNI of $581).

3.6.1 A Utility-Based Approach

The HDI is of course somewhat arbitrary in its weighing of different components of development. Recently a number of economists have examined a more theoretically grounded approach toward synthesizing the value of gains in quality and quantity of life. Key papers in this literature include Becker et al. (2005), Murphy and Topel (2006), and Jones and Klenow (2010). All these papers use a similar theoretical structure, which I discuss below.

Murphy and Topel construct a framework for valuing improvements in overall longevity as well as progress against specific diseases. They demonstrate that the value of health gains is larger, the higher is lifetime income (because more utility is derived per year alive) and are also larger, the greater is the existing level of health (because it is more valuable not to die of a particular disease if you are less likely to die of something else). With a calibrated version of the model, they calculate the value of additional life years produced by health improvements in the United States for every decade in the 20th century. They call the value of these improvements "health capital." They find that for the first half of the century, annual investment in health capital was only slightly less than conventionally measured GDP. In other words, almost half of properly measured GDP consisted of investments in health capital. By the last decades of the century, the fraction of properly measured GDP made up of such investments had fallen to roughly 20%. Even in this later period, the total value of health capital gains greatly exceeded medical expenditures.

Becker et al. employ a very similar approach, with a focus on inequality and income convergence among countries. They construct a measure of full income growth that incorporates the value of life expectancy gains in money-metric terms. Looking over the period 1960–2000, they find that in the poorer half of their sample, the annual growth rate of the part of full income that was due to health was 1.7% per year, vs. 0.4% per year in the richest half. Since the two parts of the sample had relatively similar growth rates of GDP, convergence in full income was mostly driven by health. Further, in the poorer half of the sample, about 40% of full income growth was due to longevity improvement—a result that is similar to Murphy and Topel's finding for the US in the first half of the 20th century.

Jones and Klenow also look at cross-country data, examining both levels and growth rates of welfare. In addition to incorporating longevity into their calculations, they adjust their welfare measure for two other factors: within-country inequality of consumption and the average level of leisure. However, their finding is that by far the greatest contributor to welfare differences between rich and poor countries, other than consumption itself, is longevity (looking among rich countries, this is not the case, as longevity does not vary much while inequality and leisure do). The welfare differences that Jones and Klenow find are enormous, even by the standards of cross-country differences in income. For example, the average of income per capita in sub-Saharan Africa in their data is 5.3% of the US level, but the average level of welfare is 1.1% of the US level. Looking at welfare growth over time, their results are only partially consistent with the two papers discussed above. Over the period 1980–2000, longevity growth in the US contributed 1% point to annual welfare growth of 2.7% per year. However, looking across countries, they do not find evidence that convergence of welfare over the period examined greatly exceeded convergence of GDP per capita.

3.6.1.1 Underlying Theory

Health directly affects individual utility both by enhancing the quality of life (holding consumption constant) and by raising the quantity of life. Here I focus solely on the latter channel. The starting point for a theory that combines utility from consumption and length of life is to examine individual choices in which the two are traded off against each other. Consider a person who is faced with the opportunity to avoid taking a small risk to his life in return for a small payment. Let ϵ be the probability of death and x be the payment that makes the individual indifferent. The value of a statistical life (VSL) is defined as:

$$VSL = \frac{x}{\epsilon}.$$

VSL is most commonly estimated by looking at the wage premium associated with jobs that carry extra risk of mortality. The dollar value of marginal improvements in mortality can be directly assessed simply by using estimates of VSL. However, to assess the value of infra-marginal changes in mortality, and to compare changes in mortality to changes in consumption, one needs to impose more structure. A starting point is to assume that VSL is determined by setting equal the expected loss in utility from premature death and the addition utility from consuming x. Labeling V as the expected future utility, we have:

$$\epsilon V = (1 - \epsilon)u'(c)x.$$

In practice, the $(1-\epsilon)$ term on the right-hand side of this equation can be ignored, since VSL is only measured in cases where ϵ is close to zero. The term V incorporates utility from both quality and quantity of life.

To incorporate quantity of life into a parametric utility framework requires that one insert another parameter into the utility function. A simple approach is to simply include a parameter $\bar{u}$, which can be interpreted as the "utility of being alive" that is addition to any utility from consumption. Allowing the consumption component of utility to be of the CRRA form, for example, we have:

$$U = \frac{c^{1-\sigma}}{1-\sigma} + \bar{u},$$

where utility from not being alive is normalized to zero.

To show how VSL can illuminate the relationship between life extension and growth, I examine a greatly simplified version of the model presented in the Murphy-Topel and Becker et al. papers. Following the "perpetual youth" approach of Blanchard (1985), I consider an individual who has constant mortality probability ρ and thus life expectancy of $1/\rho$. He has constant labor income and discounts future utility at rate θ, which is equal to the interest rate. Further, I assume that there is an actuarially fair annuity market. In such a setting, the optimum will be to maintain constant consumption, equal to the wage. Putting all this into the equation above and re-arranging:

$$VSL = \frac{x}{\epsilon} = c^{\sigma}\,\frac{\left(\frac{c^{1-\sigma}}{1-\sigma}\right) + \bar{u}}{\rho + \theta}.$$

In turn, we can solve for the parameter $\bar{u}$ in terms of the value of a statistical life as well as the other, more standard components of the utility function:

$$\bar{u} = VSL \times c^{-\sigma}(\rho + \theta) - \left(\frac{c^{1-\sigma}}{1-\sigma}\right).$$

With these parametric estimates in hand, one can undertake a number of quantitative exercises. The first is to calculate the value of consumption at which individuals are indifferent between being alive or dead. For the value of VSL I use \$4 million, which is broadly consistent with the literature for the United States according to Jones and Klenow.[32] Personal consumption expenditures per capita in the United States in 2012 were approximately \$35,500. I use this figure for c, ignoring issues such as economies of scale in household production and the life cycle pattern of consumption expenditures. I use a value of $\rho = 0.0133$, to give life expectancy of 75 years, and a pure time discount rate of $\theta = 0.02$.

[32] Murphy and Topel use \$6.3 million as the average VSL for adults aged 25–55. As they point out, estimates of VSL generally do not adjust by age – an approach that makes little sense in this utility framework, since an older person who dies is losing out on less utility than a young person. In their calibrated model, VSL falls from \$7 million at age 30 to \$5 million at age 50 and \$2 million at age 70. The perpetual youth model I use here avoids this issue.

Table 3.2 Implications of variations in curvature of the utility function

σ	$\bar{u}$	Break even consumption ($)
0.8	−10.11	34
log	−6.72	830
1.5	0.03055	4286
2	0.000134	7465

The results of this exercise are very sensitive to the curvature of the utility function. Table 3.2 shows the implied value of $\bar{u}$ for different values σ, inverse of the intertemporal elasticity of substitution (Both Becker et al. and Murphy and Topel use 0.8 as their preferred value. Jones and Klenow use 1.0.) Since it is measured in utility terms, $\bar{u}$ itself is not very meaningful. The third column of the table shows the level of consumption at which individuals are indifferent between being alive or dead (labeled "break even" consumption), which can be seen to vary enormously with the value of σ. For some values of σ that would be considered empirically reasonable, the level of consumption at which life is not worth living is quite high.

I use this framework to carry out two exercises. The first is to re-visit the equivalence between increases in consumption and increases in life expectancy. Again, I consider the increase in life expectancy that provides increased utility equal to a 1% increase in c. This is derived by differentiating lifetime utility with respect to $ln(c)$ and with respect to $(1/\rho)$ and taking their ratio. The formula is:

$$\textit{gain in life expectancy} = \frac{c^{1-\sigma}\left(1+\frac{\theta}{\rho}\right)^2}{(\rho+\theta)\left(\frac{c^{1-\sigma}}{1-\sigma}+\bar{u}\right)} \times .01.$$

I show the tradeoff at values equal to the US level, and then one-half, one-quarter, one-eighth, and one-sixteenth, and one thirty-second of that level. In each case, I calculate the rise in life expectancy (in years) that is equivalent to a 1% increase in the consumption measure c. (The entire exercise is conducted holding initial life expectancy constant at its US level of 75 years. Obviously, poor countries also have lower life expectancies than rich. However, in this setting, the effects of these differences are relatively muted.) I conduct the exercise for the same four values of σ considered above. Table 3.3 shows the results. The first column shows that for the parameters used in the calibration, people in the United States are made equally well-off by an increase in consumption of 1% and a gain in life expectancy of half a year. Compared to the HDI calculations discussed above, then, the model here weighs life relatively less. However, this is largely a matter of parameterization. For example, raising the value of a statistical life in the US to around

Table 3.3 Gain in life expectancy equivalent to 1% rise in consumption

	Consumption					
σ	35,000	17,750	8,875	4,437.5	2,218.75	1,109.38
0.8	0.499	0.525	0.558	0.602	0.662	0.747
1 (log)	0.499	0.612	0.791	1.12	1.91	6.46
1.5	0.499	0.906	2.14	53.6	–	–
2	0.499	1.36	9.92	–	–	–

$20 million would set the implied gain in life expectancy equivalent to a 1% rise in consumption in the US equal to the value in the HDI. The more important results in Table 3.3 have to do with the variation in outcomes as income changes. One implication is that as a country grows richer, people are willing to give up more income in return for a given increment in health (this is the reciprocal of the number shown in the table). This effect is stressed by Hall and Jones (2007) as an explanation for increases in health spending as countries get richer: the marginal utility of consumption within a year declines as consumption rises, but the marginal utility of extra life years does not decline as life gets longer.[33] As the table also shows, this effect is magnified, the more curved is the within-period utility function, that is, the larger is the value of σ. In their baseline quantitative analysis, which forecasts that health spending in the United States could reach 30% of GDP by the middle of the 21st century, Hall and Jones use a value of $\sigma = 2$.[34] I return to other implications of the table in a moment.

The other, related exercise is to calculate the ratio VSL/c, which can be thought of as the value of a statistical life relative to the individual's ability to pay. It is not surprising that VSL rises with income, because richer people have more money to spend on everything. By contrast, the behavior of VSL/c gives more insight into the underlying economics. The values of this ratio are presented in Table 3.4 for the same values of σ and c that were considered in the previous table.

These two tables convey an interesting point. In this framework, the value that individuals place on extra years of life relative to income is strongly dependent on the level of income. As income falls, people raise the gain in life expectancy that is equivalent to an income increase, and similarly, they lower the ratio of the value of a statistical life to consumption. The reason, in both cases, is that according to the model, people who are significantly poorer than Americans get relatively little utility per period alive. Thus they value additions to consumption (raising the utility per year) far more than they value adding extra years of life. As the tables show, these effects are exacerbated for higher

[33] This result holds for standard, additively separable preferences. Bommier (2006) presents an alternative, intuitively appealing approach that allows for decreasing returns to lifetime as well.

[34] In addition to the curvature of the utility function, the optimal share of spending on health in their model depends critically on the elasticity of health status with respect to health status, which they estimate declines in the level of health spending.

Table 3.4 Ratio of value of statistical life to annual consumption

	Consumption					
σ	**35,000**	**17,750**	**8,875**	**4,437.5**	**2,218.75**	**1,109.38**
0.8	112.7	107.1	100.8	93.4	85.0	75.4
1 (log)	112.7	91.9	71.1	50.3	29.5	8.7
1.5	112.7	62.1	26.3	1.0	–	–
2	112.7	41.3	5.7	–	–	–

values of the σ, the inverse of the intertemporal elasticity of substitution. Indeed, many of the results in the tables seem downright crazy. For example, assuming log utility, the model implies that a person in a country with income per capita equal to 1/32 of the US level would be indifferent between increasing consumption by 1% and raising life expectancy by six years. Taking this model seriously would give the policy implication that aid to the poorest countries should be aimed at raising consumption far more than toward saving lives. However, the model clearly has something wrong with it—a point to which I return below.

3.6.1.2 Compensating and Equivalent Variations

The above analysis considers marginal changes in consumption and life expectancy. To evaluate non-marginal changes in life expectancy and consumption, authors in this literature have used the mechanisms of compensating and equivalent variation.[35] Denote life expectancy as e and expected lifetime utility as $V(e, c)$. Consider a comparison of two countries (or a single country at two points in time), denoted a and b, where a will serve as the benchmark. The equivalent variation measure asks how much consumption would have to be adjusted downward in country a such that expected utility in the two countries was equal:

$$V(e_a, \lambda_{ev} c_a) = V(e_b, c_b).$$

The compensating variation measure, in contrast, asks how much consumption has to rise in country b in order to set expected utility equal in the two cases:

$$V(e_a, c_a) = V\left(e_b, \frac{c_b}{\lambda_{cv}}\right).$$

Using the model of perpetual youth presented above, we can solve explicitly for both of these:

$$\lambda_{ev} = \frac{1}{c_a}\left(\left[\left(\frac{\rho_a + \theta}{\rho_b + \theta}\right)\left(\frac{c_b^{1-\sigma}}{1-\sigma} + \bar{u}\right) - \bar{u}\right](1-\sigma)\right)^{\frac{1}{1-\sigma}},$$

[35] This treatment closely follows Jones and Klenow.

$$\lambda_{cv} = c_b \left(\left[\left(\frac{\rho_b + \theta}{\rho_a + \theta} \right) \left(\frac{c_a^{1-\sigma}}{1-\sigma} + \bar{u} \right) - \bar{u} \right] (1-\sigma) \right)^{\frac{1}{\sigma-1}} .$$

Important differences between the two measures arise when the level of flow utility (that is, utility from consumption plus $\bar{u}$, the utility from being alive) is near zero in a poor country. For example, consider the comparison of the United States and Zambia. Using data from the Human Development Report, GNI in the two countries is $43,480 and $1,358, respectively, while life expectancy is 78.8 and 49.4. The ratio of GNI (which I use as a proxy for consumption) in Zambia to that in the US is 3.1%. Assuming log utility, the value of λ_{ev} is 2.8%, reflecting only a small adjustment for the mortality gap: since life is barely worth living in Zambia, according to this calculation, the additional loss that would be incurred by someone from the US in switching to Zambian consumption and life expectancy (rather than just Zambian consumption) is relatively small. By contrast, the value of λ_{cv} is 1.3%, reflecting the fact that in order to give a Zambian lifetime utility equal to someone from the US, his annual flow utility would have to be raised enough to compensate for his lower life expectancy.[36]

3.6.1.3 Variation in the Value of a Statistical Life Across Countries

Many of the problems in the above framework can be related to a single cause: the use of the valuation of a statistical life in the United States to impute a value of $\bar{u}$, the utility of being alive, that is then imported to other countries or time periods. People in the United States behave in a manner that suggests that they would rather be dead than consume at a level that characterizes many people in the developing world, but there is little reason to think that many people in developing countries feel the same way.

Direct evidence on *VSL* bears out this prediction. Cordoba and Ripoll (2013) examine data from Viscusi and Aldy (2003) on measures of *VSL* in a scattering of countries at different income levels, as well as *VSL*s of different income groups within the United States. Their analysis of the data leads them to conclude that the ratio of VSL/c is actually falling in the level of consumption, although looking at their data it seems equally reasonable to conclude that the ratio of *VSL* to consumption simply does not vary with consumption. In either case, however, the implication of the standard

[36] Jones and Klenow get much larger differences between CV and EV. In the most extreme case of Malawi, the two differ by a factor of 17. One reason that they get such large differences is that in their formulation there is no pure time discount factor, which leads to a larger effect of life expectancy on expected lifetime utility. To give an example, consider countries with life expectancies of 100 and 50 years. In the Jones-Klenow setup, the CV measure will increase in the low life expectancy country's consumption such that flow utility is twice as high as in the country with high life expectancy. By contrast, in the model I present, with a time discount rate of 2%, the CV measure will increase flow utility in the low life expectancy country to be one and one third times as large as in the high life expectancy country. In practice, Jones and Klenow present the geometic average of λ_{ev} and λ_{cv} as their main result, but they note that most of their conclusions hold using either measure alone.

model presented above, which is that VSL/c should be rising with the level of consumption, seems to be soundly rejected. Further, as Cordoba and Ripoll note, there is no evidence that poor people have negative values of VSL, as the standard theory predicts.

How can we reconcile these observations with the implications of the tried-and-true utility model? Cordoba and Ripoll propose to solve the problem by looking at a non-expected utility model, in which the coefficient of relative risk aversion is decoupled from the intertemporal elasticity of substitution. In their model, individuals have a high level of risk aversion toward the state of the world in which they are not alive, but a relatively high intertemporal elasticity of substitution. The specific mechanism that gets the result that VSL/c decreases as countries get richer runs through life expectancy, which in the data is correlated with income. In their view, the marginal willingness to pay for an extra chance of survival decreases with the probability of survival; in other words, a person will pay more to raise their chances of surviving from 5% to 6% than from 95% to 96%. Cordoba and Ripoll also note that their model matches reality better than the standard model in another dimension, specifically the preference of individuals for late rather than early resolution of uncertainty regarding risk of mortality.

An alternative approach to explaining the behavior of VSL across countries is proposed by Prinz and Weil (2013), who ground their approach in a simple model of habit formation, along the lines of Carroll et al. (2000). Consider an individual with instantaneous utility function:

$$u(c) = \frac{\left(\frac{c}{z^{\gamma}}\right)^{1-\sigma}}{1-\sigma} + \bar{u}, \tag{3.10}$$

where z denotes habitual consumption and $0 \leq \gamma \leq 1$ denotes the degree of importance of habit formation. In the case of "external habits," z is determined by the average level of consumption in a country. An individual contemplating a small risk to his life in return for a small monetary benefit will take the value of z as fixed. Thus, for example, the values of "break even" consumption in Table 3.2 at which a person would be equally happy dead or alive, calculated based on the observed VSL in the United States, are correct for someone with the US stock of habits. However, a person in a poor country, with a lower stock of habits, would have higher utility, and thus be happier alive than dead, at these same consumption levels.

For a given value of γ, and under the assumption that within a country z is equal to c, one can calculate $\bar{u}$ as well as the other quantities derived above, such as the trade-off between increased life expectancy and consumption. To give an example of how the Prinz-Weil approach leads to more sensible values for VSL in poor countries, I repeat the exercise of Table 3.4, looking at the ratio of VSL to annual consumption, assuming habit formation of $\gamma = 0.5$. Table 3.5 shows the results. Unlike the original version of the table, the ratio of VSL to consumption rises far more modestly with

Table 3.5 Ratio of value of statistical life to annual consumption with $\gamma = 0.5$

	Consumption					
σ	**35,000**	**17,750**	**8,875**	**4,437.5**	**2,218.75**	**1,109.38**
0.8	112.7	110.1	107.1	104.0	100.8	97.2
1 (log)	112.7	102.3	91.9	81.5	71.1	60.7
1.5	112.7	85.2	62.1	42.7	26.3	12.6
2	112.7	70.9	41.3	20.4	5.7	–

income, and there are fewer cases where VSL is negative.[37] Allowing for habit formation, the ratio of VSL to consumption for a country with $\frac{1}{32}$ of US consumption, in the case of log utility, is 61. The corresponding ratio without habit formation is 8.7.

This being said, however, the Prinz-Weil approach cannot, at least by itself, explain the observation that Cordoba and Ripoll make, that VSL/consumption does not vary at all with income, unless one is willing to make the extreme claim that the degree of habit formation is one. This extreme case would imply that people in poor countries are just as happy with their level of consumption, adjusted for habits, as people in rich countries. An alternative explanation is that the values of VSL/consumption observed in the data are partially a result of the habit formation effect and partly a result of something else, for example, higher expected consumption growth in poor than rich countries.

3.7. CONCLUSION

Income and health are strongly correlated. Looking across countries, higher income per capita is correlated not only with life expectancy, but with numerous other measures of health status. Within countries, there is also a strong correlation between an individual's place in the income distribution and his or her health outcomes. This within-country correlation is particularly strong in developing countries.

Comparing growth of income with improvements in health outcomes, things are a bit more complicated. In the short run, there is at best a weak correlation between changes in income and changes in life expectancy. Indeed, there are many examples of dramatic improvements in health taking place in the absence of notable income growth, and similarly of episodes of rapid income growth that are not accompanied by health improvements. On the other hand, we know that prior to the Industrial Revolution levels of income and life expectancy were roughly the same throughout the world while today the two are strongly correlated, and further that the pattern of initial divergence

[37] The formula for $\bar{u}$ is:

$$\bar{u} = c^{-\sigma-\gamma+\sigma\gamma}(\rho+\theta)VSL - \frac{c^{(1-\sigma)(1-\gamma)}}{1-\sigma}.$$

and later catch-up on the two series look similar. All these facts suggest that in the very long run income growth and health improvement are indeed correlated.

As is often the case in economics, the observation that income and health are correlated, is only the beginning of the discussion. Such a correlation can be induced by causation running in either direction, as well as by the effects of some third factor. A priori, there are good reasons to think that all of these are possibilities. People who are healthier can work harder and learn more in school; and where people live longer they will be incentivized to invest more in education. Thus, we would expect better health to cause economic growth. On the other hand, higher income allows individuals or governments to make investments that yield better health. Finally, differences in the quality of institutions (looking across countries), in human capital (looking across individuals), or in the level of technology (looking over time) can induce correlated movements in health and income. Further complicating the inference problem are the dynamic effects built into many of the potential causal channels. For example, improvements in health may only result in increased worker productivity with a lag of several decades. Similarly, when life expectancy rises there can be increases in population growth that may temporarily reduce income per capita.

The causal relation that has been most widely studied by researchers in this area is the effect of health improvements on economic growth at the country level. This is an issue with direct policy relevance. If improving health leads to growth, this would be a reason, beyond the welfare gain from better health itself, that governments might want to make such investments. However, the evidence for such an effect of health on growth is relatively weak. Cross-country empirical analyses that find large effects for this causal channel tend to have serious identification problems. The few studies that use better identification find small or even negative effects. Theoretical and empirical analyses of the individual causal channels by which health should raise growth find positive effects, but again these tend to be fairly small. Putting the different channels together into a simulation model shows that potential growth effects of better health are only modest, and arrive with a significant delay.

Regarding causality running from income to health, at least at the level of countries, there is also little evidence of much effect in the short run. For developing countries, there exists a large stock of health technologies that can be applied to great effect at low cost. Political will and institutional efficiency are more important than GDP in determining health. Looking across individuals, it is harder to sort out the extent to which it is knowledge of health improving behaviors or economic wherewithal (which is correlated with human capital) that is more important in contributing to the correlation between health and income. Possibly, this even differs as a function of level of economic development (as does the effect of health on income at the individual level).

In the short run, then, at least as regards differences among countries, one is forced to the conclusion that the strong relationship between income and health is a product of some other factors. The same political will and institutional efficiency that lead to better health also lead to higher income, most of the time, but with some important exceptions.

Looking at historical changes, however, the picture is different. It is hard to escape the conclusion that in the long run, improvements in health have indeed been the result of economic growth. It is not hard to identify the scientific discoveries, medical advances, and public health initiatives that have produced enormous health gains in the most advanced countries. These achievements seem unlikely to have occurred outside the context of industrialization. As a counterfactual, it is possible to imagine a history in which economic growth (technological advance; accumulation of physical and human capital; institutional change; and so on) took place roughly as we have observed it, but in which life expectancy and other measures of health remained stuck at their 18th-century levels. But it is not similarly possible (at least for me) to imagine a history in which knowledge regarding health advanced and was implemented as it has been in the absence of economic growth.

In contrast to the uncertainty about causality, analysis of the welfare effects of health improvements is much more straightforward: they are very large. Depending on the period being examined, the welfare gain from better health may be as large or larger than the welfare gain from rising consumption.

ACKNOWLEDGMENTS

I am grateful to Daniel Prinz, Sanjay Singh, Lindor Qunaj, Alex Drechsler, Young Min Kim, and Yunan Ji for research assistance, and to Lars Lonstrop for comments.

REFERENCES

Acemoglu, D., Johnson, S., 2007. Disease and development: the effect of life expectancy on economic growth. Journal of Political Economy 115 (6), 925–985.

Acemoglu, D., Johnson, S., 2013. Disease and development: a reply to Bloom, Canning and Fink. Mimeograph.

Aghion, P., Howitt, P., 1992. A model of growth through creative destruction. Econometrica 60 (2), 323–351.

Ahuja, A., Wendell, B., Werker, E.D., 2007. Male circumcision and AIDS: the macroeconomic impact of a health crisis. Working Paper 07–025, Harvard Business School.

Almond, D., 2006. Is the 1918 influenza pandemic over? long-term effects of in utero influenza exposure in the post-1940 US population. Journal of Political Economy 114 (4), 672–712.

Almond, D., Currie, J., 2011. Killing me softly: the fetal origins hypothesis. Journal of Economic Perspectives 25 (3), 153–172.

Almond, D., Chay, K.Y., Lee, D.S., 2005. The costs of low birth weight. Quarterly Journal of Economics 120 (3), 1031–1083.

Alsan, M., 2012. The effect of the tsetse fly on african development. mimeograph.

Ashraf, Q.H., Lester, A., Weil, D.N., 2009. When does improving health raise GDP? In: NBER Macroeconomics Annual 2008, vol. 23. National Bureau of Economic Research, pp. 157–204 (August).

Ashraf, Q., Weil, D.N., Wilde, J., 2013. The effect of fertility reduction on economic growth. Population and Development Review 39 (1), 97–130.

Baten, J., Crayen, D., Voth, H.-J., forthcoming. Numeracy and the impact of high food prices in industrializing Britain, 1780–1850. Review of Economics and Statistics.

Becker, G.S., Philipson, T.J., Soares, R.R., 2005. The quantity and quality of life and the evolution of world inequality. American Economic Review 95 (1), 277–291.

Behrman, J.R., Rosenzweig, M.R., 2004. Returns to birthweight. Review of Economics and Statistics 86 (2), 586–601.

Ben-Porath, Y., 1967. The production of human capital and the life cycle of earnings. Journal of Political Economy 75, 352–365.

Bils, M., Klenow, P. J., 2000. Does schooling cause growth? American Economic Review 90 (5), 1160–1183.

Black, S.E., Devereux, P.J., Salvanes, K.G., 2007. From cradle to the labor market? the effect of birth weight on adult outcomes. Quarterly Journal of Economics 122 (1), 409–439.

Blanchard, O.J., 1985. Debt, deficits, and finite horizons. The Journal of Political Economy 93 (2), 223–247.

Bleakley, H., 2007. Disease and development: evidence from hookworm eradication in the American south. Quarterly Journal of Economics 122 (1), 73–117.

Bleakley, H., 2010. Malaria eradication in the americas: a retrospective analysis of childhood exposure. American Economic Journal: Applied Economics 2 (2), 1–45.

Bleichrodt, N., Born, M., 1994. A meta-analysis of research on iodine and its relationship to cognitive development. In: The damaged brain of iodine defficiency. Cognizant Communication, pp. 195–200.

Bloom, D.E., Canning, D., Graham, B., 2003. Longevity and life cycle savings. Scandinavian Journal of Economics 105 (3), 319–338.

Bloom, D.E., Canning, D., Sevilla, J., 2004. The effect of health on economic growth: a production function approach. World Development 32 (1), 1–13.

Bloom, D.E., Canning, D., Fink, G., forthcoming. Disease and development revisited. Journal of Political Economy.

Bommier, A., 2006. Uncertain lifetime and intertemporal choice: risk aversion as a rationale for time discounting. International Economic Review 47 (4), 1223–1246.

Caldwell, J.C., 1986. Routes to low mortality in poor countries. Population and Development Review 12 (2), 171–220.

Carroll, C.D., Overland, J., Weil, D.N., 2000. Saving and growth with habit formation. American Economic Review 341–355.

Case, A., Paxson, C., 2010. Causes and consequences of early-life health. Demography 47 (1), S65–S85.

Case, A., Lubotsky, D., Paxson, C., 2002. Economic status and health in childhood: The origins of the gradient. American Economic Review 92 (5), 1308–1334.

Case, A., Fertig, A., Paxson, C., 2005. The lasting impact of childhood health and circumstance. Journal of Health Economics 24, 365–389.

Caselli, F., 2005. Accounting for cross-country income differences. In: Aghion, P., Durlauf, S.N. (Eds.), Handbook of Economic Growth, vol. 1. North-Holland, pp. 679–742.

Cervellati, M., Sunde, U., forthcoming. Life expectancy, schooling, and lifetime labor supply: Theory and evidence revisited. Econometrica.

Cordoba, J.C., Ripoll, M., 2013. Beyond expected utility in the economics of health and longevity. working paper 13008, Iowa State University.

Costa, D., Steckel, R., 1997. Long-term trends in health, welfare, and economic growth in the United States. University of Chicago Press pp. 47–90.

Currie, J., Vogl, T., 2013. Early-life health and adult circumstances in developing countries. Annual Review of Economics 5, 1–36.

Cutler, D.M., Meara, E., 2004. Changes in the age distribution of mortality over the twentieth century. In: Wise, D.A. (Ed.), Perspectives on the Economics of Aging. University of Chicago Press, pp. 333–365.

Cutler, D., Miller, G., 2005. The role of public health improvements in health advances: the twentieth-century United States. Demography 42 (1), 1–22.

Cutler, D., Deaton, A., Lleras-Muney, A., 2006. The determinants of mortality. Journal of Economic Perspectives 20 (3), 97–120.

Cutler, D., Fung, W., Kremer, M., Singhal, M., Vogl, T., 2010. Early-life malaria exposure and adult outcomes: evidence from malaria eradication in India. American Economic Journal: Applied Economics 2 (2), 72–94.

Daley, T.C., et al. 2003. IQ on the rise: the Flynn effect in rural Kenyan children. Psychological Science 14, 215–219.

Deaton, A.S., 2003. Health, inequality, and economic development. Journal of Economic Literature XLI, 113–158.

Deaton, A., 2006. The great escape: a review of Robert Fogel's the escape from hunger and premature death, 1700–2100. Journal of Economic Literature 44 (1), 106–114.

Deaton, A.S., 2007. Height, health, and development. PNAS 104 (33), 13232–13237.

Deaton, A., Dreze, J., 2009. Food and nutrition in india: facts and interpretations. Economic and Political Weekly XLIV 7, 42–65.

Deaton, A.S., Paxson, C., 2001. Mortality, Education, Income, and Inequality among American Cohorts. University of Chicago Press, pp. 129–170.

De la Croix, D., Licandro, O., 2012. The longevity of famous people from Hammurabi and Einstein. Discussion Paper 52.

Easterly, W., 1999. Life during growth. Journal of Economic Growth 4 (3), 239–275.

Eppigg, C., Fincher, C.L., Thornhill, R., June 2010. Parasite prevalence and the worldwide distribution of cognitive ability. Proceedings of the Royal Society of Biological Sciences 277, 3801–3808.

Eveleth, P., Tanner, J., 1990. Worldwide Variation in Human Growth, second ed. Cambridge University Press.

Floud, R., Fogel, R., Harris, B.H., Hong, S., 2011. The Changing Body: Health, Nutrition, and Human Development in the Western World since 1700. Cambridge University Press.

Flynn, J.R., 1987. Massive IQ gains in 14 nations: what IQ tests really measure. Psychological Bulletin 101, 171–191.

Fogel, R.W., 1994. Economic growth, population theory, and physiology: the bearing of long-term processes on the making of economic policy. American Economic Review 84 (3), 369–395.

Fogel, R. W., and Engerman, S., 1992. The slave diet on large plantations in 1860. In R. W. Fogel, R. Galantine, R. Manning, eds., Without Consent or Contract: Evidence and Methods. W.W. Norton, New York.

Fogel, R.W., 1997. Secular trends in nutrition and mortality. In: Rosenzweig, M., Stark, O. (Eds.), Handbook of Population and Family Economics, vol. 1A. Elsevier, pp. 433–481.

Gallup, J.L., Sachs, J., 2001. The economic burden of malaria. American Journal of Tropical Medicine and Hygiene 64 (1), 85–96.

Galor, O., Weil, D.N., 1996. The gender gap, fertility, and growth. American Economic Review 86 (3), 374–387.

Gollin, D., Zimmermann, C., 2007. Malaria: Disease impacts and long-run income differences. Working papers 2007–30, University of Connecticut, Department of Economics.

Gwatkin, D.R., Rutstein, S., Johnson, K., Suliman, E., Wagstaff, A., Amouzou, A., 2007. Socioeconomic differences in health, nutrition, and population, within developing countries. World Bank, Washington, DC.

Haines, M.R., 2001. The urban mortality transition in the United States, 1800–1940. Annales de demographie historique 101, 33–64.

Hall, R.E., Jones, C.I., 2007. The value of life and the rise in health spending. The Quarterly Journal of Economics 122 (1), 39–72.

Hazan, M., 2009. Longevity and lifetime labor supply: evidence and implications. Econometrica 77 (6), 1829–1863.

Hazan, M., Zoabi, H., 2006. Does longevity cause growth? a theoretical critique. Journal of Economic Growth 11, 363–376.

Hwang, J.-Y., Shin, C., Frongillo, E.A., Shin, K.R., Jo, I., 2003. Secular trend in the age at menarche for South Korean women born between 1920 and 1986: the Ansan study. Annals of Human Biology 434–442.

Jayachandran, S., Lleras-Muney, A., 2009. Life expectancy and human capital investments: evidence from maternal mortality declines. Quarterly Journal of Economics 124 (1), 349–397.

Jones, C.I., 1995. R&D-based models of economic growth. Journal of Political Economy 103 (4), 759–784.

Jones, C.I., Klenow, P.J., 2010. Beyond GDP? welfare across countries and time. NBER Working Papers 16352, National Bureau of Economic Research.

Jones, G., 2011. IQ and national productivity. In: Durlauf, S.N., Blume, L.E. (Eds.), New Palgrave Dictionary of Economics Online Edition.

Kalemli-Ozcan, S., 2002. Does the mortality decline promote economic growth? Journal of Economic Growth 7 (4), 411–439.

Kalemli-Ozcan, S., Weil, D.N., 2010. Mortality change, the uncertainty effect, and retirement. Journal of Economic Growth 15 (1).

Kalemli-Ozcan, S., Ryder, H.E., Weil, D.N., 2000. Mortality decline, human capital investment, and economic growth. Journal of Development Economics 62 (1), 1–23.

Kiszewski, A., Mellinger, A., Spielman, A., Malaney, P., Sachs, S.E., Sachs, J., 2004. A global index representing the stability of malaria transmission. American Journal of Tropical Medicine and Hygiene 70 (5), 486–498.

Knaul, F.M., 2000. Health, Nutrition, and Wages: Age at Menarche and Earnings in Mexico. Inter-American Development Bank.

Li, N., Gerland, P., 2011. Modifying the Lee-Carter method to project mortality changes up to 2100. Paper Presented at the 2011 Annual Meeting of the Population Associatin of America.

Livi-Bacci, M. 1997. A Concise History of World Population. John Wiley and Sons, Hoboken, NJ.

Lorentzen, P., McMillan, J., Wacziarg, R., 2008. Death and development. Journal of Economic Growth 13 (2), 81–124.

Lucas, R., 2000. Some macroeconomiocs for the 21st century. Journal of Economic Perspectives 14 (1), 159–168.

Lucas, A.M., 2010. Malaria eradication and educational attainment: evidence from Paraguay and Sri Lanka. American Economic Journal: Applied Economics 2 (2).

Lynn, R., 1998. In Support of the Nutrition Theory. American Psychological Association, Washington, DC, pp. 207–218.

Maddison, A., 2001. The World Economy: A Millennial Perspective. OECD, Paris, France.

Malik, K., 2013. Human development report 2013 the rise of the south: Human progress in a diverse world. technical notes. Tech. Rep., United Nations, Development Programme).

Mankiw, N.G., Romer, D., Weil, D.N., 1992. A contribution to the empirics of economic growth. Quarterly Journal of Economics 107 (2), 407–437.

Martorell, R., 1998. Nutrition and the Worldwide Rise in IQ Scores. American Psychological Association, Washington, DC, pp. 183–206.

Mathers, C., Fat, D.M., Boerma, J.T., 2008. The global burden of disease: 2004 update.

McGuire, R.A., Coelho, P.R., 2011. Parasites, pathogens, and progress: diseases and economic development. MIT Press.

McKeown, T., 1976. The Modern Rise of Populations. Academic Press, New York.

McNeill, W., 1998. Plagues and Peoples. Anchor, New York.

Murphy, K.M., Topel, R.H., 2006. The value of health and longevity. Journal of Political Economy 114 (5), 871–904.

Oeppen, J., Vaupel, J.W., 2002. Broken limits to life expectancy. Science 296, 1029–1031.

Oster, E., 2012. HIV and sexual behavior change: why not Africa? Journal of Health Economics 31 (1), 35–49.

Oster, E., Shoulson, I., Dorsey, R., forthcoming. Limited life expectancy, human capital and health investments. American Economic Review.

Peltzman, S., 2009. Mortality inequality. Journal of Economic Perspectives 23 (4), 175–190.

Preston, S.H., 1975. The changing relation between mortality and level of economic development. Population Studies 29 (2), 231–248.

Preston, S.H., 1996. American longevity: Past, present, and future. Tech. Rep. Policy Brief No 7, Center for Policy Research, Maxwell School, Syracuse University.

Preston, S., Haines, M., 1991. The Fatal Years: Child Mortalitiy in the Late Nineteenth Century. Princeton University Press, Princeton, NJ.

Prinz, D., Weil, D.N., 2013. Habit formation and the value of a statistical life. mimeo, Brown University.

Pritchett, L., Summers, L.H., 1996. Wealthier is healthier. Journal of Human Resources 31 (4), 841–868.

Ribero, R., Nunez, J., 2000. Adult morbidity, height, and earnings in Columbia. In: Savedoff, W.D., Schultz, T.P. (Eds.), Wealth from Health: Linking Social Investments to Earnings in Latin America. Inter-American Development Bank.

Riley, J. C., 2005. Estimates of regional and global life expectancy, 1800-2001. Population and Development Review 31 (3), 537-543.

Sachs, J.D. (Ed.), 2001. Macroeconomics and Health: Investing in Health for Economic Development. World Health Organization, Geneva, Switzerland.

Sachs, J.D., 2003. Institutions don't rule: Direct effects of geography on per capita growth. Working Paper 9490, National Bureau of Economic Research.

Sahn, D.E., Stifel, D.C., 2003. Urban–rural inequality in living standards in Africa. Journal of African Economies 12 (4), 564–597.

Schultz, T.P., 2002. Wage gains associated with height as a form of health human capital. American Economic Review 92 (2), 349–353.

Schultz, T.P., 2005. Productive Benefits of Health: Evidence from Low-Income Countries. MIT Press.

Schultz, T.P., 2010. Health human capital and economic development. Journal of African Economies 19(3), iii12–iii80.

Shastry, G.K., Weil, D.N., 2003. How much of cross-country income variation is explained by health? Journal of the European Economic Association 1 (2–3), 387–396.

Sigman, M., Whalley, S.E., 1998. The Role of Nutrition in the Development of Intelligence. American Psychological Association, Washington, DC, pp. 155–182.

Smith, R., 2013. Longevity changes and their determinants in England and her European neighbours c. 1600–1900. Mimeo, University of Cambridge, Department of Geography.

Soares, R., 2005. Mortality reductions, educational attainment, and fertility choice. American Economic Review 95 (3), 580–601.

Soares, R., 2007. On the determinants of mortality reductions in the developing world. Population and Development Review 33 (2), 247–287.

Sohn, B., 2000. Health, nutrition, and economic growth. Ph.D. thesis, Brown University.

Steyn, M., 2003. A comparison between pre- and post-colonial health in the northern parts of South Africa, a preliminary study. World Archaeology 35 (2), 276-288.

Subramanian, S., Ozaltin, E., Finlay, J., 2011. Height of nations: A socioeconomic analysis of cohort differences and patterns among women in 54 low- and middle-income countries. PLoS One 6 (4), 1–13.

Thomas, D., Frankenberg, E., 2002. Health, nutrition, and prosperity: a microeconomic perspective. Bulletin of the World Health Organization 80 (2), 106–113.

Viscusi, W.K., Aldy, J.E., 2003. The value of a statistical life: a critical review of market estimates throughout the world. Journal of Risk and Uncertainty 27 (1), 5–76.

Vogl, T., 2012. Height, skills, and labor market outcomes in mexico. Working Paper 18318, National Bureau of Economic Research.

Weil, D.N., 2007. Accounting for the effect of health on economic growth. The Quarterly Journal of Economics 122 (3), 1265–1306.

Weil, D.N., 2008. Population aging. In: Durlauf, S.N., Blume, L.E. (Eds.), New Palgrave Encyclopedia of Economics, second ed.

Young, A., 2005. The gift of the dying: The tragedy of aids and the welfare of future African generations. Quarterly Journal of Economics 120 (2), 423–466.

CHAPTER FOUR

Regional Growth and Regional Decline

Holger Breinlich[*]**, Gianmarco I.P. Ottaviano**[†]**,**
and Jonathan R.W. Temple[‡]
[*]University of Essex, CEP and CEPR
[†]London School of Economics, CEP and CEPR
[‡]University of Bristol and CEPR

Abstract

Since the early 1990s, there has been a renaissance in the study of regional growth, spurred by new models, methods, and data. We survey a range of modeling traditions, and some formal approaches to the hard problem of regional economics; namely, the joint consideration of agglomeration and growth. We also review empirical methods and findings based on natural experiments, spatial discontinuity designs, and structural models. Throughout, we give considerable attention to regional growth in developing countries. Finally, we highlight the potential importance of processes that are specific to regional decline, and which deserve greater research attention.

Keywords

Regional growth, Regional convergence, Regional decline

JEL Classification Codes

O18, R11, R12, R13

> *Europe, as it has become more integrated, has also become more difficult to write about.*
>
> ***Perry Anderson, "The New Old World", p. xi***

4.1. INTRODUCTION

From 2006 onwards, an exhibition of photographs has toured galleries in Europe and the USA, now titled *The Ruins of Detroit*. The photographs, by Yves Marchand and Romain Meffre, show various scenes from the recent past of America's Motor City: the ruined Spanish-Gothic interior of the United Artists Theater (closed 1984), the abandoned waiting hall of Michigan Central Station (closed 1988), the derelict ballroom of the Lee Plaza Hotel (closed early 1990s), and an abandoned school book depository with its textbooks scattered and covered in debris. The photographs bring home, in a way that statistics do not, what it can mean for a city or region to experience an extended period of decline. In Detroit's case, that decline has been precipitous: from one of America's

ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53540-5.00004-5

wealthiest cities, with a city population of around 1.8 million at its peak, to a population that is now around 700,000. It has left the city with falling property values, enough vacant land to accommodate the whole of Paris, and a rate of violent crime among the highest of any American city. Economists sometimes refer to changing patterns of economic activity as "adjustment," but many of those who have lived through the city's decline will have experienced it chiefly as a tragedy.

One reason Detroit's experience has attracted such attention is that relative decline is rarely so marked, or rapid. The disparities between cities and regions are generally more stable than this, even at times of growth and structural change. But it is also true that disparities can be substantial and persistent, lasting many decades. They can often become an important part of how a country sees itself, and how it evolves over time. As Judt (1996) notes, the divisions and tensions between southern and northern Italy are a theme as old as the Italian state itself. In England, the 19th-century writer Elizabeth Gaskell published her novel *North and South* in 1855. More than 150 years later, regional differences in living standards, health outcomes, political beliefs, and social norms continue to be summarized as England's north-south divide. Similar phenomena can be seen in the developing world, sometimes on an even larger scale. China's coastal cities are more prosperous than its inland regions. In India, there are substantial disparities across states in terms of literacy rates, life expectancy and living conditions, as well as income (Drèze and Sen, 1997). Poverty rates vary widely within Brazil, with low rates in the booming south-east and much higher rates in the rural north-east (Skoufias and Katayama, 2011). The list could easily be multiplied but, as Williamson (1965) remarks, many countries have a tendency to see their own regional imbalances as uniquely pronounced and intractable.

This chapter will describe a range of models and evidence that can be used to understand regional growth and the evolution of spatial disparities over time. This is an unusually complex topic, and one that requires an eclectic approach and general equilibrium reasoning. In a long-run spatial equilibrium, households and firms must not prefer other locations to their current location. As Glaeser and Gottlieb (2009) emphasize, this implies that research on places is vitally different from research on countries, and requires that population, income, and prices are considered simultaneously. Regions are interdependent to an even greater extent than countries, and there is a real sense in which regional growth is a harder topic than national growth.

There is a further departure from the standard competitive paradigm. A long tradition in urban economics and economic geography explains the structure of cities and economic activity in terms of various externalities. These have the potential to generate inefficient and undesirable outcomes. Although some externalities are reasonably well understood, at least in theory, others are not. The example of Detroit shows how one mechanism in regional decline will be changes in crime and social norms, amplifying changes that originated elsewhere. Recent empirical work on local institutional variation within developing countries, such as Dell (2010), suggests remarkably powerful and long-lived effects of this variation. The reasons for this remain unclear, but could include not only the persistence of institutions, but also the intergenerational transmission of social norms and political beliefs. Economists have only just begun to engage with such complex forces.

These points hint that a single "canonical" model of regional growth is neither likely nor desirable. There are so many interesting research questions that it would be a mistake to seek or impose a single framework. The chapter will discuss how spatial disparities evolve over time; the circumstances in which there is a regional problem; how differences in regional living standards and productivity can arise; the data and methods used to study regional growth; and the forces that drive regional growth and regional decline.

Some of these questions are too intertwined to address sequentially. As in the literature on national economic growth, it can be a mistake to attempt a sharp distinction between growth and levels, as if these two phenomena necessarily require separate models. In practice, it makes little sense to write about regional growth without taking a view on what determines relative levels of income per capita. But we also note that regional growth does not only mean growth in average living standards. In common usage, it often means growth in relative population or total income, as a region outperforms others. One theme of our chapter is that changes in the relative sizes or population densities of regions merit more attention from researchers. For the study of decline, in particular, it is important to analyze depopulation rather than simply relative living standards.

Another complication is even more fundamental. Productivity growth is accompanied by, and to a large extent inseparable from, changing patterns of agglomeration and dispersion. Growth will respond to, and bring, changes in demand patterns; sectoral and occupational structure; skill levels; transport costs and infrastructure; financial development; and even local institutions and political economy. All of these could reconfigure the spatial structure of population and production. Yet modeling growth and agglomeration as outcomes of a joint process is far from straightforward, as Krugman (1995) noted. We call this "the hard problem" of regional economics, and review some models that seek to address it.

When we turn to the evidence, we depart from existing surveys by considering a wider range of countries. The New Economic Geography literature has tended to focus on Europe, Japan, and the USA, but the study of regional prosperity is even more important for contemporary developing countries, as Venables (2005) emphasizes. After all, some Chinese provinces and Indian states exceed many countries in population and land area. These include India's Uttar Pradesh (population around 200 million), and Maharashtra and Bihar (both in excess of 100 million). Guangdong province in China has a population of more than 100 million once migrants are included. The intrinsic importance of this should be clear, and the consideration of developing countries has a further benefit, widening the scope of the available evidence. Recent work has opened up some startling research possibilities, not least through the use of satellite data on light density at night to map activity at the sub-national level.

A final theme will be the formidable identification problems that arise in studying regional data. Some can be seen in narrow terms as spatial dependence; for example, errors in a regression model will often be correlated across regions. But more fundamentally, the requirements of spatial equilibrium will link regional outcomes and characteristics so tightly that it is rarely clear how to recover causal effects. This continues to be a major obstacle to understanding regional growth or making policy recommendations. We will discuss a range of empirical methods, which recover causal effects with varying degrees of plausibility, and the complementary role of structural models.

With all this in mind, the coverage of the chapter is intentionally broad, not to say sprawling. Over the past two decades, after long years of neglect, economists have developed a rich theoretical and empirical literature on economic geography, chiefly inspired by Krugman (1991). But concurrently, and largely independently, researchers working on growth and development have studied the effects of policy reforms and institutional variation using sub-national data. Influential papers include Banerjee and Iyer (2005), Besley and Burgess (2000), Holmes (1998), Jayaratne and Strahan (1996), and Tabellini (2010), but these are just a few examples from an increasingly extensive literature. An important aim of this chapter is to bring these various strands of research together, and use them to interpret each other.

To keep the scope of the chapter manageable, we also need to set some limits to what we cover. We emphasize work in economics, and especially recent work that takes a general equilibrium approach. This is a significant limitation, because the study of regional outcomes extends well beyond economics, to include research in geography, urban planning, sociology, statistics, and demography. From the mid-1950s onwards, elements of these approaches began to coalesce in the interdisciplinary field of regional science. This field has sometimes drawn on ideas from economics, such as applied general equilibrium modeling, adapted to include a spatial dimension.[1] It is clear, however, that the traditional methods of regional science are rarely well adapted to the study of regional growth and convergence dynamics. Their roots lie in static models, or in empirical methods that will rarely identify causal effects within a spatial equilibrium. It is also noticeable that, when textbooks on regional economics turn to growth, the approaches presented lack coherence. They range from basic trade-theoretic analyses, through closed-economy, one-sector growth models, to an emphasis on the demand-side role of regional exports and trade balances. Each time, it is all too easy to see what is missing: interesting dynamics, an explicit spatial dimension, a central role for supply adjustments and constraints. For all the benefits of an eclectic or interdisciplinary approach, many of the interesting questions demand general equilibrium reasoning, and the task is lost without it.

This should be clear if we consider one of the strongest associations in the data. Gennaioli et al. (2013a) emphasize that regional output per capita is strongly correlated with average human capital. In a regression of output per capita on average years of education and country dummies, using 1500 regions across 105 countries, they find that education explains 38% of the variation in output per capita within countries. This is

[1] An overview of quantitative methods in regional science can be found in Isard et al. (1998).

striking, but nobody would propose that regions within a country are each endowed with a fixed stock of skilled workers. Investigating the association requires models in which regional factor supplies are endogenous to the location decisions of workers and firms; Gennaioli et al. construct one such model, and alternatives will be considered below.

In summary, the study of regional growth often requires structural models that draw heavily on economic ideas; an understanding of the various forms of interdependence between regions; and empirical methods that can overcome the identification problems raised by that interdependence. We take these endeavors as our central focus, rather than the much wider literature in regional science and geography.[2] Nor do we provide a discussion of regional policies: these require an understanding of the mechanisms at work in regional growth and regional decline, but we do not develop the links explicitly. Finally, we note that some of the analytical issues overlap with those of urban economics; for a discussion of ideas specific to urban economics and the growth of cities, see the chapter by Duranton and Puga in this volume.

The remainder of the chapter considers different perspectives in turn. Section 4.2 looks at convergence and polarization, the associated methods, and some stylized facts. Section 4.3 will discuss the nature and interpretation of spatial disparities, and when they matter. The remaining sections, which are really the heart of the chapter, investigate the drivers of regional growth and decline. Section 4.4 covers an array of relevant models. Section 4.5 sets out two classes of models that consider growth and agglomeration jointly. We then review empirical methods (Section 4.6) and some of the main findings (Section 4.7). Section 4.8 briefly discusses regional decline as a distinct phenomenon, while Section 4.9 concludes.

4.2. CONVERGENCE, DIVERGENCE, POLARIZATION

Do regional economies have a tendency to move closer together, grow in parallel, or move further apart? This question has spurred many empirical studies, but is surprisingly difficult to answer. Part of the problem may lie with the question. We are interested in how a distribution of outcomes evolves over time, but distributions can behave in complex ways, and much is lost by collapsing this behavior into a crude binary opposition between convergence and divergence (Durlauf et al. 2009). As in the literature on national growth, part of the interest in these questions revolves around more complex possibilities. These include distinct "convergence clubs," or the emergence of polarization. A related question is that of mobility within the distribution. Since a detailed survey of regional convergence is already available (Magrini, 2004) we emphasize the studies that are especially relevant to understanding regional growth, or that have emerged over the last decade of research.

[2] For a broader perspective, see Clark et al. (2000). The relationship between work in economics and geography has been extensively discussed, as in Brakman et al. (2009, Chapter 12), Krugman (1995), and Ottaviano and Thisse (2005). Brakman et al. also survey regional growth, as does Harris (2010). For a discussion of the regional policy implications of recent work by economists, see Combes (2011).

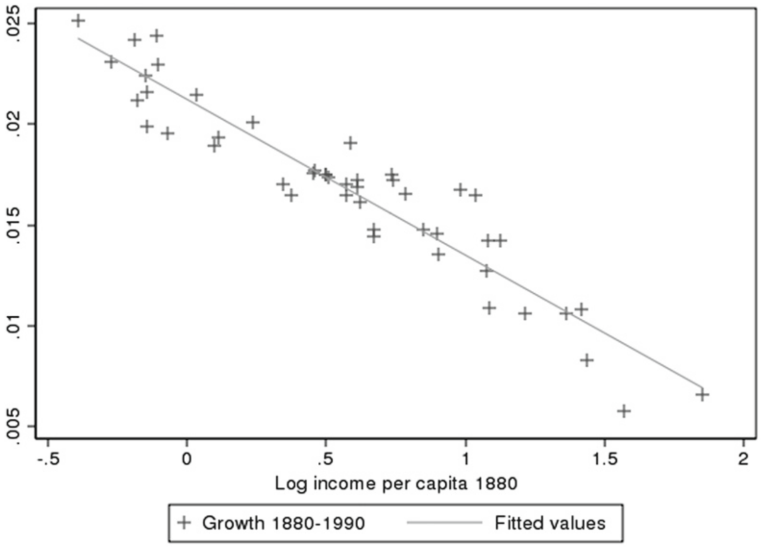


Figure 4.1 Absolute convergence?

The details of some of the more technical methods, using time-series concepts, transition matrices, or mixtures of densities, can be found in the appendix to the chapter.

4.2.1 Beta-Convergence

Do regions converge to the same level of income per capita? Our starting point is Figure 4.1, which shows annualized growth against initial income per capita for 47 contiguous US states, using data for 1880 and 1990.[3] This is similar to Figure 11.2 presented in Barro and Sala-i-Martin (2004) and uses their data. At first glance, the strength of the correlation between growth and initial income per capita is decisive evidence for the absolute convergence of regions; they draw attention to the high R^2, 0.92, of a regression close to that shown. At first glance, this suggests that regional disparities are transitory phenomena.

The figure is a little deceptive, however. To see this, consider what would happen if two regions with the same initial income differed in their growth rates, so that one region was slightly above the regression line and one slightly below. Over the 110 years between 1880 and 1990, this small difference in growth rates would compound to imply a large difference in relative levels. Hence, even the strong negative correlation in Figure 4.1 does not imply that spatial disparities will be eliminated. To show this, Figure 4.2 presents the same data in a different way, showing income per capita relative to the median region in 1990 against that in 1880, together with a 45-degree line. The shallow slope of the (dashed) regression line is consistent with mean reversion, but using the vertical axis, it

[3] The state missing from the 48 contiguous states is Oklahoma, which lacks data for 1880 given its late statehood.

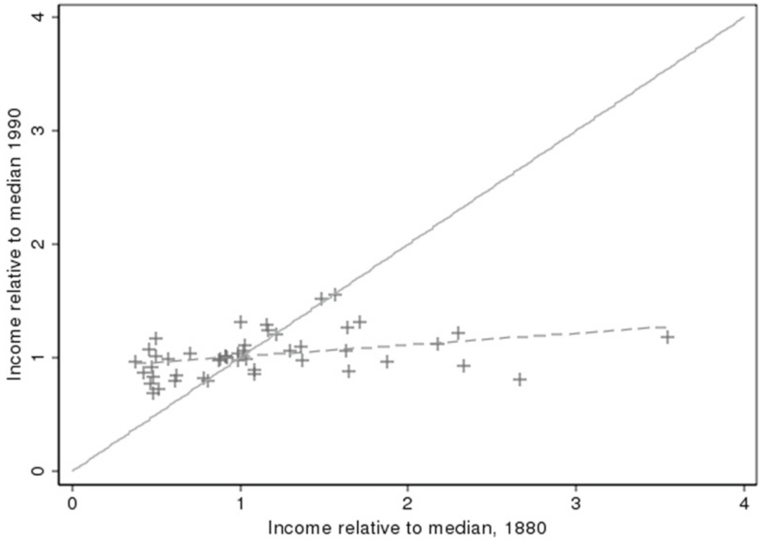


Figure 4.2 Long-run disparities.

is also clear that significant differences in living standards across US states remain. The richest state has more than twice as much income per capita as the poorest. This is contrary to what might have been expected from Figure 4.1, but in line with the common sense view that regional disparities persist over time.[4]

Worldwide, regional disparities are pervasive and substantial, especially in poorer countries. Using data on 1537 regions across 107 countries, for the year 2005, Gennaioli et al. (2013a) report that the average ratio of income per capita in the richest region to the poorest region is 4.41. The ratio is 3.77 for Africa, 5.63 for Asia, 3.74 for Europe, 4.60 for North America, and 5.61 for South America. The ratios are substantially higher in some cases, including Indonesia, Mexico, and Russia. Most of these figures do not correct for price levels, which are often higher in richer regions. Nevertheless, it seems unlikely that price differentials could account for the majority of the variation in nominal incomes.

The early literature on beta-convergence was highly successful in drawing attention to the rich interest of regional growth, and spurred a major research effort in the area. But the specific approach has also drawn criticism, partly on econometric grounds that we discuss in the appendix, and partly because the results can be hard to interpret. This is because they rely on viewing regional data through the prism of the neoclassical growth model. Indeed, it is often asserted that regional data provide an ideal testing ground for that model. These claims are misplaced, because the neoclassical model typically rules out cross-border

[4] There are some changes in rankings—the Spearman rank correlation is 0.47—and beta-convergence is sometimes argued to be informative about mobility; but it is not straightforward to map an estimated convergence rate onto a readily interpreted scale for a mobility index.

flows of goods, services, capital, and labor. The assumption that regions are closed to such flows is hardly attractive. But allowing for these flows is not straightforward, and general equilibrium models rarely lead to simple regression specifications. These perspectives suggest that beta-convergence studies will miss a great deal, and other methods are needed.

4.2.2 Inequality and Polarization

When considering regional convergence, a useful starting point is to ask whether the cross-section variability of income per capita is increasing, stable, or falling over time. The most prominent version of this is sigma-convergence, which considers the evolution of the standard deviation (σ) of the logarithm of income per capita; regions are said to be converging if the standard deviation is falling over time (Barro and Sala-i-Martin, 1991). This measure of regional inequality is not Lorenz-consistent, but could be replaced with the Gini coefficient, the coefficient of variation (as in Williamson, 1965), or the Theil measures. The Theil measures, and other members of the generalized entropy class, have the significant advantage for regional analysis that they are decomposable; see Cowell (2011) for a textbook treatment.

A further question is whether or not to weight regions by their populations; Milanovic (2005b) provides a discussion of this. If the aim is to capture the spatial inequality perceived by a randomly drawn individual, then weighting by population is natural, as in Williamson (1965) and many subsequent papers. But for the analysis of regional growth, a researcher might be interested in the effects of physical geography, institutions, and policies. In that case, it might be sensible to give regions equal weight even when they vary in size, rather than allow the results to be dominated by the characteristics of the largest regions.

As a first step in a descriptive exercise, the study of inequality measures is often valuable. But a given time path for regional inequality could correspond to a variety of underlying processes, with different long-run implications. A small group of regions may diverge from a larger group; as a result, measured inequality could increase even while a large number of regions grow in parallel. A related possibility is that the distribution becomes polarized. This term is often used rather loosely, to indicate some degree of high or rising inequality across regions. The view that deregulated market economies give rise to excessive polarization, in various senses, is especially common on the political left (for example, Dorling, 2011); others use the term to indicate a "disappearing middle" or "clustering around extremes."

On a more formal definition, polarization can be seen as concerned with multiple modes, the distance between these modes, and the distribution of probability mass around them. Drawing on Duclos et al. (2004), imagine that there is reduced variation in living standards at two different ranges of the regional income distribution. This is likely to reduce inequality, but polarization increases, since the contrast between the two groups is made sharper and more visible. More generally, polarization—in this technical sense—can increase even as the cross-section dispersion in living standards falls. These arguments suggest the need to look beyond inequality, using methods reviewed in the appendix to this chapter.

4.2.3 Findings

It seems inevitable that regional disparities will sometimes be compounded by growth and agglomeration. Seen against the long span of human history, current disparities may be a comparatively recent phenomenon; Bairoch (1993) argued that there was considerable uniformity in development levels in the early modern period (say, 1500–1800). One empirical approach relates regional inequality to the national level of development, as in the classic paper by Williamson (1965). He hypothesized an inverse-U relationship, with regional inequality rising and then falling as development proceeded. More recently, Barrios and Strobl (2009) examine the relationship using data for 12 European countries over 1975–2000. The data plotted in their Figure 4.2 suggest that regional inequality is increasing at lower levels of development, before either leveling off or reducing somewhat, but rarely returning to its initial level. For a much larger set of countries, Lessmann (2011) finds some evidence for the inverse-U relationship over 1980–2009, with regional inequality peaking at a development level close to that of, say, Mexico or the Czech Republic. He also finds some evidence that regional inequality increases at very high levels of GDP per capita (roughly, Canada's level).

A more common approach in the literature, following Barro and Sala-i-Martin (1991), has been to consider the evolution of regional inequality over time. Sometimes data on regional output are used, and sometimes data on income. For a few countries, including China and Indonesia, region-specific price deflators are available. For developing countries in particular, the treatment of natural resource revenues can be important to the results. Differences across regions in age structure, employment rates, and part-time work are another complicating factor. We defer a more thorough discussion of regional data until Section 4.6 below.

A mixture density approach (see the appendix) has been applied to European regions by Pittau (2005) and Pittau and Zelli (2006). Their work suggests a multimodal structure for the 1970s and early 1980s, and represents the distribution as a mixture of two well-separated normal densities. These two clusters later converge. They also find that, from the mid-1990s, a small group of very rich regions (Brussels, Hamburg, Île de France, and Luxembourg) moves further ahead, a result also highlighted by Enflo (2010). This suggests recent polarization, and indicates that using a benchmark region or weighted average to assess convergence is risky. Evidence for polarization also emerges from other studies, including Canova (2004).

An alternative approach is based on transition matrices or stochastic kernel densities. In applications of these methods, the stationary distribution for the US states appears to be unimodal (Johnson, 2000) while the stationary distribution for European regions is more likely to appear bimodal. But one of the most sophisticated time-series studies of the US, that by Carvalho and Harvey (2005), finds that the two richest macro-regions, New England and the Mid-East, have pulled away from others over time. This indicates some degree of polarization may be emerging for the US as well as Europe.

For Japanese prefectures, Sala-i-Martin (1996) found a sharp decline in the dispersion of average personal income between 1940 and 1955, and a smaller decline in the 1970s. At a more disaggregated level, Seya et al. (2012) found a decline in the log variance across Japanese municipalities over the 1990s, and a slight increase in the 2000s. For Russia, the usual finding is that regional disparities increased sharply in the first years of the post-Soviet era (for example, Fedorov, 2002) and were high in comparison to many other countries; more recently, over the course of the 2000s, they seem to have fallen (Guriev and Vakulenko, 2012).

The experiences of many developing countries are at least as interesting. The literature on China, in particular, is extensive. It typically finds divergence for the 1970s, followed by a period of convergence in the wake of agricultural reforms, and then further divergence during the rapid industrialization of the 1990s (for example, Weeks and Yao, 2003). For the 1990s onwards, the fast growth of the coastal provinces is often emphasized, consistent with a story in which market access has promoted industrial development. Démurger et al. (2002) note the importance of three exceptionally rich provinces, Beijing, Tianjin, and Shanghai, in raising the overall degree of regional inequality. But even excluding these provinces, regional inequality rose over the 1990s. It is substantially higher than India's, a finding that is conventionally explained in terms of barriers to mobility within China (for example, Gajwani et al. 2006).

Milanovic (2005b) studies regional inequality in the five federations of Brazil, China, India, Indonesia, and the US, over the 1980s and 1990s. For Brazil, he finds no clear trend; Azzoni (2001) studies Brazil for a longer period (1939–1995) and finds an overall decline, although one interrupted by a sharp increase in the 1970s. For India and Indonesia, Milanovic finds regional inequality to have increased. Hill et al. (2008) also study Indonesia, but find that the coefficient of variation of non-mining output per capita was broadly stable over 1975–2004, despite fast growth.[5] The reason for the inconsistent findings is not clear, and more generally, Milanovic (2005b) notes that the field lacks a consistent terminology and approach.[6]

Convergence methods have sometimes been applied to variables beyond income or output. Evans and Karras (1996) found evidence for rapid conditional convergence of TFP for the contiguous US states, and similarly rapid mean reversion for the returns to capital, computed using data on capital-output ratios and factor shares. Overall, their

[5] If mining is included, the coefficient of variation declined sharply in the 1970s and 1980s.

[6] The picture may be worse than he suggests: some papers on convergence make basic errors in the use of inequality measures, such as calculating the standard deviation of income rather than log income, or calculating the coefficient of variation using log income.

results point to inter-region capital mobility, but also that US states are converging to region-specific steady-states determined by long-run differences in productivity.

The convergence behavior of regional house prices has also been studied. This is interesting in its own right, but also because housing costs will be determined jointly with incomes, and hence informative about the mechanisms generating spatial disparities. Using a model-based approach to US data from 1975, Van Nieuwerburgh and Weill (2010) emphasize that inequality in (quality-adjusted) house prices across US regions has risen substantially. By calibrating a model with a spatial equilibrium, they argue that increased dispersion of regional productivity is needed to explain the US data. Their analysis demonstrates the benefits of studying movements in cross-section dispersion using a structural model, rather than treating the study of convergence as solely an econometric problem; we return to this point later.

4.3. DO REGIONAL DISPARITIES MATTER?

At first glance, differences in average living standards across regions contribute to overall inequality, and involve some degree of unfairness or injustice, with social and political consequences. An individual born in a depressed region may have fewer opportunities and poorer life chances than an individual born in a more prosperous region of the same country. A widely held view is that uneven regional development can undermine social cohesion and generate political tensions. This seems especially likely in those developing countries where regional disparities coincide with the spatial distribution of ethnic groups or natural resources. But the issue is not confined to developing countries, as witnessed by regional tensions within Belgium and Italy, among other countries. Judt (1996) prophesized that the project of European integration would eventually be undermined by polarization between dominant "super-regions"—such as Baden-Württemberg, Rhône-Alpes, Lombardy, and Catalonia—and an economically depressed periphery.[7] The European Union has made strengthening cohesion across the regions of its member states a major priority, enshrined by treaty, and around a third of the central EU budget is directed at policies to reduce spatial disparities.

4.3.1 Composition Effects and Welfare

Much of this is familiar, but care is needed over the meaning of "living standards," and their comparison across space. In simple models, if workers can move freely between regions, then returns to worker characteristics are equalized across space, at least in the long run.[8] Spatial differences in average income per capita do not map straightforwardly

[7] In his view, the richer regions would be likely to identify themselves ever more closely with Europe, while the less successful periphery would see increasing scepticism and resentment about the European project, and perhaps a resurgence of nationalism.

[8] For a model that does give rise to spatial variation in skill premia, see Davis and Dingel (2012).

into spatial inequality in life chances, welfare differences, or life satisfaction. In a spatial equilibrium, regional disparities will often reflect composition effects. An agricultural region may have relatively low income per capita not because it is "depressed," inefficient, or its workers underpaid, but because the relatively unskilled account for a high share of its employment, or because its activities are relatively labor-intensive.

These points apply more broadly, and their importance is supported empirically. Acemoglu and Dell (2010) show that approximately half of the within-country, between-region inequality in labor income in the Americas (Canada, US, and Latin America) can be accounted for by differences in workers' education and experience. Another branch of the literature uses household-level data to account for regional differences in household composition, and to estimate how the returns to characteristics vary across locations; relevant studies include Nguyen et al. (2007) for Vietnam, and Skoufias and Katayama (2011) for Brazil. Nguyen et al. find that, in their data for 1993, the urban-rural gap in household consumption per capita is primarily due to differences in covariates such as education, ethnicity, and age, and this is true throughout the distribution. In survey data for 1998, however, there is more evidence for differences in returns to covariates between the urban and rural sectors. In the case of Brazil, Skoufias, and Katayama find that differences in household composition account for most of the inter-regional differences in consumption, but find some evidence for differences in returns across metropolitan areas, and between metropolitan and non-metropolitan urban areas.[9]

The importance of composition effects for income comparisons is clear. Less widely appreciated, a similar argument applies to spatial comparisons of happiness or well-being. The literature often assumes that a spatial equilibrium requires utility to be equalized across locations, but this will only happen if individuals are homogeneous. When individuals differ, average well-being is also likely to differ in equilibrium (Deaton and Dupriez, 2011). To give a concrete example, if retirees are happier than the middle-aged, and especially likely to move to Florida, then average self-reported happiness will tend to be relatively high in Florida. Again, compositional effects give rise to differences in average outcomes.

Although these effects make interpretations uncertain, Pittau et al. (2010) document some interesting differences in self-reported life satisfaction across European regions, with especially wide variation in Belgium, Germany, Italy, Portugal, and Spain. For reasons that are unclear, residents of capital cities are especially likely to report low levels of life satisfaction. Knight and Gunatilaka (2011) summarize their work on happiness in China, which finds that (perhaps unexpectedly) mean urban happiness is slightly below mean rural happiness, while households of rural migrants living in the cities report lower average

[9] They attribute these return differences to agglomeration economies, but this argument is not complete, because labor mobility should equalize returns across locations even in the presence of such economies. As in the case of Vietnam, alternative explanations could include a short-run disequilibrium, or unmeasured differences in characteristics (such as those in ability or the quality of schooling); Young (2013) emphasizes this latter possibility.

happiness than other urban households. But the compositions of these populations may differ in terms of characteristics that are hard to observe.

Oswald and Wu (2010, 2011) study differences in self-reported well-being across US states, and emphasize that these differences are not modest. Their data allow them to control for a variety of personal characteristics, and therefore address composition effects. The overall argument in Oswald and Wu (2010) is that well-being, conditional on individual characteristics including income category, is correlated with measures of local amenities (measuring non-income aspects of quality-of-life) extracted separately from a compensating-differentials approach, in earlier work by Gabriel et al. (2003). This correlation is seen as independent validation of the well-being measures. For the present chapter, it is more noteworthy that the correlation is consistent with a spatial equilibrium under labor mobility, in which the income received by a given type of worker will be lower in those states with good amenities.

Composition effects have a stark consequence. If heterogeneous workers are allowed to sort across locations within a market economy, then regional disparities in average incomes and even average life satisfaction are inevitable, and a condition of efficiency. This line of argument seems to conflict with common sense: surely equality across regions is obviously desirable? What the common-sense view misses is the need to follow Sen (1980), and ask "equality of what?" One possible criterion is to compare the utilities of a given type of individual across space. But, for the reasons just explained, a long-run equilibrium which equalizes these utilities will rarely equalize average outcomes.

4.3.2 When is There a Regional Problem?

Although the logic of a spatial equilibrium is powerful, the perception remains that uneven regional development is problematic. The literature often suggests that a spatial equilibrium will be inefficient, because externalities play a central role. Outside economics, observers often suggest that regional disparities are a form of social injustice. At a minimum, economists risk underestimating the burdens of adjustment; the experience of Detroit is a salutary reminder of a harsher reality.

One concern is that growth and agglomeration in a core area could make those living in the periphery worse off, even in absolute terms. This will be a particular concern when there are barriers or frictions that restrict the mobility of individuals or firms, and labor mobility may be especially difficult for the poor.[10] But there are ways in which injustice could arise even when mobility is unrestricted. It may be that human capital is relatively costly to acquire in poorer regions; since children cannot choose where to locate, regional disparities would contribute to differences in life chances and inequality. Olivetti and Paserman (2013) argue that, given the tendency for children to remain in

[10] Guriev and Vakulenko (2012) find evidence for poverty-related immobility in Russia in the 1990s, and Phan and Coxhead (2010) for Vietnam.

the same region as their parents, regional disparities help to explain the decline in social mobility seen in the US in the first part of the 20th century. Durlauf (2012) discusses the possible relevance of poverty trap ideas at the regional level.

Another argument is that spatial equilibrium takes time to achieve. There could be lengthy periods for which utility differences persist. Those who leave declining regions are likely to experience significant disruption in their lives, relative to the residents of more prosperous regions. Moreover, life chances may be influenced, in ways that economists have rarely analyzed. This perspective could imply less emphasis on comparisons of averages, and more on regional differences in labor markets, as in the work of Overman and Puga (2002) on the spatial polarization of European unemployment rates. But even this argument is not straightforward; in models based on matching frictions, if workers can move freely between locations, the asset value of unemployment will be the same across locations.[11]

4.3.3 The Origins of Regional Disparities

What are the underlying causes of regional differences in prosperity? Later sections of the chapter will consider this question in detail. But a useful first step is to set history to one side, and investigate the relative contributions of proximate influences, such as physical or human capital. For the purpose of an accounting exercise, we can adopt the simplifying device of a regional production function:

$$p_i Y_i = p_i A_i K_i^{\alpha} \left(h_i L_i\right)^{1-\alpha}, \tag{4.1}$$

where Y_i denotes the aggregate level of output of region i, p_i the price of output net of trade costs incurred by local firms, A_i is total factor productivity, K_i is the region's capital stock, and $h_i L_i$ is the region's supply of effective units of labor, where h_i is the average human capital level. Our primary interest in this section is to see how much of the variation in $p_i Y_i$ can be explained by variation in the factors on the right-hand side of (4.1). We have abstracted from intermediate inputs, for simplicity.

This exercise is a regional counterpart to the cross-country literature on development accounting reviewed by Caselli (2005) and Hsieh and Klenow (2010). The main finding of that literature is that international output differences are only partially explained by differences in physical capital and educational attainment, with the majority accounted for by total factor productivity, measured as a residual term. Hsieh and Klenow (2010) suggest that TFP accounts for 50–70% of cross-country output differences, with human and physical capital only accounting for 10–30% and 20%, respectively. Should we expect similar findings at the regional level?

[11] See Satchi and Temple (2009) and Kline and Moretti (2013) for related analyses. In a dual economy model with matching frictions, the "urban" region could mean one city, or urban areas at multiple locations, each with the same labor market tightness, but of indeterminate relative size.

Aiello and Scoppa (2000) and Scoppa (2007) investigate this for Italy, and Tamura (2012) for the US. For some countries, regional data are more detailed than cross-country data; human and physical capital may be available over longer periods of time, and at lower levels of aggregation. A disadvantage is that regional price levels (p_i) are often unobserved, although Tamura (2012) uses (limited) information on regional price dispersion in the US.[12] Broadly speaking, these studies tend to confirm the cross-country finding that TFP is more important than human and physical capital in explaining output differences. However, this conclusion is sensitive to the way in which human capital is measured. For example, Scoppa (2007) finds that using quality-adjusted education levels can raise the contribution of human capital to over 50%.

Other papers also find that regional prosperity and average human capital are strongly associated. This includes work on spatial sorting, drawing on the urban economics literature. Combes et al. (2008) use a large panel of French employees (close to 20 million observations for the years 1976–1998) to isolate the importance of skill composition in explaining local variation in wages. They find that up to half of the variation in wages across 341 French employment areas can be explained by differences in skills. For 119 areas of Great Britain, Rice et al. (2006) find a smaller, but still substantial, role for occupational composition.[13]

The evidence points in a consistent direction. In the study of regional prosperity, a key question is why skilled individuals are more likely to choose to locate in some regions than others. To answer this question needs general equilibrium models that can map regional characteristics into endogenous outcomes such as the supply of skills in each region, other factor supplies, and (in some cases) the endogenous determination of total factor productivity. Much of the rest of the chapter will be about this endeavor.

Before we describe the relevant theories, there is another point to note. We have discussed differences in outcomes across locations, but we have not allowed the nature of those locations to play a determining role. At least since Adam Smith's *Wealth of Nations*, it has been acknowledged that regional outcomes are related to their physical geography. It seems equally clear that outcomes are related to the outcomes of neighbors, and more broadly, to economic geography. To illustrate this, we plot regional GDP per capita against distance from Luxembourg, for a large number of European regions; see Figure 4.3, reproduced from Breinlich (2006). The strong correlation suggests that models of regional prosperity will need to engage with physical and economic geography.

[12] We discuss regional price deflators in the appendix to this chapter. Another issue for some studies is that the assumption of Cobb-Douglas technologies is restrictive; Bernard et al. (2013) find evidence against this assumption for the US.

[13] These findings are based on a decomposition of earnings differences, but as in cross-country variance decompositions, it is not clear how covariance terms should be treated. Duranton and Monastiriotis (2002) study changes in regional wage inequality in the United Kingdom, between 1982 and 1997, and especially the divergence between London and other regions. This was mainly driven by a stronger rise in education levels in the capital and an increase in relative skill premia.

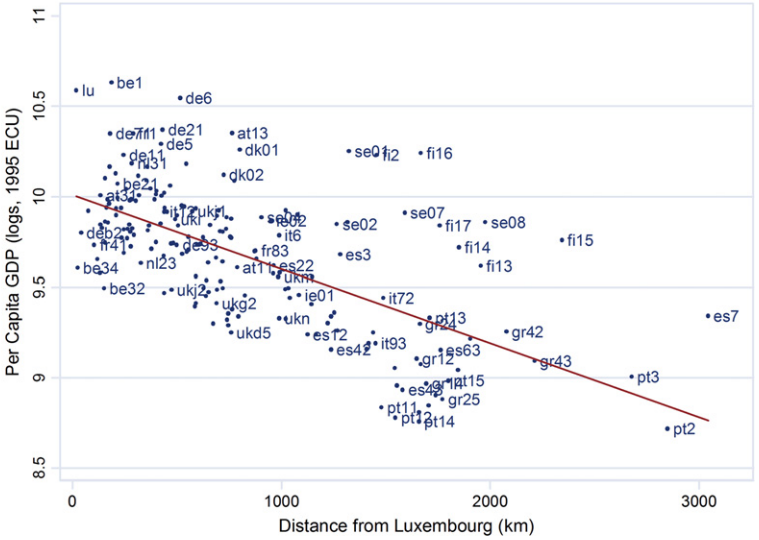


Figure 4.3 Regional GDP and proximity to Luxembourg.

A range of other observations support the idea that regional prosperity is connected with geography. There are clear spatial correlations of activity within most countries. Activity tends to gravitate toward areas with relatively good transport links, and close to large markets. It is unlikely to be an accident that much of China's industrial development has been concentrated on its coast, or that Brazil's interior is poorer than its coastal cities. Even for a well-integrated, developed economy such as the US, much activity is located on the coast, while population movements are influenced by aspects of physical geography, such as climate.

4.4. MODELS OF REGIONAL PROSPERITY

Geography clearly matters, but understanding its implications often requires formal models. As in other general equilibrium contexts, such as the study of international trade, informal reasoning can easily go astray. Yet for a long time, the study of regional prosperity was rather overlooked by economists. In the post-war period, it tended to attract those who were sceptical about some of the tools and findings of conventional economic theory. The complexities of the regional growth process were interesting in themselves, but could also be used to illustrate reservations about economic theory (and general equilibrium theory in particular) that applied more widely.

Perhaps the best-known contribution along these lines is that of Myrdal (1957). He argued that market forces would lead to divergence between regions and ongoing disequilibrium, driven by cumulative causation, or the tendency for a change in a given direction (such as an increase in one region's productivity) to instigate further changes in the same direction. Kaldor (1970) took up this theme, emphasizing that localized industrial development generates cumulative advantages, due to various forms of increasing returns. But Hirschman (1958) had already countered that such views went too far. Although some degree of uneven development was inevitable, it would be self-limiting: as regions moved apart, there would be powerful forces working to limit further divergence.

The recent development of the literature has revealed some truth in both sets of positions. A contemporary treatment will rarely see an inherent contradiction in the coexistence of feedback effects and the study of an equilibrium; but it is likely to follow Myrdal in stressing the importance of these feedback effects for analysing regional prosperity. Agglomeration may promote further agglomeration, and initiate other changes that are part of a larger, self-sustaining process. Further, Myrdal and Kaldor were interested in the idea that the forces which drive growth and agglomeration are not readily separable, and this recognition continues to pose a major challenge for theorists.

4.4.1 A Theoretical Challenge

The formal general equilibrium models of the 1950s and 1960s typically assumed perfect competition, so that firms are price takers in markets for output and factors. But when firm and individual location decisions are introduced in realistic ways, perfect competition can rarely accommodate an interesting equilibrium. Consider what happens when output depends on capital and labor, and these factors can move freely between regions. In that case, workers and firms will all move to whichever region has the highest total factor productivity.

This argument can be expressed in more formal terms. For location decisions to be non-trivial, imagine that individuals and firms must choose an address, and these choices affect their utility and profitability, respectively. This assumes some indivisibility in the way they spread their consumption and production activity across space, and assumes—for space to play a determining role—that there are transaction or transport costs involved when consumption and production are geographically separated. These are realistic assumptions, but they cause the competitive paradigm to break down (Starrett, 1978). A sufficient condition for this breakdown is that different locations have the same characteristics. Then, either there is no equilibrium with perfect competition, or all individuals and firms gather in a single location. More formally, if preferences are monotone, space is homogeneous, and transport is costly, there is no competitive equilibrium which involves transportation.

To generate an interesting location problem, localized externalities and indivisibilities at the level of individual firms and workers are necessary. But it is costly transportation that ultimately gives substance to the effects of geography. In the absence of transport costs, space is immaterial even when individuals have distinct locations. Scotchmer and Thisse (1992) call this the folk theorem of spatial economics. It underpins two laws of economic geography set out by Prager and Thisse (2012): not all activities are available everywhere (the first law); and what happens close to us is more important than what happens far from us (the second law).[14]

These ideas play a fundamental role in the economic analysis of spatial equilibrium. In the literature that followed Krugman (1991), they have been studied by allowing for increasing returns and market size effects. This has been achieved by sacrificing generality, using various simplifying devices. But making even these models dynamic is not straightforward. Theorists face a trade-off, balancing the demands of realistic geographic and spatial considerations against the simplicity needed for a manageable dynamic analysis. Since agglomeration and growth are complex phenomena on their own, often models with proper geography lack interesting dynamics, while models with interesting dynamics lack proper geography. We use this idea to organize our discussion of various formal models. We first consider growth models that are largely without spatial considerations (growth without geography); multi-sector models with some limited spatial content or implications (growth with limited geography); and spatial models largely without dynamics (geography with limited growth). The hardest problem, to model growth and agglomeration jointly, is deferred until Section 4.5; we call this geography and growth.

4.4.2 Growth Without Geography

Asked to consider regional prosperity, some economists have settled on a default approach, which is to see whether progress can be made by ignoring space altogether. At its extreme, this approach treats regions as if they are separate countries. Their locations may differ, but space has no determining role. Trade in goods, and the movement of factors between regions, are either frictionless or (more commonly) ruled out altogether. We call this form of approach growth without geography.

In particular, various neoclassical growth models remain an organizing framework for some research. They are used to think about the determinants of productivity levels, and to motivate many of the convergence studies discussed in Section 4.2 above. This might be a useful place to start, but it is no place to end. By construction, the models cannot account for the patterns of spatial dependence seen in the data, or changes in the spatial distribution of economic activity. The neoclassical growth models have dynamics—capital accumulation, productivity growth—but, to borrow the words from a popular song, they don't know much about geography.

The most sophisticated defence of the neoclassical growth model is that given by Barro and Sala-i-Martin (2004). They argue that, provided capital and labor are not perfectly mobile, the main consequence of factor flows between regions is to modify the rate at which regional economies converge to their steady-states. For example, Shioji (2001)

[14] Also note Tobler's first law of geography: everything is related to everything else, but near things are more related than distant things.

develops a growth model with exogenous public capital, and private capital that is mobile between regions but subject to adjustment costs. This leads to a conditional convergence equation where the steady-state depends on the equilibrium return to private capital, and the region's stock of public capital. This is a more detailed treatment than many in the literature, but problems arise if workers and firms make location decisions based on the spatial distribution of activity, or other characteristics of distinct locations. The logic of a spatial equilibrium requires location decisions and growth to be analyzed jointly, and the neoclassical growth model rules this out.

4.4.3 Growth with Limited Geography

Recent models of agglomeration and growth often imply a core-periphery structure, sometimes corresponding to a division between rural and urban regions. This connects with an older class of models, the dual economy tradition, where urban manufacturing and services coexist with rural agriculture. These models have long been studied within development economics, and on the borders of growth economics and trade theory. Historically, dual economy models have been studied mainly in terms of comparative statics: for example, the effect of a given productivity change, or factor accumulation, in changing the equilibrium.[15] Recent work has given greater emphasis to structural transformation as an ongoing, dynamic process. These models provide some insights into spatial equilibrium and the relative growth of rural and urban regions; they also provide a laboratory for developing some basic intuitions, before turning to richer models with more geographical content.

Strictly speaking, in the traditional approach to dual economies, the goods of the respective sectors are labeled, but not the locations of consumption and production, and firm location decisions are not modeled.[16] In the newer models from economic geography, the core-periphery structure emerges endogenously. In contrast, the older models can be seen as reduced forms, in which urban locations happen to have substantial advantages.

Typically, the agricultural and urban sectors are each modeled as producing an homogeneous good, under conditions of perfect competition. The relative price of the agricultural good is either determined exogenously (by the world prices facing a small open economy) or determined by utility maximization (in a closed economy). Less often, the agricultural and urban goods are treated as perfect substitutes. The location decisions of workers play a key role, so that one endogenous variable is the allocation of workers across the two sectors, and another is the equilibrium wage differential between the sectors/regions.

[15] See Temple (2005) for a survey that emphasizes their empirical applications.

[16] In those dual economy models which incorporate migration costs for workers, it would often be most natural to interpret agricultural production as taking place at a single point, and urban manufacturing and services production all taking place at another single point.

In the simple case where wages equal marginal products, and labor mobility equalizes wages, this maximizes aggregate output, in the absence of externalities or distortions. At the same time, the average product of labor will typically differ across sectors; regional differences in average productivity are a condition of efficiency rather than a sign of its absence. In richer dual economy models, however, inefficiency can easily arise, and urban regions may be too small or too large relative to rural regions.[17]

Some of the most interesting extensions to these models start with the urban labor market. In the model of Harris and Todaro (1970), a fixed urban wage leads to urban unemployment, and migration takes place unless expected utilities are equalized across the rural and urban sectors.[18] In many dual economy models with urban unemployment, productivity growth in the urban region will induce a migration response that increases the number of urban unemployed—the Todaro paradox. In contrast, productivity growth in the rural region will increase rural wages and relieve the pressure on cities, leading to better outcomes in the urban labor market and smaller regional disparities.

An especially rich approach to dual economies has been developed by Lagakos and Waugh (2013). They consider a general equilibrium Roy model, in which heterogeneous workers sort across sectors according to their comparative advantage. One implication of a Roy model is that all but the marginal worker will strictly prefer their current sector to the alternative. A calibrated version of the model can explain a large wage gap between agricultural and non-agricultural workers, without having to appeal to barriers to labor mobility, the traditional approach in dual economy models. It can also explain why international variation in agricultural productivity is usually found to be much larger than international variation in non-agricultural productivity (Caselli, 2005).

Even simple two-sector models demonstrate the importance of general equilibrium reasoning, while allowing for multiple sources of growth. In most of the models, the respective paths of rural and urban regions depend partly on rates of technical progress, and partly on capital accumulation. The accumulation of capital often leads to relative expansion of the sector/region which uses it more intensively (usually, but not always, the urban region). For the basic 2×2 model of trade theory, with two goods and two factors, where the factors are both mobile between sectors, this result is the standard Rybczynski effect. A version of that effect reappears in models with alternative labor market assumptions, such as the open economy version of the Harris-Todaro model studied by Corden and Findlay (1975).

In recent years, attention has shifted to dynamic versions of these small-scale general equilibrium models. These can be used to study structural transformation, and primarily the shift out of agriculture, as part of a transition toward a balanced growth path. Kongsamut et al. (2001) showed that this required strong assumptions that are unlikely

[17] Different versions of this can be seen in Graham and Temple (2006) and Satchi and Temple (2009).

[18] Approaches with endogenous wages often have similar implications; see Bencivenga and Smith (1997), Moene (1988), and Satchi and Temple (2009), among others.

to hold in practice. The approach of Ngai and Pissarides (2007) shows how to combine differential productivity growth rates with a balanced growth path, at the expense of restrictive assumptions on production technologies. Their (closed economy) model can explain ongoing declines in the relative price of the manufacturing good and the employment share of that sector.

Other recent work has started to combine dual economy ideas with models from urban and regional economics. Murata (2008) introduces a new mechanism for structural transformation, which draws on the New Economic Geography literature. In his model, a fall in transport costs increases the size of the market for non-agricultural goods, and—by lowering prices and raising real incomes—prompts a demand shift toward non-agricultural goods. Henderson and Wang (2005) construct a model of the rural-urban transformation which draws on dual economy ideas, but extended to consider the endogenous evolution of distinct cities, and allowing the formation of new cities. Michaels et al. (2012) study the US evolution of populations in rural and urban areas from 1880 to 2000, explaining the observed patterns partly in terms of structural transformation.

Rural-urban income differences can make a substantial contribution to overall inequality.[19] Dual economy models can be used to study this, and changes in the relative productivity of different sectors as development proceeds. There is long-standing evidence that structural change is associated with increases in the relative labor productivity of agriculture; see, for example, Temple and Woessmann (2006). Not all dual economy models readily generate this pattern, which makes it a useful test. Gollin et al. (2004) argue that introducing home production leads to a better explanation of the data. More broadly, one weakness of dual economy models is that not much attention has been paid to the modernization of agriculture; Yang and Zhu (2013) is a recent exception.

Although dual economy models have spatial implications, the locations are only differentiated by rural or urban activity, which limits their usefulness for understanding regional prosperity. The model of Gennaioli et al. (2013a) moves further in the required direction: workers and firms have distinct addresses, and decide where to locate. There are two possible types of region, productive and unproductive; at each location, there is a fixed supply of land and housing, and hence some part of the population remains in the less productive regions. The regions all produce the same good, which is freely traded internally. A key margin in the model is that especially able workers will self-select into entrepreneurship, and more able entrepreneurs run larger and more productive firms. Relative to most dual economy models, this gives greater importance to the stock of human capital, an idea that Gennaioli et al. investigate empirically. But the tractability of the model inevitably comes at a price. Although locations exist as discrete points with fixed stocks of land and housing, there is no role for transport costs, and hence the model cannot explain the spatial correlation of activity that is so apparent in the data.

[19] See Milanovic (2005a,b) and Young (2013), among many others.

Most dual economy models contrast agriculture with non-agriculture, but the divergent paths of manufacturing and services are increasingly important, for developing countries, as well as those that are developed. Desmet and Rossi-Hansberg (2009) argue that the age of a sector matters for the dynamics of agglomeration, defining a sector's age as the time that has elapsed since the last major innovation, such as electrification (for manufacturing) or IT (for services). In its early stages, a major innovation will spur geographic concentration, because knowledge spillovers are important as the technology is refined. But as further development of the technology slows down, concentration gives way to dispersion. Desmet et al. (2012) use this framework to analyze the evolution of employment density across districts of India, where growth has been associated with the rapid expansion of services in particular.

Much of the work discussed thus far originates in development economics. A parallel literature describes the interactions between growth and the size distribution of cities, drawing on work in urban economics. In Eaton and Eckstein (1997) and Black and Henderson (1999), localized externalities sustain the emergence of cities and generate increasing returns at the aggregate level, so that agglomeration triggers growth. Gabaix (1999) and Eeckhout (2004) show how models featuring exogenous localized growth and localized externalities can generate the stable distribution of city sizes observed in the data. In particular, they seek to explain why the upper tail of the distribution is approximately Pareto—so Zipf's law holds—although both very small and very large cities are systematically under-represented.

The random growth approach, revived in the recent literature by Gabaix (1999), starts with an initial arbitrary distribution of city sizes and lets each city grow at an arbitrary mean rate, around which cities are hit by period-to-period shocks. It then allows the cities to evolve freely and studies the conditions under which their limit size distribution mimics the observed one. Eeckhout (2004) assumes that total factor productivity in a city is determined by a positive localized externality that increases with city size and an exogenous process of localized technological change. In particular, letting $A_{i,t}$ be the productivity parameter reflecting the technological advancement of city i at time t, Eeckhout assumes that the law of motion of $A_{i,t}$ is given by $A_{i,t} = A_{i,t-1}(1 + \sigma_{i,t})$ with each city experiencing an exogenous technology shock $\sigma_{i,t}$. City-specific shocks are symmetric as well as identically and independently distributed with mean zero and $1 + \sigma_{i,t} > 0$. This law of motion implies that $\log(A_{i,t})$ follows a unit root process. There is clearly no growth in productivity in aggregate but, under appropriate functional forms, the model converges to a long-run distribution of city sizes whose upper tail is Pareto. City growth is proportionate, as also observed in reality.

Here, growth determines agglomeration, but the growth process itself is treated as exogenous. To fill this gap, Duranton (2007), Rossi-Hansberg and Wright (2007), and Córdoba (2008) propose models that generate growth processes consistent with specific features of the observed invariant distributions of city sizes. In all three contributions, growth leads to agglomeration. Duranton notes that it may be easier to match the city size distribution than it first appears, suggesting more attention is needed to the empirical relevance of the different possible mechanisms. Rossi-Hansberg and Wright address a particular conundrum for spatial theories of growth: the inherent tension between local increasing returns, implied by the existence of cities, and aggregate constant returns, implied by balanced growth. They show that variation in the urban structure through the growth, birth, and death of cities can be seen as the margin that eliminates local increasing returns, to yield constant returns to scale in the aggregate. Their model produces a distribution of city sizes that is consistent with the real one, and whose dispersion is also consistent with the dispersion of productivity shocks found in the data.

The close connection between these models and specific features of the data is attractive. But their usefulness for studying regional growth is ultimately constrained, because the models do not investigate how cities will be distributed across space. As a result, key features of the observed geographical distribution of economic activities, and their evolution through time, are absent.

4.4.4 Geography with Limited Growth

We now turn to more complex models, which draw heavily on ideas from international economics. The models we review are predominantly static. Nevertheless, it is often argued that static models can be used to understand the steady-state implications of dynamic processes, otherwise too complex to analyze. In the wake of a major change, such as a fall in transport costs or an improvement in total factor productivity, the outcome of regional adjustment processes can be understood in terms of the changing steady-state of a static model. We call this approach geography with limited growth.

Traditionally, this approach has been the backbone of the economics of agglomeration (Fujita and Thisse, 2002). These models address a fundamental question: can economic interactions generate spatial patterns of activity that are not determined solely by differences in exogenous fundamentals? Will asymmetric patterns of activity emerge even when locations are symmetric? As discussed by Ottaviano and Thisse (2004), the fact that economic activities are unevenly distributed in space is hardly surprising, given that locations differ in their climates, degrees of accessibility, and endowments of productive resources. All these features can be classified under the common label of "first nature." These features have undoubtedly played an important role in explaining economic history, not least in the early stages of economic development. Exogenous spatial heterogeneity is the cornerstone of neoclassical models of international trade, and land use-models in the tradition of von Thünen.

But another driving force of economic history has been the ongoing search for safe and cheap ways to move materials and products from one location to another. One consequence is that the spatial distribution of economic activity will not map directly against the spatial distribution of natural advantages. When workers maximize utility and firms chase profits, this will generate endogenous patterns of economic activity across space. This idea is captured by the concept of second nature, the forms of economic geography that emerge as the outcome of human actions. Modern theories of agglomeration study the relevant forces, unveiling how spatial patterns of activity depend partly on exogenous spatial heterogeneity, and partly on a range of other variables, not least transport costs.

Second nature geography is the outcome of an inherently dynamic process but, as already noted, it can be understood partly by means of tractable static models. At least since Marshall (1920), various second-nature forces have been studied by economists, geographers, and regional scientists, stemming from different types of localized technological and pecuniary externalities. For instance, technological externalities associated with production are stressed by modern urban economics, while pecuniary externalities associated with imperfect competition are stressed by spatial competition theory and work in economic geography (see Rosenthal and Strange, 2004).

The literature on these questions is vast, and a thorough assessment is beyond the scope of this chapter. Extended discussions can be found in Fujita and Thisse (2002), Combes et al. (2008), Neary (2001), Prager and Thisse (2012), and various volumes of the *Handbook of Regional and Urban Economics*. Here we want to highlight the findings most relevant to understanding regional prosperity. These can be summarized in terms of the so-called spatial question in economic theory (Ottaviano and Thisse, 2001). This question has two sides, one positive and one normative. On the positive side, the question at stake is whether the agglomeration of economic activities can be explained in terms of an explicitly defined market mechanism. On the normative side, if observed patterns of economic activity can be seen in terms of market outcomes, the question at stake is whether such outcomes are likely to be efficient. The answers are "yes" on the positive side and (usually) "no" on the normative one. A range of models link agglomeration and market forces, but typically these models are built on externalities and distortions that lead to some degree of inefficiency (for example, Ottaviano and Thisse, 2005).

As discussed by Ottaviano and Thisse (2005), the relative importance of technological and pecuniary externalities depends on the spatial scale of the analysis. According to Anas et al. (1998), cities are replete with technological externalities. The same holds in local production systems (Pyke et al. 1990). Besides local public goods, communication externalities are of particular interest. These could be critical in services such as management, administration, research, and finance. Knowledge, ideas and, above all, tacit information, can be considered as impure public goods that generate spillover effects from one firm or organization to another. If economic agents possess different pieces of information, pooling them through informal communication channels can benefit many, hence the importance of proximity (Feldman, 1994). Thus, to explain geographical clusters of somewhat limited spatial dimension, such as cities and industrial districts, it seems reasonable to appeal to technological externalities. In modeling terms, these can often be accommodated in the competitive paradigm. Future work is likely to draw on the economics of networks, to consider the dissemination of information within and across regions.

But when one turns to a larger geographical scale, alternative mechanisms come into play, and ones that are not easily accommodated in conventional general equilibrium models. Direct physical contact seems unlikely to explain major agglomerations such as the US manufacturing belt, or Western Europe's concentration of economic activity along the "Hot Banana" urban corridor, stretching from northern England to northern Italy. Instead, economists have sought to explain large-scale agglomeration in terms of pecuniary externalities. These arise from imperfect competition, in the presence of market-mediated linkages between firms and consumers/workers. The relevant models are often grouped under the banner of the New Economic Geography (NEG), which emerged in the early 1990s. This approach draws heavily on analytical tools and ideas from the theory of international trade (see, in particular, Helpman and Krugman, 1985). These tools are used to study the movements of goods, services, and factors within countries, and to explain agglomeration as the outcome of endogenous processes in which cumulative causation often plays a role.

This literature was founded by Krugman (1991), who develops a model in which agglomeration arises through the mobility of labor. This mobility endogenously generates variations in market size that promote further agglomeration since, in the presence of transport costs, firms want to locate near large markets. Spatial agglomeration can also rise through input-output linkages, in which the location choices of firms influence the size of the market for other firms and/or input costs (Venables, 1996). Some ideas have also been borrowed from urban economics: congestion and rising land rents can be introduced to offset the intrinsic advantages of particular regions, as in Helpman (1998) and Gennaioli et al. (2013a), among others.

Endogenous agglomeration arises because mobile factors like to cluster, and this can polarize the regional landscape between an active "core," and a "periphery" in which immobile factors face lower real remuneration. The emergence of a core-periphery structure typically depends on the level of trade frictions. In the absence of congestion in the use of land or other non-tradable and non-replicable resources, low trade frictions foster agglomeration, as immobile demand in the periphery can be serviced from the core (Krugman, 1991; Krugman and Venables, 1995). When congestion matters, the opposite is true: if trade frictions are low, the high local cost of non-tradables pushes mobile factors away from the core (Helpman, 1998). In the general case, agglomeration is more likely to emerge for trade frictions that are neither too low nor too high (Puga, 1999; Ottaviano et al. 2002).[20]

[20] See Fujita et al. (1999), Baldwin et al. (2003), Ottaviano and Thisse (2004), and Combes et al. (2008) for detailed accounts of NEG models.

Early NEG models were mainly aimed at explaining the "residual variation" of economic activities across locations based on second nature forces, classified as promoting either agglomeration or dispersion. Later models have increasingly brought first nature into the picture. To see what might be learnt about regional growth from this form of approach, we provide a quick sketch of the framework used in Rice andVenables (2003).[21] Assume two types of workers, skilled and unskilled, each with Cobb-Douglas utility functions based on four goods—housing, an international tradable good, a good that can be traded domestically but not internationally (e.g. certain financial services) and a good that cannot be traded domestically (e.g. restaurant meals or haircuts). The goods other than housing are each produced using Cobb-Douglas technologies. The international tradable, and the non-tradable, are treated as homogeneous, produced under constant returns to scale and perfect competition. The nationally traded good is produced under monopolistic competition as in Dixit and Stiglitz (1977). Finally, assume that workers can freely move between cities. Other things equal, they will migrate to cities with some intrinsic advantage, such as better amenities, until the advantage is offset by higher commuting costs and higher land rents. All rents are distributed to workers as a lump sum, in proportion to wages.

Under these assumptions, the prices of all goods are the same in all locations, but the skill mix of the labor force in each city is indeterminate. Two cities with different relative endowments of skilled labor will produce different quantities of the two traded goods, but have the same factor prices—a version of the factor prize equalization theorem of trade theory, applied within a country. Cities may then differ in terms of not only skill mix, but also in size and GDP per employee; and many different configurations of these outcomes are possible. But in the model, the free movement of labor implies that the utility of a given type of worker must be the same in all locations. And since wages are equal across locations, housing costs plus commuting costs must also be equal across cities. In this case, cities that are relatively skill abundant and high income must also be low density, so that the relatively high housing demand of skilled workers is offset by lower commuting costs.

At first glance, there is no regional problem, because the mobility of labor ensures that the utilities of a given type of worker are equalized across locations. But as new workers enter the labor force, the attainment of equilibrium relies on migration: each new generation has to relocate to restore the balance between the production structure of individual cities and their endowments of skilled relative to unskilled labor. In a richer model, reallocations could involve significant costs.

This simple setup has some counterfactual predictions: for example, the model predicts a negative correlation between GDP per employee and density, but the correlation in the data is often thought to be positive. A richer model gives one city/region an intrinsic advantage that will make it larger in equilibrium, and introduces transport costs for the

[21] See also Overman et al. (2010), who develop a diagrammatic approach to economic linkages across space.

nationally traded good, which is assumed to be skill-intensive. The larger market of the dominant city makes it a profitable location for the nationally traded good, bidding up wages (and GDP per employee) in the dominant city. Given transport costs, the price of the nationally traded good is lower in the dominant city, and workers are attracted to the city by higher wages and lower prices, until these advantages are offset by higher housing costs. Again, utilities are equalized in equilibrium, but changes in underlying parameters, such as transport costs or the intrinsic advantages of the dominant city, will generate population and asset price movements.

An alternative modification assumes that one city has a productivity advantage in the production of the (skill-intensive) international tradable; no transport costs; and commuting costs that are equalized across cities. Now consider an increase in the traded-sector productivity advantage of a dominant city. This will raise wages in the dominant city, crowding out the nationally traded sector, and attracting workers until the high wages are fully offset by a higher price of the non-traded good and higher housing costs. This latter version of the model generates positive correlations between density, GDP per employee, wages, average skills, price levels, and housing costs, which may often be the empirically relevant case. But there is no distinctively spatial pattern to the process of agglomeration.

For now it is interesting to consider what has been learnt from the sketch above. There may be disparities in skill endowments and GDP per employee across regions, but regions that appear advantaged may also have higher housing costs and higher prices for goods that are not traded across regions. Utilities are equalized in equilibrium by assumption. But spatial disparities continue to have relevance for policy-makers, not least if there are costs of adjustment. Changes in parameters—such as the productivity advantage of one city—might induce a lengthy transition process that has relatively modest ultimate benefits. In a numerical example in Rice and Venables (2003), a relatively modest change in traded-sector productivity can generate large population movements. This process of transition and adjustment may be associated with equilibrium utility levels that are only modestly higher than before. It is noteworthy that, in a quantitative exercise based on US data, Desmet and Rossi-Hansberg (2013b) find a similar result: eliminating differences in productivity or amenities across US cities would lead to major population movements, but modest welfare gains. For China, the welfare gains are estimated to be larger by an order of magnitude.

To date, the literature on geography and trade is long on models that study the mechanics of agglomeration in abstract landscapes, but remains short on the development of realistic quantitative versions of those models. This is largely because of their complexity, which often restricts the analysis to a small number of symmetric regions. Recent studies have started to fill this gap. In so doing, they have borrowed from the new literature on international trade in which the cross-country productivity distribution is endogenous; the literature reveals new sources of gains from trade under perfect competition

(Eaton and Kortum, 2002) and imperfect competition (Bernard et al. 2003; Melitz, 2003; Melitz and Ottaviano, 2008).[22]

A recent example of this work is Donaldson (2010), who develops a Ricardian trade model to study the effects of the Indian railway network, introduced by the British when India was under colonial rule. The model draws on Eaton and Kortum (2002) and features many regions, many commodities, and costly trade. Regions are assumed to have different productivity levels across commodities, generating opportunities to exploit comparative advantage through trade. When two regions become linked by a railway, their bilateral trade cost falls and this allows for specialization according to comparative advantage. The empirical implementation of the model allows Donaldson (2010) to quantify the extent to which the railway network improved India's trading environment, in terms of lower trade costs, smaller inter-regional price gaps, and larger trade flows between regions and internationally. Further, he also investigates how much of the estimated reduced-form welfare gains plausibly arise from newly exploited gains from trade. In particular, he finds that those gains account for virtually all of the observed reduced-form impact of railways on real income estimated from the data.

More generally, Redding (2012) also develops a tractable model of regional economic geography based on Eaton and Kortum (2002), suitable for quantitative investigations. He studies the general equilibrium of an economy with an arbitrary number of regions connected by an arbitrary pattern of geographical trade costs, modeled as iceberg transport costs that may differ between all bilateral pairs of regions. Labor is assumed to be mobile across regions. The productivity level of each region is drawn from a Fréchet distribution. Regions with higher productivity pay higher wages, which attracts population until the higher wages are offset by higher living costs. Regions with good market access (low transport costs) will have low prices for traded goods, and again this is offset by population movements that drive up housing costs. Hence, in equilibrium, welfare is equalized across locations, but regions that are productive or well situated have higher nominal wages, larger populations, and higher housing costs. The model is especially well suited for studying the effects on regional growth of an economy-wide trade liberalization, or more generally, a fall in external trade costs. A liberalization of trade will lead to an endogenous internal reallocation of population, implying a combination of regional growth and regional decline.

An alternative vein of research gives more attention to the relationship between first nature geography and the decisions of mobile, and heterogeneous, people and firms under imperfect competition. Melitz (2003) and Melitz and Ottaviano (2008) provide a basis for this class of models. A key distinction is whether the heterogeneous characteristics of agents are assumed to be revealed to them before, or after, their location decisions. Sorting

[22] For a discussion of these new sources of gains from trade see Arkolakis et al. (2012) and Melitz and Redding (2013).

models study how heterogeneous agents, aware of their characteristics *ex ante*, will sort themselves into locations of varying sizes (Nocke, 2006; Baldwin and Okubo, 2006; Davis and Dingel, 2012; Okubo et al. 2010; Picard and Okubo, 2012).[23] In contrast, selection models study what happens when heterogeneity materializes *ex post*, after agents have already committed to their locations: they can then self-select across whatever economic activities are available in those locations.

Behrens and Robert-Nicoud (2012) present a selection model where *ex ante* identical individuals decide whether to move from a common rural hinterland to cities. Their heterogeneity is revealed after this decision has been made, and the decision itself is assumed to be irreversible, which rules out sorting. They show that larger market size increases productivity partly through a finer division of labor driven by pecuniary externalities (richer availability of intermediates) and partly through a selection process. Meanwhile, higher productivity increases market size by providing incentives for rural-urban migration. Behrens et al. (2010) analyze both sorting and selection in a model where agglomeration is driven by technological externalities. They distinguish between *ex ante* heterogenity (talent), known to agents before they decide where to locate, and *ex post* heterogeneity (luck), revealed to agents after their location decisions have been made. Agents choose locations based on their talent, while luck influences subsequent occupational choices. More talented agents stand a better chance of finding more productive occupations in larger locations; this complementarity between talent and market size leads to the sorting of more talented agents into larger markets. Then, more demanding selection in more talented locations implies that average productivity is higher in these locations. Higher productivity, in turn, complements the agglomeration benefits of larger locations, and so markets with greater concentrations of talent are larger in equilibrium. Markups are constant, as in Melitz (2003). This implies that, conditional on sorting and agglomeration, selection becomes independent of market size.

Similarly to Behrens and Robert-Nicoud (2012), Ottaviano (2012) dispenses with *ex ante* heterogeneity (and first nature asymmetries) in order to investigate how firm heterogeneity influences the aggregate balance between agglomeration and dispersion forces, in the presence of pecuniary externalities. This is a selection model based on Melitz and Ottaviano (2008). A further departure from the analysis of Behrens and Robert-Nicoud (2012) is that the model allows location decisions to be reversible, and whether regions are characterized as urban or rural is determined endogenously by those decisions.[24] The emergence of agglomeration is driven by pecuniary rather than technological externalities. Markups are determined endogenously with larger market size leading to lower

[23] While other papers focus on firm heterogeneity on the supply side, in terms of productivity, the distinctive feature of Picard and Okubo (2012) is their study of heterogeneity on the demand side, in terms of tastes.

[24] Behrens et al. (2011) take a similar approach in their study of spatial frictions, allowing for the joint determination of location sizes, productivity levels, markups, wages, consumption diversity, and the number and size distribution of firms.

markups. This implies that, differently from Behrens et al. (2010) but as in Behrens and Robert-Nicoud (2012), selection is still more demanding in larger markets, even after conditioning out sorting and agglomeration.

Combes et al. (2012) also extend the model of Melitz and Ottaviano (2008) to allow for agglomeration economies driven by technological externalities. They estimate the relative importance of selection and agglomeration in determining the spatial distribution of firm productivity levels. Following Melitz and Ottaviano (2008), they rule out labor mobility across locations, although extensions to the basic framework can be made.[25] To distinguish between agglomeration and selection effects, they nest a generalized version of Melitz and Ottaviano (2008) and a model of agglomeration in the spirit of Fujita and Ogawa (1982) and Lucas and Rossi-Hansberg (2002). In larger (more dense) locations, the firm productivity distribution is left-truncated due to more demanding selection, but also right-shifted and dilated due to agglomeration. Combes et al. (2012) show how to estimate these effects by studying how the quantiles of the log productivity distribution in a large city will be related to the quantiles of the log productivity distribution in a small city. They estimate the relationship from data on French employment areas, and find no difference in the left-truncation of the log productivity distribution between dense and less dense areas. This suggests that the firm selection mechanism cannot explain spatial productivity differences across these areas. As they acknowledge, this result might not generalize to countries that are less well integrated than France, or where firms charge prices that differ across locations. Even for the French case, it does not rule out selection effects altogether, since their intensity could be the same across locations.

4.5. GEOGRAPHY AND GROWTH

We now consider growth models that make space for space: dynamic models of the growth process in which space plays a determining role. We have labeled this the "hard problem" of regional economics. Desmet and Rossi-Hansberg (2012a, p. 2–3) provide a clear statement of the problem:

> *Incorporating a continuum of locations into a dynamic framework is a challenging task for two reasons: it increases the dimensionality of the problem by requiring agents to understand the distribution of economic activity over time and over space, and clearing goods and factor markets is complex because prices depend on trade and mobility patterns. These two difficulties typically make spatial dynamic models intractable, both analytically and numerically.*

[25] In a separate online appendix (http://diegopuga.org/papers/selectagg_webapp.pdf), they show how their model can be extended to include worker mobility, consumption amenities, and urban crowding costs, without affecting the key equilibrium equations on which their empirical analysis is based. They restrict their attention to a situation in which there exists a unique stable spatial equilibrium with (asymmetric) dispersion. In contrast to Ottaviano (2012), whether heterogeneity fosters agglomeration or dispersion is beyond the scope of their paper.

One reason the problem becomes intractable is that, if we think of a dynamic model as one with forward-looking investment decisions, then agents must anticipate the solutions for future prices, and hence the equilibrium patterns of trade and mobility, at all future dates. As Desmet and Rossi-Hansberg note, the only way forward is to simplify the problem. Our review distinguishes between two families of models, dynamic NEG models and dynamic sequential market clearing (SMC) models. These vary in whether locations are ordered in space as in the real world. Though the task of combining agglomeration forces with interesting long-run dynamics and growth paths is far from complete, these two families of models represent the current frontier of theoretical research on regional growth. With this in mind, we set the ideas out in detail.

4.5.1 Non-Ordered Space

We first consider non-ordered space. An influential literature in trade theory has developed dynamic models with two or more countries. The relevant contributions include Grossman and Helpman (1991),Young (1991),Ventura (1997), Eaton and Kortum (1999), and Cuñat and Maffezzoli (2007). In these models, either autarky is compared to free trade or, when trade costs are introduced, countries are not ordered in space. From the viewpoint of spatial economics, the most attractive members of this family are the dynamic versions of NEG models. These typically feature a small number of locations (in most cases only two) that exchange goods and ideas in the presence of frictions. In the wake of Krugman (1991), localized pecuniary externalities drive the agglomeration of production. Endogenous growth is introduced by adding innovation in product variety with technological externalities, as in Grossman and Helpman (1991). The localized nature of these externalities, due to frictions in the exchange of ideas between regions, drives the agglomeration of innovation and can lead to cumulative causation in the location of production and innovation.

Baldwin and Martin (2004) survey several different specifications of these models and tease out their main insights. Cumulative causation implies the joint agglomeration of innovation and production. Aggregate growth is then driven by factor accumulation in a small subset of regions. This leads to growth poles and growth sinks. However, due to the localized nature of the technological externalities in innovation, the endogenous emergence of regional disparities is accompanied by faster aggregate growth and higher welfare in all regions.

Minerva and Ottaviano (2009) present a simple unifying model with two regions that encompasses a variety of insights from this line of research in a parsimonious way. It highlights the implications of geography for the dynamic process of regional growth. In this model, the geographical element arises partly due to costs of trading goods across regions (transport costs) and partly from barriers to exchanging ideas (communication costs). The model illustrates how agglomeration and growth can reinforce each other, giving rise to the cumulative causation that Myrdal envisaged.

An important limitation should be acknowledged at the outset, which is that the model rules out labor mobility across regions. Relaxing this assumption is not straightforward, as we discuss later. It has been relaxed in an alternative class of models, based on sequential market clearing. These models, reviewed later, can accommodate labor mobility, congestion in land use and a large number of regions, without sacrificing analytical tractability. At the same time, the account of growth in such models tends to be more stylized than the one we describe here.

Following Minerva and Ottaviano (2009), let us assume that there are two regions, north and south. To abstract from first nature, the exogenous attributes of the two regions are the same. First, they are populated by an identical number Q of geographically immobile workers. As each worker supplies one unit of labor inelastically, Q is also the regional endowment of labor. Second, regions are endowed with an identical initial stock of knowledge capital K_0. Through time, profit-seeking R&D laboratories create additional knowledge capital that is freely mobile between regions. In so doing, they finance their investments through bonds, with riskless return $r(t)$ at time t, sold to workers in a perfect inter-regional capital market. Henceforth, in the presentation of the model we will focus on north. Analogous expressions will apply to south.

Transport costs and localized spillovers play a key role in the analysis. Workers consume two goods, a homogeneous "traditional" good Y and a horizontally differentiated "modern" good D, with preferences given by the following utility function:

$$U = \int_{t=0}^{\infty} \log\left[D(t)^{\alpha} Y(t)^{1-\alpha}\right] e^{-\rho t} \mathrm{d}t. \tag{4.2}$$

In (4.2) $D(t)$ represents the CES consumption basket of the different varieties of good D:

$$D(t) = \left[\int_{i=0}^{N(t)} D_i(t)^{1-1/\sigma} \mathrm{d}i\right]^{1/(1-1/\sigma)}, \quad \sigma > 1, \tag{4.3}$$

where $D_i(t)$ is the consumption of variety i and $N(t)$ the total number of varieties in the economy.

Given a unit elasticity of intertemporal substitution, intertemporal utility maximization determines the evolution of expenditures according to the Euler equation:

$$\frac{\dot{E}(t)}{E(t)} = r(t) - \rho, \tag{4.4}$$

where $E(t)$ is individual expenditure. The Cobb-Douglas instantaneous utility function is then maximized when the shares α and $1-\alpha$ of individual expenditures $E(t)$ are allocated to the consumption of the modern and traditional goods respectively. In turn, the fraction $\alpha E(t)$ is distributed across the varieties of the modern good depending on their relative prices. This gives individual demand:

$$D_i(t) = \frac{p_i(t)^{-\sigma}}{P(t)^{1-\sigma}} \alpha E(t). \tag{4.5}$$

In (4.5) $P(t)$ represents the exact price index associated with the CES consumption basket (4.3):

$$P(t) = \left[\int_{i=0}^{N(t)} p_i(t)^{1-\sigma} \mathrm{d}i \right]^{1/(1-\sigma)}, \tag{4.6}$$

so that σ measures both the own- and the cross-price elasticities of demand.

The production of the traditional good is characterized by perfect competition and constant returns to scale with labor as its only input. An appropriate choice of units means that the unit labor requirement can be set to 1. This implies that the profit-maximizing price of Y equals the wage. The traditional good is assumed to be freely traded between and within regions. Hence, both its price and the wage are equalized across regions. Selecting good Y as the numéraire pins down the common wage to 1.

The production of the modern varieties is characterized by monopolistic competition and increasing returns to scale. These arise from the presence of a fixed cost incurred in terms of one unit of knowledge capital per variety. Variable costs are incurred, instead, in terms of β units of labor per unit of output. Due to the fixed capital requirement, at any instant t the total number of varieties available in the economy is determined by the aggregate knowledge capital stock $K^w(t)$. In equilibrium there is a one-to-one relation between firms and varieties, and so the total number of firms $N(t)$ is equal to $K^w(t)$. In turn, due to the free mobility of knowledge capital, the entry decisions of firms will determine where varieties are actually produced, and we use $n(t)$ to denote the number of northern firms and varieties. Entry is free, and at any given instant there are many potential entrants. These need knowledge capital to start producing. In the presence of a capital supply that is fixed at any given instant, competitive bidding by entrants transfers all operating profits to capital owners.

Geography is introduced in the product market, by assuming that trade flows of differentiated varieties face iceberg transport costs, within and between regions. The size of the internal transport cost differs across regions: in north and south, firms have to ship $\tau_N > 1$ and $\tau_S > 1$ units, respectively, in order to deliver one unit to their domestic customers. As for inter-regional trade, the delivery of one unit requires the shipment of $\tau_R > 1$ units, regardless of the direction of trade. Within-region shipments are less costly than inter-regional ones, and this cost advantage is more pronounced for north. Hence, we have $\tau_N < \tau_S < \tau_R$. This ranking of the transport cost parameters identifies north as the developed core and south as the developing periphery.

All firms in both markets face the same constant elasticity of demand σ and the same marginal production cost β. Hence, profit maximization leads to the same producer price (mill price) for all firms as a constant markup over marginal cost $p = \sigma\beta/(\sigma - 1)$. The corresponding consumer prices (delivered prices) simply reflect differential transport costs: $p_N = p\ \tau_N, p_S = p\ \tau_S, p_R = p\ \tau_R$. With these prices, operating profits are $\pi(t) = \beta x(t)/(\sigma - 1)$. Here, $x(t)$ denotes firm output inclusive of the quantity lost in transit, and the price index (4.6) can be rewritten as $P(t) = pN(t)^{\frac{1}{1-\sigma}} \left[\delta_N \gamma(t) + \delta_R(1-\gamma\,(t))\right]^{\frac{1}{1-\sigma}}$, where $\gamma(t) = n(t)/N(t)$ is the share of firms located in North and $N(t) = K^w(t)$ is the total number of firms as well as the total stock of knowledge capital. The parameters $\delta_N \equiv (\tau_N)^{1-\sigma}$, $\delta_S \equiv (\tau_S)^{1-\sigma}$, and $\delta_R \equiv (\tau_R)^{1-\sigma}$ measure the efficiency of internal and external transportation with $0 < \delta_R < \delta_S < \delta_N < 1$.

The national capital stock $K^w(t)$ is accumulated through profit-seeking R&D by perfectly competitive laboratories facing constant returns to scale. Knowledge spillovers are assumed to increase the productivity of researchers as knowledge accumulates, and this sustains growth in the long run. Geography is introduced in the knowledge capital market through a specification of the R&D technology that encompasses both localized knowledge spillovers (Martin and Ottaviano, 1999) and intermediate business services (Martin and Ottaviano, 2001) as captured by the following constant-returns-to-scale production function:

$$\dot{K}(t) = A(t) \left[\frac{D(t)}{\varepsilon}\right]^{\varepsilon} \left[\frac{Q_I(t)}{1-\varepsilon}\right]^{1-\varepsilon}, \tag{4.7}$$

where $\dot{K}(t) \equiv dK(t)/dt$ is the flow of knowledge created at time t, $Q_I(t)$ is labor employed in R&D, $D(t)$ is the basket of business services, and $\varepsilon \in (0,1)$ is the share of business services in R&D. Note that the basket of business services is assumed to be the same as the consumption basket, for analytical convenience. In (4.7) $A(t)$ refers to the North's total factor productivity in R&D and is assumed to be an increasing function of the total stock of knowledge $K^w(t)$ as embodied in the operations of modern producers. Specifically, the region-specific level of productivity in R&D is given by $A(t) = A\ K^w(t)^{\mu} \left[\omega_N \gamma(t) + \omega_R(1-\gamma(t))\right]^{\mu}$, where A is a positive constant. Here $\mu \in (0,1)$ measures the intensity of the knowledge spillovers, whose geographical diffusion is hampered by frictional communication costs. Their spatial decay is regulated by the ω's. It is assumed to be steeper between regions than within them, and steeper in south than in north, reflecting their different development stages. Hence, $0 < \omega_R < \omega_S < \omega_N < 1$. The larger ω, the lower the corresponding communication costs.

Both transport and communication costs create an incentive for innovation to cluster where production also disproportionately happens. To see this, we can use profit-maximizing prices and the equilibrium wage to compute the marginal cost associated with (4.7) as:

$$\begin{aligned} F(t) &= \frac{P(t)^{\varepsilon} w^{1-\varepsilon}}{A(t)} \\ &= \frac{\eta}{N(t)\left[\omega_N \gamma(t) + \omega_R(1-\gamma(t))\right]^{1-\frac{\varepsilon}{\sigma-1}} \left[\delta_N \gamma(t) + \delta_R(1-\gamma(t))\right]^{\frac{\varepsilon}{\sigma-1}}}, \end{aligned} \tag{4.8}$$

where $\eta = p^{\varepsilon}/A$ is a positive constant, and we have imposed the constraint $\mu + \varepsilon/(\sigma-1) = 1$ so that in the long run the economy follows a balanced growth path. This constraint preserves the incentive to invest in R&D in the long run, as the marginal cost of innovation decreases over time at the same rate as its benefit measured by the value of a firm.

Inspecting (4.8) reveals that, given the rankings of ω's and δ's, the marginal cost of innovation is lower in north provided it hosts a larger number of firms. As we will see, this is indeed the case in equilibrium, as lower internal transport costs increase the size of the local market. Hence, due to perfect competition among laboratories, in equilibrium they will all be located in north. Even though long-run growth is entirely driven by northern innovators, they are still financed in the inter-regional capital market by both northern and southern workers. This implies that, in equilibrium, the value $v(t)$ of a unit of knowledge capital has to satisfy an arbitrage condition. It requires the bond yield $r(t)$ to be equal to the percentage return on investment in knowledge capital, consisting of the percentage capital gain $\dot{v}(t)/v(t)$ and the percentage dividend $\pi(t)/v(t)$:

$$r(t) = \frac{\dot{v}(t)}{v(t)} + \frac{\pi(t)}{v(t)}, \tag{4.9}$$

where $v(t) = F(t)$ as, due to perfect competition in R&D, profit-maximizing laboratories price knowledge capital at marginal cost.

Finally, the model is closed by imposing that in equilibrium product and labor markets clear. Consider the product market first. Substituting the profit-maximizing prices into demands (4.5) allows us to state the market-clearing conditions for northern and southern firms as:

$$\begin{aligned} x(t) &= \frac{p^{-\sigma}\delta_N}{P(t)^{1-\sigma}}\left[\alpha E(t)Q + \varepsilon F(t)\dot{N}(t)\right] + \frac{p^{-\sigma}\delta_R}{P^{*}(t)^{1-\sigma}}\alpha E^{*}(t)Q, \\ x^{*}(t) &= \frac{p^{-\sigma}\delta_S}{P^{*}(t)^{1-\sigma}}\alpha E^{*}(t)Q + \frac{p^{-\sigma}\delta_R}{P(t)^{1-\sigma}}\left[\alpha E(t)Q + \varepsilon F(t)\dot{N}(t)\right], \end{aligned} \tag{4.10}$$

where an asterisk flags southern variables. Only northern demand is augmented by intermediate expenditures $\varepsilon F(t)\dot{N}(t)$ as R&D is active only in north. Turning to the labor market, this clears when the total endowment of labor $2Q$ is fully employed in innovation $Q_I(t) = (1-\varepsilon)F(t)\dot{N}(t)$, in modern production $Q_D(t) = [(\sigma-1)/\sigma]\,[2\alpha E(t)Q + \varepsilon F(t)\dot{N}(t)]$, and in traditional production $Q_Y(t) = 2(1-\alpha)E(t)Q$:

$$2Q = \frac{\sigma-\varepsilon}{\sigma}F(t)\dot{N}(t) + 2\frac{\sigma-\alpha}{\sigma}E(t)Q. \tag{4.11}$$

We now study agglomeration and growth. The market clearing conditions for products and labor can be used to highlight how growth affects location and, vice versa, location affects growth. We focus on a balanced growth path with constant expenditures and a constant growth rate of knowledge capital $g = \dot{K}^{w}(t)/K^{w}(t) = \dot{N}(t)/N(t)$. Constant expenditures imply $\dot{E} = 0$ so that, given (4.4), we have $r = \rho$. Further, FN and γ are constant, and hence the evolution of the value of knowledge capital is determined by the growth rate of knowledge capital through the implied change in the marginal cost of R&D, $\dot{v}/v = \dot{F}/F = -g$. In other words, the marginal benefit of innovation (v) and its marginal cost (F) both fall at the same constant rate.

The arbitrage condition (4.9) implies that, in equilibrium, all firms achieve the same level of profits and hence the same scale of output wherever they are. Then we can use (4.10) to determine this common output scale as:

$$x = [(\sigma - 1)/\beta\sigma]\,[(2\alpha EQ + \varepsilon FNg)/N]\,. \tag{4.12}$$

This can be used to rewrite (4.9) as a function of E, g, and FN. The resulting expression can be solved together with labor market clearing (4.11) to show that, in equilibrium, expenditure equals permanent income:

$$2EQ = 2Q + \rho FN, \tag{4.13}$$

and the growth rate satisfies:

$$g = \frac{\alpha}{\sigma - \varepsilon}\frac{2Q}{FN} - \rho\frac{\sigma - \alpha}{\sigma - \varepsilon}. \tag{4.14}$$

Substituting (4.8) into (4.14) shows that location affects growth through the marginal cost of innovation FN net of the spillover from accumulated knowledge capital:

$$g = \frac{\alpha}{\sigma - \varepsilon}\frac{2Q}{\eta}\left[\omega_N\gamma + \omega_R(1-\gamma)\right]^{1-\frac{\varepsilon}{\sigma-1}}\left[\delta_N\gamma + \delta_R(1-\gamma)\right]^{\frac{\varepsilon}{\sigma-1}} - \rho\frac{\sigma-\alpha}{\sigma-\varepsilon}. \tag{4.15}$$

In particular, more agglomeration in north makes innovation less costly and hence leads to faster growth.

The joint solution of the product market clearing conditions (4.10) determines not only the firms' common output scale, but also the share of northern firms, as:

$$\gamma = \frac{1}{2} + \frac{1}{2}\frac{\delta_R(\delta_N - \delta_S)}{(\delta_N - \delta_R)(\delta_S - \delta_R)} + \frac{\delta_N\delta_S - \delta_R^2}{(\delta_N - \delta_R)(\delta_S - \delta_R)}\left(\theta - \frac{1}{2}\right). \tag{4.16}$$

In (4.16) $\theta = (\alpha EQ + \varepsilon FNg)/(2\alpha EQ + \varepsilon FNg)$ is the northern share of expenditures in the modern sector, after taking into account that $E = E^*$ since regions share the same initial endowments. It depends on the endogenous variables E, FN, and g. However, using (4.14) it can be expressed as a function of g only, thus allowing us to rewrite (4.16) as:

$$\gamma = \frac{1}{2} + \frac{1}{2}\frac{\delta_R(\delta_N - \delta_S)}{(\delta_N - \delta_R)(\delta_S - \delta_R)} + \frac{1}{2}\frac{\delta_N\delta_S - \delta_R^2}{(\delta_N - \delta_R)(\delta_S - \delta_R)}\frac{\varepsilon}{\sigma}\frac{g}{g+\rho}. \tag{4.17}$$

This shows that growth affects location through its influence on the northern share of expenditures. In particular, faster growth increases the northern expenditure share as innovation takes place only in north, which leads more firms to locate there.

Expressions (4.15) and (4.17) highlight a crucial result: agglomeration (larger γ) and growth (larger g) are jointly determined. Although the two do not interact dynamically, this can still be seen as a form of cumulative causation: forces which promote growth indirectly promote agglomeration, and vice versa. The outcome is a trade-off for policy-makers, between promoting growth and reducing regional disparities. Further insights into the role of geography are readily gained by focusing on two extreme cases that arise when the cost of innovation is determined by communication costs only ($\varepsilon = 0$) or by transport costs only ($\varepsilon = \sigma - 1$) as in Martin and Ottaviano (1999, 2001), respectively. If $\varepsilon = 0$, lower communication costs within north foster growth but have no impact on agglomeration. The same applies to lower inter-regional communication costs. In contrast, lower communication costs in south have no impact as long as no innovation takes place there. Moreover, changes in transport costs affect location, but have no impact on growth. If $\varepsilon = \sigma - 1$, reductions in inter-regional and intra-north transport costs promote agglomeration in north as well as growth; reductions in intra-south transport costs promote relocation from north to south, but also hamper growth.

As we noted previously, this analysis has ruled out labor mobility, which is hard to accommodate in multi-region endogenous growth models. In principle, mobility could be introduced as in Fujita and Thisse (2003) but, absent congestion in land use, this would simply lead to the clustering of all factors in a single region. Allowing for congestion in land use could avoid this outcome, but leads to a model that is analytically intractable. Studies that allow for labor mobility in a multi-region endogenous growth model, under perfect foresight, include Walz (1996) and Baldwin and Forslid (2000). As discussed by Fujita and Thisse (2002), the assumption of costless migration in Walz (1996) leads to bang-bang behavior that does not accord with reality. Migration is gradual in Baldwin and Forslid et al. (2003), at the expense of analytical complexity. For reasons of tractability, Fujita and Thisse (2003) focus on a steady-state equilibrium in which the spatial distribution of skilled workers is time-invariant. Although they provide a stability analysis, the details of the transition process are not studied.

4.5.2 Ordered Space

Dynamic NEG models enhance our understanding of the common forces underlying growth and agglomeration. As argued by Desmet and Rossi-Hansberg (2010), however, their focus on a small number of locations misses the richness of the observed geography of economic activities, and limits their empirical applications. Generalizing them to more than a few regions introduces problems of analytical tractability, especially when one allows for frictions in the mobility of capital (Baldwin et al. 2001) or labor (Fujita and Thisse, 2003). Some progress could still be made through numerical methods, as shown by Fujita et al. (1999) for static models in a continuous space, but work in this vein remains limited.

A small number of papers study a fully dynamic setup with a continuum of locations: these include Brito (2004), Brock and Xepapadeas (2008, 2009), and Boucekkine et al. (2009). They typically focus on the allocation problem of a social planner but, absent more structure, it is hard to extract general insights. The main problem is that, in order to make decisions, forward-looking agents need to understand the whole distribution of economic activities over space and time implied by each feasible action.

Desmet and Rossi-Hansberg (2010, 2012a) advance an alternative approach, initially proposed in Rossi-Hansberg (2005), that is analytically tractable when space is continuous and one-dimensional. To reduce the complexity of the problem, they model a situation in which agents do not have to consider the future allocation paths, because these paths are beyond their control and do not affect their returns from current decisions. Hence, though forward-looking, agents solve static problems. This is achieved by imposing enough structure on the diffusion of technology or on the mobility of agents and the way property rights over land are allocated among them. This approach generates a dynamic process in which locations continuously change in occupational structure and employment density, but the aggregate economy converges to a balanced growth path.

In a simplified version of their model, Desmet and Rossi-Hansberg (2010) study an economy in which all markets are perfectly competitive; locations accumulate technology by investing in innovation in one homogeneous-good industry and by receiving spillovers from other locations; factor mobility is frictionless; and trade is the result of agents holding a diversified portfolio of land across locations. Land is given by the unit interval $[0, 1]$, time is discrete, and total population is $\overline{L}$. The one-dimensional space $[0, 1]$ is divided in connected intervals (counties), each administered by a local government.

Consumers-workers in location l solve the utility maximization problem:

$$\max_{\{c(l,t)\}_0^\infty} E \sum_{t=0}^{\infty} \beta^t U\left(c(l,t)\right),$$

subject to:

$$w(l,t) + \frac{\overline{R}(t)}{\overline{L}} = p(l,t)c(l,t)\, \forall l,t,$$

where $U\left(c(l,t)\right)$ is the instantaneous utility of consumption $c(l,t)$ in period t, β is the discount factor, and E is the expectation operator. Consumption incurs a price $p(l,t)$. Income consists of the wage $w(l,t)$ and the share $1/\overline{L}$ of total land rent $\overline{R}(t)$ under the assumption that consumers hold a fully diversified portfolio of land across locations. Due to free labor mobility, in each period t, utility is the same everywhere.

Production employs labor and land with technology:

$$x\left(L(l,t)\right) = Z(l,t)L(l,t)^{\mu},$$

where $x\left(L(l,t)\right)$ is output per unit of land, $Z(l,t)$ is total factor productivity, and $L(l,t)$ is employment per unit of land with $\mu \in (0,1)$. The profit-maximization problem of a firm can be stated as:

$$\max_{L(l,t)} (1 - \tau(l,t)) \left[p(l,t)\, Z(l,t) L(l,t)^{\mu} - w(l,t) L(l,t) \right],$$

where $\tau(l,t)$ is a tax on profits, levied by the local government of the county to which location l belongs, in order to finance investment in process innovation leading to an improved level of total factor productivity equal to $z_l Z(l,t)$.

In particular, the local government can buy a probability $\phi \in [0,1]$ of innovating at a cost $\psi(\phi)$ per unit of land proportional to wages, with $\psi'(\phi) > 0$ and $\psi''(\phi) > 0$. Successful innovation allows the government to draw z_l from a Pareto distribution with c.d.f. $F(z) = 1 - z^{-a}$ with $z \geq 1$. Under the assumption of risk neutrality, the local government of county G with land measure I then solves:

$$\max_{\{\phi(l,t)\}_{l \in G}} \int_{l \in G} \frac{\phi(l,t)}{a-1} p(l,t) Z(l,t) L(l,t)^{\mu} dl - I\psi(\phi(l,t)), \tag{4.18}$$

where $\psi(\phi(l,t))$ is government investment in location l at time t, $\phi(l,t)$ is the probability that the government gets to draw from $F(z)$ in location l at time t, $1/(a-1)$ is the expected value of the total factor productivity gain for location l at time t conditional on the government getting to draw from $F(z)$ in that location. In other words, the local government spends on R&D to maximize the expected increase in the output value of its county net of the investment cost. The fact that the maximization problem (4.18) is static follows from a key assumption on the diffusion of innovation, which makes the best technology available to all neighboring locations with a one-period delay with respect to the innovator. Matched with the assumption that counties are small, the one-period delay implies that a county's innovation decision today does not affect its expected level of technology tomorrow. Interestingly, (4.18) exhibits a scale effect as high-price, high-productivity, and high-employment density locations will optimally innovate more.

As in the dynamic NEG framework presented earlier, Desmet and Rossi-Hansberg (2010) introduce geography through communication and transport frictions that hamper the geographical mobility of goods and ideas. For ideas, at time t, before the innovation decision, location l has access to the best spatially discounted technology available of the previous period, so *ex ante* $Z(l,t)$ equals:

$$Z^{-}(l,t) = \max_{r \in [0,1]} e^{-\delta|l-r|} Z(r, t-1),$$

where $\delta > 0$ measures the steepness of the spatial decay of diffusion. Based on this technology consumers costlessly relocate, which ensures that utility is the same across all locations, and wages are set. The fact that consumers hold fully diversified portfolios of land in all locations implies that they need not be forward-looking when deciding where to locate. After consumers move, counties invest in innovation, and production takes place using the new technology $Z^{+}(l,t)$ so that *ex post* $Z(l,t)$ equals $Z^{+}(l,t)$.

Due to land portfolio diversification, rents are redistributed from high-productivity to low-productivity locations, which therefore run trade surpluses and deficits respectively. Turning to the product market, transport costs again take the iceberg form: if one unit of the product is shipped from l to r, only $e^{\kappa|l-r|}$ units reach their destination. Hence, with perfect competition we have $p(r,t) = e^{-\kappa|l-r|}p(l,t)$.

In equilibrium, labor and product markets clear. In the case of labor, at each point in time the market clearing condition is:

$$\int_0^1 L(l,t)dl = \overline{L}.$$

The market clearing condition in the product market is less straightforward. Following Rossi-Hansberg (2005), it is stated sequentially. In particular, one can start at one end of the one-dimensional space interval and accumulate production minus consumption in a given market (properly discounted by transport costs) until one reaches the other end of the interval. At the boundary, for markets to clear, excess supply has to be equal to zero. Formally, let $H(l,t)$ define the stock of excess supply accumulated from location 0 to location l. By construction, $H(l,t)$ is defined by the initial condition $H(0,t) = 0$ and the differential equation:

$$\frac{\partial H(l,t)}{\partial l} = x(l,t) - c(l,t)L(l,t) - \kappa\left|H(l,t)\right|,$$

where $x(l,t) = x(L(l,t) - \psi(l,t)/p(l,t),t)$ so that, at each location, we add to the stock of excess supply the amount of local output and subtract the amount of local consumption. We then need to adjust for the fact that if $|H(l,t)|$ is not zero and we increase l, we have to ship the stock of excess supply over a longer distance. This implies a per-unit cost in terms of the good equal to κ due to the iceberg transport costs. In the end, the good market clears if $H(1,t) = 0$.

At any period t, the instantaneous equilibrium of this economy can be computed easily. Before innovation takes place, workers decide where to live. Although the realizations of innovation are random, counties are small, so that there is no aggregate uncertainty. This allows workers to anticipate prices correctly. In addition, workers observe wages and land rents. Once innovation is realized, one can compute actual production, actual distributed land rents, and trade. The resulting prices should then be consistent with those used by workers when they decided where to live. Since decisions depend only on current outcomes, computing an equilibrium involves solving a functional fixed point each period. The dynamic growth process is determined by the sequence of those static points.

As usual, the spatial distribution of producers and workers results from the balance between agglomeration and dispersion forces. The diffusion of technology promotes agglomeration, as high levels of local employment raise the incentives to innovate. Due to spatial decay in the diffusion of innovation, productivity is higher in locations close to high-employment clusters, which attracts employment and fosters more innovation. This agglomeration force is opposed by local congestion, as employment density reduces labor productivity. This arises because, with constant returns to labor and land, and given that land cannot be accumulated locally, there are local diminishing returns to labor. This form of local congestion tends to spread employment across locations given identical technology levels.

Growth is linked to geography because more uniform, but weaker incentives to innovate are associated with dispersion; whereas agglomeration is associated with fewer, but more active innovation centers. As a result, when activity is spatially dispersed, innovation relies more on the extensive margin (how many locations innovate) whereas the intensive margin (how much each location innovates) plays a key role when activity is agglomerated. Easier diffusion makes the extensive margin less important and aggregate growth is generally higher with agglomeration.

Growth is also higher for higher transport costs, as these lead to more concentrated production. In this respect, higher transport costs entail static losses but dynamic gains, through more agglomeration and thus innovation. This is different from the NEG framework discussed earlier, in which higher transport costs promote dispersion and slower growth. The difference is explained by the fact that Minerva and Ottaviano (2009) do not model locally non-reproducible land, so that no congestion arises from its use. This parallels the opposite predictions of the static models of Krugman (1991) and Helpman (1998) discussed earlier in the chapter.

The model of Desmet and Rossi-Hansberg (2010) implies that the concentration of employment in neighboring locations leads to more innovation and faster growth. This effect is due to local density in a given location, and diffusion from locally dense neighbors. Desmet and Rossi-Hansberg (2012a) present a more general version of their framework in which two industries, manufacturing and services, interact because of trade. This extension reveals another channel through which agglomeration and growth are connected. Due to perfect competition, locations specialized in manufacturing exhibit higher producer (mill) prices of services. This happens because low transport costs in serving local consumers in the manufacturing cluster allow service providers in those locations to remain competitive in terms of customer (delivered) prices, despite higher producer prices. Manufacturing clusters will therefore have an incentive to import services from other locations. Their demand for imported services will, however, fall with distance due to growing transport costs, so that locations closer to manufacturing clusters will tend to have higher employment, higher prices, and greater innovation in services. Accordingly, the co-agglomeration of different industries is an additional source of local growth and innovation. This trade channel works on top of diffusion, and is reminiscent of the distinction between transport and communication costs drawn by Martin and Ottaviano (1999, 2001) in their dynamic NEG models.

In a quantitative exercise, Desmet and Rossi-Hansberg (2012a) show that their model can help to explain the evolution of the US economy over the last half-century. In particular, it can generate the reduction in the manufacturing employment share, the increased spatial concentration of services, the growth in service productivity starting in the mid-1990s, the rise in the dispersion of land rents in the same period, and several other spatial and temporal patterns.

In contrast to the model we presented above, where innovation is decided by local governments, Desmet and Rossi-Hansberg (2012a) explicitly model innovation as the outcome of firms making profit-maximizing choices. To produce, firms need to compete for non-replicable land. Since innovation can increase the productivity of that non-replicable land, firms realize they can enhance their bid for land by innovating. As a result, firms may optimally choose to innovate, in spite of the market being perfectly competitive and all profits being bidden away through land rents. The role of land in generating innovation in a perfectly competitive environment is discussed in further detail in Desmet and Rossi-Hansberg (2012b).

Moreover, Desmet and Rossi-Hansberg (2012a) show how the reallocation of employment toward services ultimately accelerates innovation in some locations specializing in services; from then onwards, service productivity increases together with manufacturing productivity, leading to a balanced growth path. Hence, their model is a full-fledged endogenous growth model with spatial heterogeneity, and one that can accommodate both structural transformation and a balanced growth path. The methods that Desmet and Rossi-Hansberg (2010, 2012a) use to deal with growth in an ordered geographical space are fairly straightforward to apply, relative to the underlying complexity of the problem. However, they can only be used in one-dimensional (or two-dimensional and symmetric) compact geographical spaces, and extending this approach to non-symmetric, two-dimensional space would be a challenge.

4.6. REGIONAL PROSPERITY: DATA AND METHODS

A common thread runs through many of the models we have considered: what happens at each location is a function of the outcomes and characteristics of all other locations. This raises a formidable identification problem for empirical researchers who want to isolate causal mechanisms, and the available empirical methods differ in how persuasively they achieve this. This section will first discuss the available data, and then some leading methods. Some of the most important studies are based on natural experiments, with estimates often obtained by difference-in-differences; since these methods are well known, we do not cover them in detail. For an extended discussion of the natural experiment approach in regional economics, see Holmes (2010). Some examples, and discussion, can be found in Diamond and Robinson (2010).

4.6.1 Data

Historically, one obstacle to work on regional growth has been the scattered nature of the available data. Researchers on national growth have long been able to draw on the Penn World Table and the World Development Indicators, but there is no close equivalent for sub-national data. Recently this has begun to change, in contributions by Gennaioli et al. (2013a,b), Lessmann (2011), and Mitton (2013). The regional data sets of Gennaioli et al. (2013a) and Mitton (2013) are especially comprehensive; the first covers 1569 regions from 110 countries, which together account for 74% of the world's land area and 97% of its GDP. Mitton's data set is broadly similar in coverage, but partially corrects for internal variation in the cost of living, using data on living costs compiled for a number of cities by the Economic Research Institute. More detailed data on output deflators and regional living costs are typically unavailable, however, as we discuss in the appendix.

These data sets are cross-sections; Lessmann (2011) has compiled a panel data set on regional inequality, but for a smaller number of countries. For a few countries, long-run data sets have been compiled going back to the 19th century, such as the work of Turner et al. (2007) and Mitchener and McLean (1999, 2003) on US data; the latter papers use some data on prices. For some countries, the populations of cities have been used to proxy regional development over centuries; see Acemoglu et al. (2011) and Cantoni (2010) for examples and references.

The increased availability of establishment-level data for some countries can be used to address some research questions. Another recent development is the ability to analyze data at smaller spatial scales even for developing countries. Harari and La Ferrara (2013) illustrate the potential of this approach: they study civil conflict in Africa at the sub-national level, based on areas that are 1 degree of latitude by 1 degree of longitude, and relating conflict to localized crop failures or climate shocks. Moving to a smaller scale requires a careful approach to spatial dependence and clustering; Barrios et al. (2012) is a recent treatment of this issue. One way to use data at small scales is to aggregate them up to a regional level, the origin of some variables in the Gennaioli et al. (2013a) and Mitton (2013) data sets. For discussion of the use of geographical information systems in regional economics, see Overman (2010).

One approach of particular interest, emerging from an interdisciplinary research effort, has been to use satellite data on light density at night to develop measures of income or population density at the sub-national level. As Chen and Nordhaus (2011) and Henderson et al. (2012) emphasize, this is especially attractive for measuring growth in countries where spatially-disaggregated statistics are unreliable or not available. One application would be to map changes in regional income for countries where hard-to-measure activity, like subsistence agriculture or an urban informal sector, is significant. Relative to the use of official data, the approach also allows population density and income to be estimated for smaller spatial scales. For example, using data on light density for 22,850 sub-national units in developing countries, Hodler and Raschky (2010) study whether foreign aid is disproportionately allocated to the home regions of national leaders. Michalopoulos and Papaioannou (2013) use light density to study whether regional outcomes in Africa are related to pre-colonial institutions, and the local traditions of political centralization in particular.

Measurement issues, especially for developing countries, require thought about how the data relate to the research questions of interest and the concepts used in theoretical models. For example, flows of remittances between regions, which can be significant, will influence regional income. The measured output of some regions can be heavily influenced by natural resource revenues, which will typically be transferred out of the region. To give a concrete example, the treatment of mining output for Indonesia influences findings about regional inequality and convergence (Hill et al. 2008). These points also suggest the importance of considering whether the data at hand correspond most closely to the regional equivalent of GDP (output), or GNP (income). The former is most relevant for productivity comparisons, the latter for studying regional differences in living standards. Whichever concept is adopted, measurement error in regional data is likely to be a significant problem, and its consequences remain under-explored by applied researchers.[26]

4.6.2 Spatial Econometrics

If we recognize that regions are interdependent, statistical analysis has to proceed carefully. Outcomes at one location (for example, for productivity) will be closely linked to the outcomes and characteristics of other regions. This implies that the data-generating process will be characterized by spatial dependence; ignoring this dependence is risky, which is clear from a time-series analogy. A good econometrician knows that serial correlation is not solely an issue for inference, but often indicates that the empirical model has been misspecified. This is why econometricians are wary of mechanical autocorrelation corrections, or exclusive reliance on clustering the standard errors. Related points apply to spatial data, and yet many economists continue to analyze regional data as if spatial dependence is a second-order problem. In fairness, it is true that spatial dependence is inherently harder to address than time-series dependence, because the one-dimensional ordering in time does not apply in the spatial case.

The field of spatial econometrics typically addresses this problem by pre-specifying the relative strengths of interactions between regions, using the device of a spatial weight matrix. The entries in the matrix are often based on distances between locations or the existence of shared borders, although there is nothing in the approach which requires the interactions to be determined by physical geography. The literature is large and growing fast, and we highlight only the areas most relevant to the discussion later in the chapter.

[26] Additional measurement issues are discussed in some of the contributions in Kanbur and Venables (2005). Measurement errors are likely even in the official data of developed countries. Cameron and Muellbauer (2000) examine this issue for the UK, by comparing the UK's Regional Accounts with alternative sources of information on earnings.

This brief introduction draws partly on Anselin (2001) and especially the longer survey by Anselin (2006).

For the case of N regions, a cross-section model with a spatial lag is conventionally expressed in matrix notation as:

$$y = \rho Wy + X\beta + \varepsilon, \tag{4.19}$$

where W is an $N \times N$ spatial weight matrix (typically normalized in some way) and ρ indexes the strength of the spatial spillovers. Given $\rho \neq 0$, the spatial lag will be correlated with the disturbances ε, because the above model implies:

$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\varepsilon. \tag{4.20}$$

Expanding each inverse implies that y at each location is a function of X and ε at all locations, so that the effects of the explanatory variables and the errors are transmitted across space rather than confined to each region. A corollary is that Wy in (4.19) is necessarily endogenous and hence OLS estimates of that model will be inconsistent. The literature has developed alternative procedures for estimating such models, using either maximum likelihood or instrumental variables.

Alternatively, we could allow for spatial dependence in the errors rather than in the dependent variable, using the spatial error model:

$$y = X\beta + u, \tag{4.21}$$

$$u = \lambda Wu + \varepsilon. \tag{4.22}$$

This is more closely related to a spatial lag model than it may seem. If we note that $u = (I - \lambda W)^{-1}\varepsilon$ we can use this in $y = X\beta + u$ and then have:

$$y = X\beta + (I - \lambda W)^{-1}\varepsilon,$$

which implies:

$$y = \lambda Wy + X\beta - \lambda WX\beta + \varepsilon.$$

This model could be estimated with or without the implied parameter restrictions. This is usually referred to as the spatial Durbin model, by analogy to the derivation of common factor restrictions in time series models by Durbin (1960). One interpretation is that a spatial lag helps to address the issue of omitted variables that are spatially correlated, but this is only true if the spatial dependence corresponds to the relative interactions embedded in the weight matrix W.

There are two main interpretations of what spatial econometrics achieves. The first is that the spatial dependence is not itself of direct interest, but must be addressed to obtain reliable estimates of the parameters. In practice, some parameters become much harder to interpret when a model incorporates spatial spillovers. A common example would be attempts to link parameter estimates to the rate of convergence in a neoclassical growth model. When regional income levels are influenced by the income levels of neighboring regions, the theoretical counterpart of an estimated convergence rate is unclear, because the neoclassical growth model sits uneasily with the reality of interdependent regions.

An alternative interpretation is that spillovers are of direct interest. In that case, the estimate of ρ is seen as directly informative and not just a nuisance parameter. The problem here is that spatial econometric models are silent on mechanisms, and without a mechanism, adding a spatial lag of the dependent variable will often seem too ad hoc to be informative; Gibbons and Overman (2012) argue along these lines. They suggest that, for many applications, it would be more sensible to emphasize spatial lags of the explanatory variables. That approach is simpler to implement, and often easier to connect to theoretical models.

Another frequent criticism of the spatial approach is that the researcher's choice of weight matrix W is necessarily arbitrary, because there are many different possibilities. This criticism might sometimes go too far. There is a sense in which imposing $\rho = 0$ is an arbitrary choice too. Even a model with a mis-specified weight matrix may have better properties than a model which does not acknowledge spatial dependence at all. Approaches based instead on structural models, such as the use of measures of market potential, also impose restrictions on the data that are best seen as maintained assumptions, and that are open to question. Given the inevitable uncertainty over the appropriate weight matrix, one way to make the analysis less arbitrary is to use Bayesian Model Averaging, as in Crespo Cuaresma and Feldkircher (2013) and LeSage and Fischer (2008). This allows a range of specifications to be considered, while formally acknowledging the researcher's uncertainty about the model and the nature of the spatial interactions.

As things stand, there are clear divisions in the literature about the usefulness of these methods. Corrado and Fingleton (2012), Gibbons and Overman (2012), and LeSage and Fischer (2008) provide extensive discussion, from a variety of perspectives. That opinion is divided can be seen from the different paths taken in the applied literature. The spatial econometric papers take care over dependence, but often adopt rather mechanical hypotheses about regional growth and the nature of spillovers. In contrast, many papers by growth economists and development economists put forward interesting hypotheses, but largely ignore the issue of spatial dependence, or adopt corrections such as spatially clustered standard errors that do not address underlying problems with the regression specification. One improvement would be to adopt a spatial equivalent to HAC estimators of standard errors, such as that developed by Kelejian and Prucha (2007); but this continues to emphasize the problems for inference rather than the structure of the estimated model.[27]

In recent panel data studies, a common approach to error dependence has been to interact time dummies with one or more regional characteristics. Versions of this are adopted in Acemoglu et al. (2011), Burgess and Pande (2005), Burgess et al. (2005), and

[27] In Conley (1999), if an overidentified GMM approach is taken, spatial dependence is an issue for estimation as well as inference. More generally, a model which does not allow for spatial dependence is likely to be incomplete, again suggesting that spatial dependence matters for point estimates as well as standard errors.

Cantoni (2010) among others. We call this an assumption of proportional time effects. It can be seen as a special case of the common factor structures studied in the macroeconometric literature, where the error term has a component $\phi_i f_t$. Here, f_t is a vector of common factors, the effects of which are allowed to vary across the regions i by means of the (row) vector of factor loadings, ϕ_i. This could be a natural route to take for regional data. For example, an urban core of manufacturing and services might be strongly correlated with the national business cycle, while an agricultural periphery would be less correlated with the business cycle and more strongly correlated with climate variation and world food prices. In principle, a factor structure could account for much of the spatial dependence in the data. So far, there has not been much work analyzing regional data using these methods, but the techniques are developing rapidly and could be important for regional panel data models in particular. For surveys, see Eberhardt and Teal (2011) and Sarafidis and Wansbeek (2012).

4.6.3 Regional Growth Regressions

A substantial fraction of the work on regional prosperity, especially that for developing countries, is based on cross-section or panel data growth regressions. Assessed as a whole, the literature inherits many of the issues of interpretation that have undermined the cross-country study of economic growth. As Durlauf et al. (2005, p. 558) argue, the problem is not only that some regression-based studies are unreliable. A further problem arises on the consumption side: it can be hard, when presented with a particular study, to tell whether it has been executed well or badly. This means that even the best studies may be assigned relatively little weight.

Relative to the cross-country literature, the use of regional data may be much less vulnerable to omitted variables. A common argument is that factors such as institutions and cultural norms vary greatly across countries, but less so within them. But the fact that regions are within the same polity can be a double-edged sword, because they may influence each other, and be subject to common shocks, to a much greater extent than countries. In some ways, the legacies of the cross-country literature have been unfortunate. Empirical studies often treat the units as essentially independent, or take the neoclassical growth model as the starting point, either explicitly or implicitly; this approach has problems at the country level, and seems all but untenable for regional data.

As already noted, many of the regression-based studies by growth economists and development economists fail to address spatial dependence. It is common for researchers to analyze variables or interventions which are highly correlated spatially, but the estimates may then be confounded by omitted spillovers, or spatially correlated variables such as aspects of physical geography or market access. At least some of these are known to be important features of the data, and there could be gains from combining the hypotheses of these studies with methods from spatial econometrics and macroeconometrics.

Another fundamental issue receives even less emphasis in the literature: the basic causal structure implicit in a regression sits uneasily with a spatial equilibrium. The most obvious and well-known problem is that many regional characteristics, such as average education or financial depth, are not fixed endowments, but endogenously determined outcomes. But this also hints at a deeper identification problem, and one that has been less widely noted. At first glance, regression-based methods give simple answers about the determinants of regional prosperity. But their interpretation is complicated by endogenous agglomeration. When a given variable changes, this could reconfigure spatial patterns of activity in ways that (for example) amplify the effects of minor differences, just as agglomeration can amplify minor differences in physical geography. This makes it hard to interpret estimated associations between regional growth and explanatory variables. To make this point concrete, consider the estimated growth effect β of a one-unit change in a given variable X for one region. If X increased by the same amount for all regions, would the growth effect be β for each region? This is rarely clear, but then it is hard to interpret the results from regression-based studies. Put differently, it is not clear what is being assumed about the simultaneous role of changes in the spatial distribution of activity. In the context of a spatial equilibrium, this attempted distinction is artificial and impossible to maintain, but that is precisely the point. It complicates the interpretation of much empirical work, not least its consequences for policy. It also indicates the benefits of structural models, where meaningful counterfactuals are much easier to construct.

Another distinctive feature of regional data is that the cross-section dimension and the time dimension often have broadly similar magnitudes. This suggests that panel time-series methods, such as those introduced by Pesaran and Smith (1995) and Pesaran et al. (1999), could be natural candidates for the analysis of regional data. For reasons that are not fully clear, few studies have applied these methods to regional questions; exceptions include Cameron and Muellbauer (2001) for Britain, and Trivedi (2006) for India. The first of these briefly explores time-series specifications in which the dependent and independent variables are formed as deviations from the average values of contiguous regions.

4.6.4 Structural Models

We have repeatedly emphasized the dangers of analyzing spatial data without thinking in terms of a spatial equilibrium. That might suggest abandoning regressions in favor of calibrating or estimating structural models, often drawing heavily on the work we described in Section 4.4 of the chapter. Examples of this approach include Donaldson (2010), Redding (2012), and Van Nieuwerburgh and Weill (2010). The quantitative use of structural models has many advantages: general equilibrium effects are accounted for, parameters should have a clear interpretation, and progress can be made even when some data (such as regional price levels or productivity levels) are lacking, by inferring these from other outcomes. Further, the use of a structural model allows counterfactual simulations and the quantification of welfare effects, both of which are attractive when policies are to be assessed. Holmes (2010) emphasizes that the approach can be used to evaluate policies that have never previously been implemented. He discusses the approach further, as does Combes (2011).

We review several of these studies below. If structural models have a weakness, it is the uncertainty over whether it is the model speaking or the data; the list of maintained assumptions is often extensive, and the data may know more than the model can say. There is a complementary role for reduced-form approaches, partly in drawing attention to interesting associations, and partly as a check on the maintained assumptions of any given structural model, as in Donaldson (2010). But it seems clear that structural models will have an important, even pre-eminent, place within the best future work on regional data.

4.6.5 Spatial Discontinuity Designs

We now discuss a method which has become a powerful way of identifying causal effects in the recent literature, and which drives some of the most important papers. This is to look for institutions or policies that sometimes stop (or change) at the borders of regions, and quantify their causal effects by comparing outcomes either side of the border. We illustrate the application of this method, and some of its pitfalls, using the classic paper by Holmes (1998).

Holmes was interested in whether state-level policies influence the location of manufacturing activity. It had long been known that manufacturing activity had grown slowly in the industrial north of the US and more rapidly in other regions, including in the right-to-work states which weakened unions by state legislation that outlawed closed shops. But simple correlations between regional outcomes and a right-to-work indicator are not all that informative about causal effects, given that regions may differ in other ways, such as geography and climate. Holmes' answer to this problem was to identify sets of counties adjacent to borders, where right-to-work laws applied one side of the border and not the other. Using the presence of these laws as a proxy for more generally pro-business policies, he found large effects: manufacturing's share of employment increased by about a third on crossing from an anti-business state to a pro-business state. As Holmes (1998, p. 671) explains, the power of this approach is that:

> *…at state borders, the geographic determinants of the distribution of manufacturing—for example, climate, soil fertility, access to transportation, and the level of agglomeration benefits—are approximately the same on both sides of the border. What differs at the border is policy.*

In what follows, we call this approach a spatial discontinuity design. It has since been applied in other contexts, including to political institutions and financial reform. For now, we note that Holmes' paper not only demonstrates the power of this method, but also provides a careful account of its limitations. He notes that the effects of policy differences far from the border may be smaller than the effects close to the border. After all, a firm may be more influenced by policy differences between locations that are close to one another (and hence similar in terms of market potential) than between locations that are further distant. With this in mind, Holmes interprets his estimates as upper bounds on the effects of a statewide policy change. The same issue substantially complicates a welfare analysis. A policy difference that shifts a firm from one location to another could have minor effects on welfare (for example, if a firm chooses to locate one side of a border rather than another) or major effects (for example, if policy differences compounded the decline of America's northern cities).

A second issue is that borders are not randomly generated. This means that an identifying assumption—geographic characteristics are the same either side of the border—will not always hold. The example in Holmes (1998) is that some state boundaries coincide with discontinuities in nature represented by mountain ranges and coal veins. As a result, he drops some observations, but acknowledges that there may be other unknown instances in the data. Although this is a limitation, spatial discontinuity designs are likely to hold various other characteristics constant, to an extent that is otherwise hard to achieve. They represent one of the most informative methods for learning about regional prosperity.

4.6.6 Synthetic Controls

For some regional questions, an approach based on spatial discontinuity may be either infeasible or uninformative. This is especially likely when a researcher is interested in events or characteristics confined to a single region, or a small number of regions. As an example, consider a researcher interested in the effect of localized conflict on a single region's prosperity. There is not an obvious way to construct a counterfactual. Comparing outcomes with those of a neighboring region may not work, because there is no guarantee that the two regions will share similar characteristics. The alternative is a less formal case study, but that has problems of its own (see Temple, 1999).

Abadie and Gardeazabal (2003) considered this problem and introduced a method for constructing a synthetic control, which can be compared to the region of interest. In an application to Spanish regions, their specific aim was to quantify the economic effects of Basque terrorism on the Basque Country. To do this, they compared the evolution of Basque Country outcomes with a weighted average of other Spanish regions: the synthetic control. The weights were chosen so that the characteristics of the synthetic control resembled those of the Basque Country in the years before terrorism. The synthetic control can be seen as an approximation to the required counterfactual, the Basque Country without terrorism.

More formally, consider a case where there are J control regions available (in their case, the Spanish regions other than the Basque Country). The treated region has a set of K characteristics stored in a $(K \times 1)$ vector X_1. The J control regions have corresponding pre-treatment characteristics stored in a $K \times J$ matrix X_0. Drawing on ideas in the statistical literature on matching, the suggestion of Abadie and Gardeazabal is to choose a $(J \times 1)$ column vector of weights $W = (w_1, \ldots, w_J)'$ in order to minimize:

$$(X_1 - X_0 W)' V (X_1 - X_0 W),$$

subject to $w_i \geq 0$ and $\sum w_i = 1$, and where V is a diagonal matrix with non-negative components, which weight the different characteristics.

Hence, as well as choosing the set of K relevant characteristics, the researcher has to decide how to weight them. The diagonal elements of V could be based on subjective judgments about their relative importance. In their own application, Abadie and Gardeazabal use a more objective approach, and choose the elements of V so that GDP per capita in the synthetic control is close to that of the Basque country for the pre-treatment years. Once a researcher has chosen or estimated V, and obtained the weights W, outcomes can be compared between the region of interest and the synthetic control. For example, the GDP per capita of the control will just be a weighted average of the GDP per capita of the J regions, where the weights are the (possibly zero) individual elements of W.

The synthetic control method lends itself to graphical comparisons of outcomes, and robustness tests using the placebo approach familiar from the treatment effects literature. Applications to regional data are currently limited, but the method is especially likely to be useful when the number of regions is small, or the treatment of interest is confined to a small number of regions. It also provides a bridge between the regression-based methods favored by economists, and the more qualitative, case-study approaches favored in some other disciplines. An introductory overview of the method by Abadie et al. (2012) makes this point in relation to political science.

4.7. WHAT DETERMINES REGIONAL PROSPERITY?

We now turn to the empirical evidence on regional growth. Following the precedent of the cross-country literature, our use of the term "growth" is deliberately elastic. We use the term to encompass the study of influences on levels (or relative levels) as well as influences on steady-state growth rates. In fact, most of what we have to say has more bearing on the former, and so "prosperity" might be a better term.

There is a second ambiguity, to a far greater extent than in the cross-country literature. Regional growth sometimes refers to an increase in population rather than productivity, as factors of production gravitate toward particular areas. In fact, some authors have argued that measures such as population density will better capture underlying differences in productivity and quality of life (for example, Rappaport and Sachs, 2003). Many spatial models predict that equalization of real incomes will be achieved through adjustment in nominal wages, price levels, and local population sizes. Hence, the criterion for regional success, or the best interpretation of "growth," varies across studies and research questions.

4.7.1 Physical Geography

We start with physical geography, which can influence economic activity and population density through many channels; it is a more disparate topic than it might appear at first. Among the channels highlighted in the literature are access to the coast or transport networks, and physical transport costs more broadly; climate factors such as temperature or precipitation; and disease ecology.[28] Less obviously, as we also discuss below, physical geography can be a long-run influence on cultural and social norms, and local institutional development. And it may be especially important for developing countries, some of which are much larger and more heterogeneous internally than, say, the countries of Western Europe.

Geography can be thought of as influencing prices (partly through higher transport costs for remote regions), total factor productivity (in both agriculture and industry, see Dell et al. 2012), and incentives for factor accumulation. One of the most well-known findings is that economic activity is disproportionately coastal: Gallup et al. (1999) report that the areas of the US, Western Europe, and northeast Asia that are within 100 km of the coast contain just 3% of the world's inhabited land area, but 13% of its population, and at least 32% of global GDP.

Coastal locations, by lowering the costs of external trade, can be seen as favoring high productivity. Rappaport and Sachs (2003) argue that the coastal concentration in the US derives primarily from a productivity effect. The direct benefit of coastal location will be amplified by effects on economic geography, as firms and populations form agglomerations in coastal areas. This can also introduce path dependence; Bleakley and Lin (2012) study this issue using the proximity of many American cities to historical obstacles to water navigation, where continued transport relied on overland hauling. They find that these obstacles continue to be associated with relatively high population densities, even though their direct relevance to transport costs has long since disappeared.

Some of the evidence that physical geography matters is based on studying the location of individual industries (Ellison and Glaeser, 1999; Davis and Weinstein, 2008). Much of this evidence is for developed countries, but Felkner and Townsend (2011) use detailed data for Thailand to show that enterprise locations are associated with various geographic characteristics.

When the data are analyzed at higher levels of aggregation, the interpretation is more difficult. The cross-section study by Gennaioli et al. (2013a) finds that average temperature has limited explanatory power for output per capita within countries. Mitton (2013) considers a wider range of geographic and climate variables; he finds that many are statistically significant, but collectively their explanatory power remains relatively modest. Dell et al. (2009) and Nordhaus (2006) find some effects of temperature using variation at

[28] Relevant papers include Bloom and Sachs (1998), Dell et al. (2012), and Sachs and Malaney (2002), respectively.

smaller spatial scales; in Nordhaus (2006), there are opposing effects on output per capita and output per area. The latter is relevant because, given the nature of a spatial equilibrium, some climate effects are likely to be more readily observable in relative population density, rather than relative productivity. As we have emphasized throughout, in models with heterogeneous sectors and/or mobile workers, trying to infer fundamental influences on productivity from comparisons of average output per capita is not straightforward.[29]

Instead, it should be recognized that physical geography will operate partly through the spatial distribution of the population. Some aspects of physical geography may have a limited direct effect on production costs—for example, fewer cloudy days per year—but can still influence wages and incomes, through their impact on the location decisions of utility-maximizing mobile workers (Roback, 1982). Using US data, Rappaport and Sachs (2003) point out that proximity to Great Lake or ocean coasts helps to explain population density in levels and changes; as well as a productivity explanation, there may also be a quality-of-life effect. Over the 20th century, the US saw a large-scale movement of population toward areas with good weather. Many of the northern industrial cities have lost population over time, while cities in the Sun Belt have grown. Rappaport (2007) argues that, as US incomes have risen, an income effect on the demand for good weather has been an important driver of this adjustment.[30]

These ideas are supported by studies which consider factor incomes. There is a large literature which links regional variation in wages and rents to physical geography through amenities such as better weather. A general finding is that some of the regional variation in wages can be explained by differences in climate-related amenities (Roback, 1982; Beeson, 1991). And researchers primarily interested in the effects of economic geography sometimes find effects of measures of physical geography in regional wage regressions (for example, Amiti and Cameron, 2007).

Even this brief summary hints at the difficulties of studying physical geography in the context of a spatial equilibrium. It can influence productivity directly and via economic geography, and partly through the location decisions of workers based on amenities, while path dependence complicates this even further. Various researchers have sought to cut through these complexities by studying major shocks or perturbations. Much of this work points to the sustained importance of fixed regional characteristics, even in the face of other changes. For example, Hornbeck (2012b) studies agricultural land values in the Great Plains of the US from 1945 to 2002, and shows that long-run technological progress has not diminished the importance of local environmental advantages.

[29] In one of the first studies of these questions, Warner (2002) calls a version of this problem the "mobility bias."

[30] The argument is that, as consumption goes up, the marginal utility of consumption falls and hence individuals are more willing to forego income for the sake of better weather; they migrate to regions with better weather, forcing wages in those regions downwards and house prices upwards until a spatial equilibrium is restored. See also Desmet and Rossi-Hansberg (2013b).

It is clear that, historically, some climate shocks have led to substantial population movements. Hornbeck (2012a) looks at the economic effects of the Dust Bowl, the severe drought and subsequent wind erosion of topsoil in sections of the American Plains in the 1930s. The erosion of topsoil greatly reduced agricultural productivity in the affected areas, leading to falls in the price of land, out-migration, and diverted in-migration. As Hornbeck notes, adjustment was achieved mainly by population movements.

Other major shocks have also been studied, as in the work of Davis and Weinstein (2002). They show that the relative population densities of Japanese regions have been remarkably stable over the past 8000 years, and that even large-scale shocks, such as the Allied bombing of Japanese cities during World War II, had only temporary effects on the Japanese city size distribution. The findings indicate the long-run importance of fixed characteristics of locations, including their physical geography.

Finally, a more complex set of arguments traces the influences of physical geography on local institutions; cultural and social norms; the distribution of ethnic groups; and even political trajectories. Physical geography can sometimes manifest itself in profound differences of institutions and culture, with the semi-autonomous Federally Administered Tribal Areas (FATA) of north-western Pakistan as a well-known example. More generally, as observers such as Scott (2009) have noted, state-building sometimes founders in the mountains. Herbst (2000) argues that the large interiors of some African countries, with their low population densities and disconnected peripheries, have made it difficult for governments to maintain control over their territories, and have limited the development of effective states. Even more complex effects of geography are possible. China's ethnic geography and cultural differences, it is sometimes argued, partly reflect historical differences between areas suitable for arable farming (and hence permanent settlement) and the more nomadic cultures of the pastoral areas.[31] Moreover, geography can shape the response to historical events: looking at the specific issue of rugged terrain, Nunn and Puga (2012) argue that its direct costs have been offset, in Africa's case, by the protection it offered from the slave trade, with effects that have persisted to the present day.

Taken together, these points indicate a major challenge for empirical researchers: much remains to be done in understanding when and how physical geography influences regional prosperity. And aspects of this task seem increasingly urgent, given the scope for climate change to reshape productivity levels and specialization across the world, both across and within countries. Dell et al. (2012) find that increases in temperature adversely affect output in poor countries, and may also have consequences for political stability. It should also be emphasized that, even if regional prosperity seems only modestly affected by temperature differentials between regions, the effects of climate change on national comparative advantage could be substantial. This in itself would be enough to drive new

[31] See Kaplan (2012) for an overview of this argument.

patterns of regional growth and decline for many of the world's countries, leaving aside other effects such as desertification.

Recent work by Krusell and Smith (2009), Hassler and Krusell (2012), and Desmet and Rossi-Hansberg (2013a) seeks to quantify the differential effects of climate change across distinct locations, an approach pioneered by Nordhaus in his development of the multi-region RICE model. Desmet and Rossi-Hansberg (2013a), in particular, emphasize the importance of a spatial dimension to the analysis: as the climate changes, welfare losses arise because of frictions in the movement of people and goods across locations. One consequence is that, in the presence of migration restrictions between the global south and the global north, the estimated welfare losses are much larger for the global south. Their work is based on the effects of temperature changes; for at least some countries, the uncertainties for future regional development are compounded by the possibilities of water stress, coastal flooding, changes in the incidence of extreme weather, and new risks to health.

4.7.2 Market Access

Geographers have long pointed out that access to markets influences regional output levels. For example, Harris (1954) argued that the demand facing a given region depends on the distance-weighted GDP of all other regions. More recently, the empirical literature by economists has adopted measures of market access derived from structural models in the New Economic Geography tradition. We now provide a brief review of this literature; for more detailed surveys, see Combes (2011), Combes et al. (2008), Head and Mayer (2004), and Redding (2010). Assessed as a whole, the literature strongly supports the idea that market access or proximity influences regional prosperity.

In Sections 4.4 and 4.5 of the chapter, we reviewed models in which firms in more central locations will face higher demand for their products and thus, initially, higher profits. The usual assumption of free entry will equalize rates of return across locations but, as Head and Mayer (2006) note, the adjustment could take place via local employment or production, or through changes in wages—or, more generally, the remuneration of immobile production factors. The bulk of the literature has focused on adjustment through wages, and much of our discussion will look at this mechanism. But there is also a literature which considers adjustment via employment and production changes, often drawing on models with a freely tradable numéraire sector which makes wages invariant to demand. Key papers in this "home market effect" literature include Davis and Weinstein (1999, 2003), Head and Ries (2001), and Hanson and Xiang (2004). Head and Mayer (2006) show how these papers relate to the literature on adjustment via wages. Empirically, it is often difficult to separate these two adjustment mechanisms cleanly.

Turning to adjustment via wage changes, a first strand in this literature builds on Redding and Venables (2004). Their influential paper takes the spatial distribution of production and expenditure as given, and considers the wages that firms in each location can afford to pay. Firms in more remote locations incur higher trade costs when selling their products. This lowers the value added attributable to the factors of production; labor, as the relatively immobile factor, is affected most. Hence, the wage and income levels of a region are influenced by its position relative to potential markets—in other words, by economic geography.

The empirical specification of Redding and Venables is ultimately based on the model of Krugman and Venables (1995). Symmetric, monopolistically competitive firms from a given location i sell their tradable output in N different locations subject to trade costs. Demand is of the CES form and production takes place under increasing returns to scale. If labor is the only factor of production, free entry implies the following relation between (nominal) wages in location i and the demand and prices in all regions, which Redding and Venables call the "wage equation":

$$w_i^\sigma = A \sum_{j=1}^{N} T_{ij}^{1-\sigma} E_j P_j^{\sigma-1}, \tag{4.23}$$

where w_i denotes wages, E_j and P_j are respectively expenditure on traded goods, and the CES price index in location j, T_{ij} are trade costs between locations i and j, σ the elasticity of substitution between product varieties produced by different firms, and A is a constant.[32] Trade costs take the familiar iceberg form: for every unit shipped only $1/T_{ij}$ units arrive, where $T_{ij} = 1$ would correspond to free trade.

Equation (4.23) says that wages in region i depend on the sum of expenditure in all other regions, adjusted for price differences and discounted by bilateral trade costs. Redding and Venables call the summation term in (4.23) "market access." Other authors, including Head and Mayer (2006), prefer the term "real market potential," to highlight the price component P_j absent from more traditional measures such as the Harris (1954) market potential. Redding and Venables estimate (4.23) for a cross-section of 101 developed and developing countries for the year 1996. They find that GDP per capita (used as a proxy for wages) is correlated with their measures of market access, even after controlling for characteristics such as institutions and resource endowments.

The findings suggest that relative prosperity has a spatial dimension, but the assumption that labor is immobile across locations is less attractive for regional data than cross-country data. In contrast, Hanson (2005) allows for labor mobility. In order to obtain empirically relevant spatial production patterns, with activity at each location, he follows Helpman (1998) and introduces a nontraded good (housing) to create an additional dispersion force. While real wages are equalized across regions in this model, nominal wages are still

[32] Redding and Venables also allow for technological differences between firms in different locations, intermediate inputs, and other internationally mobile primary factors. This implies that wages will also depend on technology levels and the price of intermediate inputs in each location.

a function of market access, as well as housing stocks.[33] In more detail, we have:

$$w_i^{\sigma} = B \sum_{j=1}^{N} T_{ij}^{1-\sigma} E_j^{\frac{\sigma(\mu-1)+1}{\mu}} H_j^{\frac{(\sigma-1)(1-\mu)}{\mu}} w_j^{\frac{\sigma-1}{\mu}}, \tag{4.24}$$

where B is a constant, H_j is the housing stock of region j (assumed to be in fixed supply), and μ is the expenditure share of the traded goods sector. Hanson refers to the summation term in (4.24) as the "augmented market potential" of region i, again to distinguish it from simpler measures that do not correct for price variation. He estimates (4.24) on a sample of 3075 counties in the continental United States for the period 1970–1990, and finds a strong positive correlation between changes in augmented market potential and changes in nominal wages.

These two frameworks have been used to study the geographical variation in wages and output levels for a wide range of countries, regions, and time periods. Breinlich (2006) and Head and Mayer (2006) use the Redding-Venables approach to explain the variation in output per capita and wages across European Union regions, arguing that labor mobility is relatively low. Both papers find that the measure of real market potential in (4.23) performs no better than the simpler Harris market potential, in terms of explanatory power as measured by the R^2. Using modifications of the Hanson approach, Brakman et al. (2004a) and Mion (2004) provide evidence for the importance of proximity to sources of demand for German and Italian regions respectively.

Recent research has extended the ideas to low-income and middle-income countries, often based on the Redding-Venables approach. Bosker and Garretsen (2012) find a positive correlation between market access and GDP per worker for sub-Saharan African countries. Fally et al. (2010) find a correlation between wages and market access for Brazilian states, Amiti and Cameron (2007) for Indonesian districts, and Hering and Poncet (2009, 2010) for Chinese provinces and cities.

A common finding is that, although the market access variables are significant, the magnitude of the estimated effect is substantially lower than in Redding and Venables (2004). One explanation is that (with the exception of Bosker and Garretsen) these newer studies work at a disaggregated level, using either firm- or worker-level data. This enables them to control for additional covariates which are likely to be correlated with market access, including human capital. Moreover, these papers study wage differences in a regional context, where labor mobility will promote the equalization of wages. The finding of Hering and Poncet (2010) that variation in market access has a stronger impact on wages of highly skilled workers, whose mobility is more restricted in China, lends support to this explanation.

[33] See Hanson (2005, Section 2) for a full derivation. Note that local expenditure is still taken to be exogenous despite full labor mobility.

There are some ways in which the developed-country literature needs modification when applied to developing countries or regions. For sub-Saharan Africa, Bosker and Garretsen (2012) find that the correlation between GDP per worker and market access is relatively weak. Their preferred explanation is that manufacturing, the sector to which the wage equation applies most directly, is still underdeveloped in the African case. In their study of China, Hering and Poncet (2010) show that wages in private firms, and particularly in foreign firms, react strongly to variation in a city's market access, but wages in state-owned enterprises much less so. These findings suggest that structural and institutional conditions can influence estimated relationships between wages and market access.

Since the cross-section study of Redding and Venables (2004), a number of papers have shown that the correlation holds using variation in market access over time. As mentioned previously, Hanson (2005) correlates changes in nominal wages with changes in market potential. Head and Mayer (2010) apply the Redding and Venables approach to all countries in the world with available trade data over the period 1965–2003. Breinlich (2006) and Bosker and Garretsen (2012) estimate specifications which sometimes include region and country fixed effects, respectively. The general finding is that output or income remains correlated with market access, but the correlation is substantially reduced when using the within variation.

Can the correlation between income and market access be interpreted as causal? One issue is that market access might be correlated with other fundamental determinants of local income levels, such as institutions or endowments. This can work both ways, since some determinants may themselves be influenced by market access; Redding and Schott (2003) construct a model in which incentives to acquire human capital are lower in countries with weak market access. But the fact that market access effects are weaker in the within dimension does suggest that market access may be correlated with time-invariant determinants of income levels omitted from cross-section regressions.

A further problem is that, in essence, wage equation estimates are based on regressions of "own" income (w_i) on measures of income/expenditure levels in neighboring cities, regions, or countries (E_j). But as discussed in Section 4.6.2, this leads to a correlation between regression disturbances and the market access variable, and inconsistent estimates. This is most evident from Equations (4.19) and (4.20) once we realize that market access can be seen as a spatial lag of regional expenditure levels adjusted for price differences ($E_j P_j^{\sigma-1}$), where the $T_{ij}^{1-\sigma}$ are the elements of the spatial weight matrix.[34]

One approach has been to search for instrumental variables, but the exclusion restrictions are often questionable, and the scope for finding a time-varying instrument is limited. A more promising approach is to study quasi-natural experiments in which there

[34] More precisely, we have wages or output per capita on the left-hand side of the market access equation (4.23), and regional expenditure levels on the right-hand side. In practice, however, regional wages and expenditure levels are highly correlated, and estimating a market access equation is conceptually similar to estimating Equation (4.19).

is exogenous variation in market access. The pioneering work in this area is Hanson (1996,1997), who uses the changes in market access generated by Mexico's trade liberalization in 1985. Focusing on the apparel sector, Hanson (1996) shows that the pre-liberalization period was characterized by a strong regional wage gradient, with wages declining with distance from Mexico City. The 1985 trade liberalization led to a partial breakdown of this gradient, which Hanson attributes to a relocation of apparel assembly production to regions bordering the United States. The evidence for other manufacturing sectors is weaker, although the earlier introduction of special enterprise zones near the border (the *maquiladoras* programme) led to a compression of regional wage differences (Hanson, 1997).

Another event, which seems even more likely to isolate an exogenous and sizeable change in market access, is the division of Germany in 1949 and its reunification in 1990. This is studied by Redding and Sturm (2008). They base their analysis on the model by Helpman (1998) but look at its predictions for equilibrium population sizes rather than nominal wages. They show that, consistent with the model's predictions, West German cities close to the border with East Germany experienced a substantial decline (and after reunification, recovery) in population growth relative to other West German cities.

In an extension of the Redding-Sturm approach, Brülhart et al. (2012) use the end of the Cold War, and the fall of the communist regimes in Eastern Europe, to isolate a change in market access. They study the differential impact on Austrian municipalities bordering former communist economies, relative to interior municipalities. In contrast to Redding and Sturm, they have both wage and employment data at their disposal, and can analyze adjustment through both channels. They find that wages and employment growth were both influenced positively by better market access, but the estimated impact on employment growth was about three times as large as the impact on wage growth. This again suggests that, in settings where labor mobility is important, studies focused on wages could miss important forms of adjustment.

Overall, the literature indicates that prosperity is strongly associated with market access, at a range of levels of aggregation. This association is consistent with formal models from the New Economic Geography literature, in which the link is causal. One qualification is that, in some circumstances, the effects of market access on wage and employment patterns will be observationally equivalent to the effects of technological spillovers and labor pooling (Duranton and Puga, 2004; Redding, 2010). Some papers seek to address these alternative explanations using control variables, but their treatment is often less sophisticated than the treatment of market access, and draws less heavily on structural models. Another remaining challenge, and one of particular relevance for this chapter, is to integrate the analysis of adjustments in wages and those in employment. In seeking to understand regional prosperity, it would be useful to know how labor mobility and various institutional constraints or frictions shape the relative importance of these two forms of adjustment.

4.7.3 Openness

Since the New Economic Geography borrows heavily from trade theory, a natural research topic has been the relationship between external trade, internal economic geography, and regional disparities. Fujita et al. (1999) analyze this issue in detail, suggesting that trade could work to disperse manufacturing industry as a whole, but also lead to the spatial clustering of specific industries. Given the empirical importance of market access effects, it seems inevitable that spatial patterns of activity will be influenced by the nature of external trade, as in the work of Hanson (1996) reviewed above.

Redding (2012) uses a structural model (reviewed in Section 4.4 above) to examine the effects of a fall in trade costs between the US and Canada, leaving internal trade costs unchanged. Given its greater trade intensity with US states, Central Canada would gain more than Western Canada under population immobility. But if the population is mobile across regions, the improved market access of Central Canada causes it to gain population, while Western Canada would see a decline in population. The endogenous reallocation of population continues until all Canadian regions gain equally from the fall in trade costs, in the absence of costs to mobility.

Empirical work on external trade and regional disparities has often taken a reduced-form approach. In a study of Latin America, Serra et al. (2006) argue that regional disparities modestly increased, at least temporarily, in the wake of trade liberalization; the effect seems especially marked for Mexico. A further issue, especially for developing countries, is the influence of FDI on regional prosperity. Brakman et al. (2009) review the literature on the relationship between international business, FDI, and agglomeration. A small empirical literature studies the links between FDI and regional inequality directly, particularly for the Chinese case, where FDI has been heavily concentrated in the eastern provinces (Wei et al. 2009). Lessmann (2013) studies China, and a wider sample of 55 countries over 30 years, 1980–2009; his main result is that FDI inflows may increase regional inequality in developing countries, but there is no evidence of a similar effect in richer countries.

4.7.4 Transport and Infrastructure

In his book, *The Age of Capital*, Hobsbawm (1962) briefly recounts the story of teachers sent from Rome to Sicily in the 1860s with plans to standardize the school curriculum. Differences in regional idioms, and the extent of regional insularity, were so extreme that the Sicilians mistook the teachers for visitors from England. This story can stand in for others: differences in regional dialects and social norms testify to long spans of time in which regions were not closely integrated. What changed this insularity, in Italy as elsewhere, was in large part new technologies for transport and communication. It is a truism that lower transport costs have made regions, countries and the world smaller, and played a major role in reconfiguring the spatial distribution of economic activity. Williamson (2006) provides an account of the transport revolution of the 19th century, documenting substantial falls in transport costs, driven by canal-building, steamships, and railways. In 1817, it took 52 days to ship a load of freight from Cincinnati to New York by wagon and riverboat; by 1852 this had fallen to 6 days (Williamson, 2006, p. 8).

As we saw in Section 4.4, theoretical models differ in their predictions about the effects of lower internal transport costs, partly because lower costs make it easier for the consumers of a rural periphery to be served from cities. The effect of transport costs has been central to the New Economic Geography, and detailed treatments can be found in Fujita et al. (1999) and Combes et al. (2008), among others. The ambiguity of the models makes empirical work especially important, but it is not easy to quantify the causal effect of infrastructure. Investment in transport and communications will sometimes respond to regional changes in activity or population that originated with other forces. When policy-makers are forward-looking, national and sub-national governments may invest in regions with good growth prospects, or that are politically important. The research questions are difficult, but also highly relevant: investments in infrastructure for depressed areas have often been central to regional policies in Europe, China, and elsewhere.

One approach to identifying causal effects is to construct a structural model with a role for transport, such as a computable general equilibrium model. The huge advantage of this approach is that counterfactuals can be studied, by simulating the patterns of regional development under different assumptions about transport technology or infrastructure investment. Williamson (1974) is an early example of this approach; Herrendorf et al. (2012) a more recent one, both covering the effects of 19th-century transport changes in the USA. If this approach has a weakness, it lies in the ambiguity already noted: it is not clear how to choose between models, but conclusions about the effects of transport costs are sensitive to this choice.

The work of Donaldson (2010), briefly reviewed in Section 4.4, develops a Ricardian trade model with many locations and commodities, and trade costs. He uses this to study the introduction of the railway network of pre-partition India, seen as reducing trade costs between districts. His reduced-form regression estimate is that access to the railway increases a district's real income by 16%, and he finds that lower trade costs account for the entirety of this reduced-form effect. Donaldson and Hornbeck (2012) study the 19th-century expansion of the US railway network, finding effects that are more than double those in the well-known social saving approach of Fogel (1964).

Michaels (2008) studies the introduction of the US Interstate Highway System, which connected cities and border crossings, but also lowered trade costs for the rural counties crossed by new roads. He finds that these counties experienced significant increases in trade-related activity, but without major changes in specialization in the directions predicted by trade theory.[35] Banerjee et al. (2012) study access to transport infrastructure

[35] More precisely, motivated by trade theory, he finds small increases in the wage bill of skilled workers relative to unskilled workers in skill-abundant counties, and small reductions where skills were scarce; but there is no evidence for changes in the industrial composition toward industries intensive in the abundant factor.

in China, exploiting the historical importance of connections between the major cities of the 19th century and the Treaty Ports. They find that regions closer to historical transportation networks have significantly higher levels of GDP per capita and higher average firm profits, but there is no evidence that the advantaged regions grew more quickly over the period studied (1986–2006).

For some developing countries, regional prosperity may also be influenced by energy infrastructure, and particularly the extent of electrification. It has been estimated that around a quarter of the world's population lack access to electricity. Lipscomb et al. (2013) study the long-run effects of electrification in Brazil: using spatial variation in the scope for hydropower plants, they can isolate exogenous variation in the extension of the network. Their results suggest that electrification brings significant gains in educational attainment, employment rates, and income per capita.

The effects of geographic characteristics on electrification programs have also been exploited by Dinkelman (2011) and Rud (2012). Dinkelman studies the effects of a household electrification program in South Africa, using land gradient to isolate exogenous variation in access. The results indicate significant effects on female employment, potentially due to time released from home production and increased small-scale labor demand. In a study of Indian states, Rud (2012) uses the uneven availability of groundwater for electric-pump-based irrigation schemes to instrument for the expansion of the electricity network. His panel data estimates indicate that an increase in rural connections of one standard deviation would increase a state's manufacturing output by almost 15%.

Massive investments in infrastructure often appeal to policy-makers seeking to accelerate development by concrete, visible means; Lenin once defined communism as Soviet power plus electrification. This political appeal might suggest a risk of overinvestment, and we have already seen that the effects of transport investments can be ambiguous. Even for electrification, the analysis of welfare effects becomes more complicated in a spatial equilibrium. New infrastructure can induce population movements that increase the demands on locally provided public goods. Dinkelman and Schulhofer-Wohl (2012) study the issue, again for household electrification in South Africa, and find that congestion effects can halve the estimated local welfare gains.

4.7.5 Institutions and Local Political Economy

The study of institutions has been central to recent work on comparative development. National institutions will be among the forces that shape patterns of regional specialization and relative incomes, partly because they will influence comparative advantage at the national level. But there may also be important local variation in institutions within countries, as Acemoglu and Dell (2010), Naritomi et al. (2012), and Tabellini (2010) all emphasize. Its consequences have now been investigated for countries in Africa, Latin America, and South Asia. It also has implications for the study of national development: countries could share the same rating for institutional quality—effectively a weighted average across regions—but differ in their internal institutional variation, with consequences for agglomeration and overall activity.

The idea that institutions vary within countries needs a little justification. In federal countries in particular, such as Brazil, India, and Mexico, some areas of law may be determined locally, and de jure institutions will then vary across regions. But even where de jure institutions are similar, there may be substantial inter-regional differences in how these institutions operate in practice, partly given the importance of informal institutions. Tabellini (2010) emphasizes that given institutions can function differently across locations. He suggests that the judicial system works differently in southern and northern Italy, even though the formal frameworks are similar. A further complication is that which institutions matter will depend on a region's specialization; the institutions most relevant to rural agriculture may differ from the institutions most relevant to urban firms, for example. Using surveys of public employees in Bolivia, Brazil, and Chile, Gingerich (2013) shows that perceived effectiveness varies across different government agencies within these countries.

Further, the nature of the political economy will vary across regions. This could include the extent to which local elections are free and fair, the extent of control exerted by local elites, and the effectiveness of the rule of law and the judiciary. There is now a large literature examining variation in political economy at the sub-national level, with Besley and Burgess (2002) and Baland and Robinson (2008) as just two examples. Many instances of sub-national authoritarianism have been documented for democracies in developing countries and transition economies; for example, the dominance of sub-national government by single parties was a feature of the US South until the later part of the 20th century (see Gibson, 2005). Through these mechanisms, there could be significant variation across regions in the quality of government, the provision of local public goods, and in the rule of law and contract enforcement.

Local institutions will be a determinant of the comparative advantage of regions, in the same way that national institutions appear to shape the comparative advantage of countries (for example, Nunn, 2007). One way in which the regional context differs is that individuals have considerable scope to relocate. Hence, when local public goods and amenities are better in some areas than others, and valued by individuals, migration across regions will take place until these advantages are offset by congestion—more intensive use of amenities—or higher living costs, such as housing costs. Similarly, firms that use local public goods intensively will tend to relocate to regions that provide these goods effectively. Hence, variation in local institutions will influence regional prosperity and population density.

In the past, the study of these effects has been hampered by the lack of data on institutional variation at the local level. The leading approach has been to study natural experiments. The well-known study by Banerjee and Iyer (2005) looks at a variety of outcomes across Indian districts, notably agricultural investments and productivity, and relates them to historical variation in land rights under British rule in the 19th century. Banerjee and Iyer find that outcomes have diverged: those districts where land rights were given to landlords rather than cultivators have significantly worse outcomes in the post-independence period. Some of the divergence takes place relatively late, in 1965–1980, which they attribute to the varying political trajectories of (historically) landlord and non-landlord districts. Their leading explanation is that, in districts where land rights were given to landlords, this led to a class-based and antagonistic politics, with consequences for policy priorities and public investment that have persisted for many decades. The results indicate that regional variation in public investment and development expenditure can make a material difference to outcomes at the sub-national level. It also seems clear that variation in local institutions and political trajectories have long-lived effects on regional outcomes. Further work on India by Iyer (2010), exploiting exogenous variation in direct British colonial rule versus indirect rule, reaches similar conclusions: areas that were under direct rule continued to have higher levels of poverty and infant mortality well into the post-colonial period.

Along similar lines, Naritomi et al. (2012) study local institutions in Brazil, finding that institutional quality and the distribution of land have been influenced by the distinct colonial histories of different regions. Acemoglu et al. (2012) study Colombia, identifying persistent effects of slavery on various outcomes, exploiting spatial variation in slavery associated with the presence of gold mines during the 17th and 18th centuries. Dell (2010) finds similarly long-lived effects of a forced labor scheme in Peru, the "mita." The identification strategy is a spatial discontinuity design, based on comparing outcomes either side of a section of the geographic boundary of the affected area. Although the scheme was abolished in 1812, Dell establishes that its effects can still be seen today in substantially lower consumption levels, a greater incidence of child stunting, and greater prevalence of subsistence farming in the affected districts. She argues that these effects arose because the mita districts followed a different political trajectory, based on communal land tenure, compared to non-mita districts. The latter provided a more stable land tenure system that encouraged public goods provision, including education. The main results show how local institutions can have long-lived effects on spatial disparities. Moreover, these disparities seem to correspond to differences in life chances and opportunities that have not been eliminated by the possibility of migration between regions, even over many decades.

Further evidence of the long reach of history comes from the innovative study of Michalopoulos and Papaioannou (2013), briefly mentioned earlier. They study the relationship between contemporary sub-national development in Africa and measures of pre-colonial political centralization, where the latter reflect the extent of levels of political jurisdictions above the local (village) level. The initial results are based on a sample of roughly 500–700 geographic units. The measure of contemporary development is derived from satellite data on light density at night; this also allows a higher-resolution analysis of around 66,000 geographic units. Further, they also compare light density across contiguous ethnic homelands of groups that differ in their traditions of political centralization.

The results consistently suggest that differences in light density, and hence in the density of contemporary economic activity, are related to long-standing differences in institutional traditions: areas with traditions of political hierarchy have higher development levels. As they note, the data on light density open up many further research possibilities.

In the literature to date, the leading sceptics are Gennaioli et al. (2013a) and Mitton (2013). To explore the question, Gennaioli et al. run a simple regression of regional GDP per capita on country dummies and a proxy for economic institutions constructed from sub-national data extracted from the Enterprise Survey and *Doing Business* reports; there are 496 regional-level observations, across 79 countries. They find that, although there is significant regional variation in their institutions measure, it explains very little of the regional variation in GDP per capita. Mitton (2013) obtains similar results. It is not clear how to reconcile their findings with the studies based on natural experiments, although the latter give more emphasis to political institutions (as opposed to economic institutions) and exploit exogenous variation. One possible story is that perceptions of institutions depend on a region's specialization. It might also be that measured variation in economic institutions is endogenous and linked to the scope for corruption and the politicization of economic activity, which could be greater in richer or more industrialized regions. Gennaioli et al. (2013a, p. 128) note that, on average, economic institutions are perceived as weaker in a country's richest region than in its poorest. Nevertheless, these explanations are speculative, and a good instrument for sub-national variation in economic institutions seems unlikely to emerge.

4.7.6 Culture and Social Norms

The sub-national variation in cultural and social norms, and social capital, has also been studied. The concept of social capital often lacks well-defined boundaries, but partly because of this, it is a useful umbrella term for social norms such as trust and civic engagement. A contemporary economist might also consider the density of social networks, and the quality of the links within them. Much of the recent interest in social capital can be attributed to the work of Putnam et al. (1993), who contrasted the levels of trust and civic participation between regions of Italy, and argued that these differences in social norms had far-reaching consequences, partly acting through political outcomes.

Using regional data for Europe, Tabellini (2010) analyzes the relationship between regional incomes (and growth rates) and measures of cultural norms, such as trust, respect for others, and respect for individual independence and autonomy. His study includes country fixed effects and instruments the cultural variables using long-run historical data, 19th-century literacy rates and early (1600–1850) political institutions, both measured at the regional level.[36] His results suggest that, although the regions within each country have

[36] It should be noted, however, that the political institutions measure has only limited measured variation within some countries.

long shared the same formal institutions, historical data help to explain contemporary outcomes, with the effects mediated by cultural and social norms. The tenor of these findings is consistent with Banerjee and Iyer (2005) and Dell (2010), suggesting that regional outcomes can often be traced back many decades.

Within Russia, Acemoglu et al. (2011) find long-run effects of the persecution, displacement and mass murder of Jews by the Nazis during World War II; the cities where this was most intense have grown relatively slowly, and show greater support for communist politicians. The administrative districts (oblasts) most affected have lower average wages and income per capita. Acemoglu et al. attribute these effects to the changes in social structure brought by the Holocaust.

Other work on social norms uses variation at borders. For example, Becker and Boeckh (2011) study communities in Eastern Europe either side of the border of the former Habsburg Empire. Using survey data for 2006, they find that historical affiliation of an area with the Empire is associated with higher trust and less corruption in courts and the police, even though the Empire was broken up almost a century ago, in 1918. Along similar lines, Grosfeld and Zhuravskaya (2012) use the historical partition of Poland among three Empires—Russia, Austria-Hungary, and Prussia—and find effects either side of former borders on religious beliefs, voting patterns, and political beliefs, including support for democracy. Again, it is natural to think that some of these differences would have implications for the spatial pattern of activity, and interesting that spatial differences in culture and social norms can persist for decades.

One natural question is whether religious differences influence regional outcomes. Using data on the counties of Prussia in the late 19th century, Becker and Woessmann (2009) find that Protestant counties are relatively prosperous, and attribute this to higher levels of literacy prompted by Luther's emphasis on schooling. In contrast, Cantoni (2010) exploits religious variation across the German lands of the Holy Roman Empire; using data on populations for 272 cities over 1300–1900, he finds that the paths taken by Catholic and Protestant cities and regions are virtually indistinguishable.

The literature we have reviewed emphasizes differences across regions in cultural and social norms. A less obvious argument is that differences across countries could influence the agglomeration process. For example, patterns of labor mobility may differ between societies that are relatively atomistic and individualistic, and those where close family ties are especially valued. Duranton et al. (2009) establish some interesting associations between historical family types, classified for medieval Europe, and variation in outcomes within countries. Investigating such hypotheses across countries is not straightforward, however: a cross-section analysis is limited by the small number of countries in the world, and a panel data analysis by the lack of time-series variation in cultural norms.

Nevertheless, the importance of social ties is worth stressing. Economists often emphasize the benefits of mobility, and fluid economic arrangements are seen as important for efficiency. But a highly mobile labor force is also one in which networks of family, friends, and neighborhood connections are repeatedly disrupted. Most of the theoretical models present a society that is atomized by construction, and there are no frictions that arise from ties to family and friends. It is possible that such a world could exist, but few of us would want to live there.

Recent empirical work by Belot and Ermisch (2009) and Dahl and Sorenson (2010) indicates the potential importance of social ties to location decisions. This raises the possibility of nonpecuniary externalities to mobility that are not always acknowledged. The social capital perspective complicates the picture still further: Sennett (1998) raised concerns that a modern, highly educated workforce may be mobile, but also rootless and rarely socially engaged. More broadly, economic adjustments across regions and cities are likely to have cultural and social consequences, changing the character of areas in unpredictable ways.[37] These considerations are hard to define, and might seem a topic for sociologists rather than economists, but for the inconvenient fact that welfare effects sometimes involve non-economic mechanisms.

4.7.7 Entrepreneurship, Skills and Ideas

The role of entrepreneurship in prosperity is one of the most vexed questions in regional economics. Glaeser et al. (2010) open their discussion with the following questions: Can the economic history of Detroit be told without Henry Ford and Alfred Sloan? Would Ford have achieved the same success if he had worked in Houston? Would Silicon Valley have experienced its remarkable growth without Frederick Terman and William Shockley? These questions hint at some degree of indeterminacy in the evolution of regional specialization and prosperity. They seem to open the way to a Great Man approach, in which, to misquote Thomas Carlyle, the history of regions is nothing but the biography of great men and women.

As in the more general study of history, this idea is unsettling. Taken to its extreme, it radically undermines attempts to generalize about regional growth. But equally clearly, there are limits on the extent to which individuals (and individual companies) can be decisive; Silicon Valley was more likely to take shape in California than Alaska. These considerations suggest that the idea of entrepreneurship should be invoked by historians and economists in rather different ways. The historian of a region might want to draw heavily on what Klepper (2011) terms "nano-economics," the study of specific companies and entrepreneurs, their spin-off companies, and other legacies. This endeavor would have lessons for the study of regional prosperity, but is not coterminous with it. Economists will typically want to think about models in which entrepreneurship is an outcome or mechanism within a much larger process. Put differently, explaining what happened in

[37] As an example, Solnit (2013) provides a brief account of contemporary San Francisco which emphasizes the losses that can accrue as employment patterns and living costs change.

retrospect (to Detroit, or Silicon Valley) is not the same exercise as understanding how entrepreneurship shapes regional prosperity in general, even if there is some overlap.

This hints at some difficulty in framing the relevant research questions. Glaeser et al. (2010) argue that entrepreneurs will play a crucial role in the extent to which cities and regions are economically dynamic, and they survey some of the literature in this area. They emphasize that many of the forces which could drive agglomeration—spatial differences in input availability, access to ideas, and local culture or institutions—will also influence the extent of entrepreneurship. We note a corollary: seeking to quantify "the" effect of entrepreneurship on regional prosperity makes little more sense than seeking to quantify the effect of agglomeration. Serious empirical work on entrepreneurship has to contend with its endogeneity to a range of economic and social forces, and this helps to explain why there is not more work on the topic.

Progress might depend on ingenious use of natural experiments, with Glaeser et al. (2012) as a leading example. They argue that, in the US, areas close to mines were more likely to specialize in industries like steel, with significant scale economies and dominated by large firms; as a result, the conditions for entrepreneurship were less likely to arise. They find that proximity to historical mining deposits (in 1900) is indeed associated with larger firms and fewer start-ups decades later, and use this proximity as an instrument for entrepreneurship. Across cities, entrepreneurship is strongly associated with faster employment growth even in IV estimates.

Entrepreneurship plays a central role in the structural model of Gennaioli et al. (2013a), discussed in Section 4.4.3 above. Individuals can choose between employment and entrepreneurship; more able individuals self-select into entrepreneurship and, as in Lucas (1978), especially able entrepreneurs run larger and more productive firms. The most important empirical consequence is that human capital formation is placed center-stage, as a source of highly able individuals; and traditional Mincerian wage regressions, or development accounting exercises, risk understating the effect of schooling on regional prosperity, because some of the returns to schooling are reflected in capital income rather than wages. Felkner and Townsend (2011) also construct an occupational choice model with a role for entrepreneurship, but emphasizing the role of local access to finance; they simulate the model on detailed data for Thailand, and compare the paths taken by spatial enterprise concentration with those seen in the data.

The model in Gennaioli et al. (2013a) is essentially static, and designed to explain outcomes in a cross-section of regions. But one reason for being interested in entrepreneurship is that it may help to explain why some regions successfully reinvent themselves, while others lose dynamism. It is also related to ideas about information transmission and the extent to which individuals and firms are linked through various networks; this suggests the benefits of integrating the analysis of regions with ideas from urban economics. Glaeser and Gottlieb (2009) argue that, to be successful, modern cities increasingly depend on the links between urban density and the transmission of ideas. For some countries, the spatial concentration of the highly skilled is likely to be increasing; Moretti (2012) argues this for the US.[38]

Particular instances of entrepreneurship, firm entry, and the concentration of skilled workers in particular locations, are often thought to be associated with universities, as in Stanford's influence on Silicon Valley, or the cluster of technology companies around Cambridge in the UK. Traditionally, the study of some of the effects of universities has drawn on the local multipliers and impact assessments developed in the regional science literature.[39] But quantifying the wider benefits of universities for the local transmission of ideas, innovation, firm entry, or the "creative classes" of Florida (2002) is even harder.

Universities are not randomly assigned across locations, and natural experiments are hard to find. Moretti (2004) studies the social returns to education in the US, partly by using the land-grant colleges of 1862 and 1890 to instrument for differences in the share of college graduates across cities. He finds that a higher college share not only increases the wages of less-educated workers, but also those of the well-educated, consistent with a role for human capital externalities. An alternative approach is to use more detailed data, perhaps at the establishment level, to study particular mechanisms; for example, Abramovsky et al. (2007) find that business R&D in the UK is sometimes located close to highly-ranked university research departments in related disciplines.

4.7.8 Local Financial Development

Does local financial development matter? In a well-known paper, Jayaratne and Strahan (1996) used data on US states to study the effects of bank deregulation on economic growth. Most US states began the 1970s with restrictions on the expansion of bank branches within and across state borders; over the following 25 years, the majority of these states eliminated or loosened the controls. Using an empirical model with state fixed effects, Jayaratne and Strahan find that branching deregulation significantly increased the growth rate of state personal income per capita and gross state product per capita. The evidence that this was achieved by a greater volume of commercial lending is not strong. Instead, the lifting of branching restrictions seems to have resulted in a lower share of non-performing loans in the state total, and a lower share of loans being written off each year. This evidence on the quality of lending is not conclusive, since the loan portfolios of banks may have changed in their size composition and in the riskiness of borrowers. Nevertheless, the paper suggests that local financial intermediation can influence regional prosperity.

[38] See also Ganong and Shoag (2013) for further discussion and references.

[39] Armstrong and Taylor (2000) include an introduction to these approaches, acknowledging some important objections. Using, instead, an econometric approach to local multipliers, Moretti (2010) finds that one additional skilled job in the traded sector will generate 2.5 jobs in providing goods and services.

More recently, the same policy reform has been revisited by Huang (2008) using a spatial discontinuity design. He compares performance across pairs of contiguous counties either side of a state border, where one county is affected by deregulation earlier than the other. This approach allows for heterogeneity in treatment effects over time and across states. The evidence that deregulation had economic benefits seems noticeably weaker than in the Jayaratne and Strahan study, although it is not clear whether the alternative approach to identification has led to more reliable estimates, or just to greater imprecision. The results do not appear to be driven by spillovers of deregulation across borders, since Huang also compares outcomes using hinterland counties within the still-regulated states, further from the border. He emphasizes that the instances of significant growth accelerations in his study all occur relatively late in the reform process, after 1985, and interprets this in terms of "learning by observing," so that states which liberalized later tended to have better outcomes on average.

For those developed countries without restrictions on inter-regional lending or spatial variation in regulation, a sceptic might argue that local financial development cannot matter. Comparing regions of Italy, Guiso et al. (2004) find evidence that it does. For example, the ratio of new firms to population is 25% higher in the most financially developed Italian region, compared to the least. Their study exploits a 1936 banking law, which had persistent effects on the number of bank branches, and can be used as an instrument for the exogenous supply of credit. Natural experiments have also been found in other countries. In a study based on Russian data, Berkowitz et al. (2012) use regional variation in banking that arose at the end of the Soviet era, and its establishment of specialized banks (*spetsbanks*). They find that the presence of the spetsbanks increased within-region lending to firms and individuals, but had no discernible effect on income per capita. Regions with spetsbanks are associated with increased employment rates, however.

Chen et al. (2010) study venture capital in the US, noting that venture capital firms, and venture-capital-financed companies, are heavily concentrated in just three metropolitan areas (Boston, New York, and San Francisco). They associate this with localized knowledge spillovers in sectors especially likely to draw on venture capital; with localized knowledge spillovers across venture capital firms; and with entrepreneurs that seek finance from previously successful venture capital firms. These features could lead to a virtuous circle as entrepreneurs locate businesses close to funding sources, and other venture capital firms enter at the same location; conversely, other regions may experience a vicious circle. Chen et al. suggest that policies which increase the number of venture-backed investments in a region will increase the chances of venture capital firms establishing offices in that region.

For developing countries, there are complicating factors, not least the close connections between the banking sector and the state that are found in some countries. China is an important example: Démurger et al. (2002) argue informally that the monopoly state banking system has contributed to regional inequality, by limiting access to external finance in the interior provinces and by assigning priority for lending to the state-owned enterprises in the coastal and north-eastern regions. For India, Burgess and Pande (2005) investigate the effects of a large state-led expansion in bank branches in rural areas; they find that it increased deposit mobilization and lending, and lowered rural poverty. Fafchamps and Schündeln (2013) study local financial development in Morocco, at a lower level of aggregation, corresponding roughly to a city or county: they find that access to a bank increases firm entry, raises firm growth, and lowers the likelihood of firm exit.

4.7.9 Other Policies and Regulations

We have already discussed some well-known papers on specific policies and regulations. One approach to local policy variation uses spatial discontinuity designs, as in the paper by Holmes (1998) reviewed in Section 4.6.5 above. A more recent example is the work of Duranton et al. (2011) on local taxation, based on pairing establishments across borders in the UK. But the literature on local and regional policies in developed countries is sufficiently extensive to require a dedicated survey of its own, which space does not permit.

Instead, we briefly review some evidence for developing countries, much of it based on the states of India. The case of India is interesting because policies and regulations have varied across states and over time. Aghion et al. (2008) study the effects of dismantling the License Raj, a 1951 system of controls that regulated entry and production in the formal manufacturing sector. The elimination of these barriers to investment and entry affected states differently: those industries in states with pro-employer labor market institutions grew faster than those in states with pro-worker institutions. Since pro-worker institutions seem to be directly associated with weaker industrial performance, the overall effect of de-licensing was to increase the disadvantages of states with pro-worker labor market institutions. Earlier, the panel data study of Besley and Burgess (2004) had already found that pro-worker labor market regulation was associated with lower output, employment, investment, and productivity in the formal manufacturing sector; higher output in the informal sector; and higher rates of urban poverty.

Besley and Burgess remark that, in this case, specific attempts to redress the balance of power between capital and labor seem to have worked against the interests of the poor. Their earlier panel data study, Besley and Burgess (2000), examined an alternative redistributive policy, land reform. They find that reforms which changed the terms of land contracts lowered poverty and raised agricultural wages, although this may have been accompanied by lower average income. Implementing land reform had a poverty-reducing effect equivalent to growth in income per capita of around 10%. Since the estimated effects vary with the exact type of land reform, a further lesson of their study is that the specific details of a policy intervention can matter a great deal. A remaining question raised by these papers, not straightforward to answer, is the effect of policy variation on regional disparities when states are linked in a spatial equilibrium. Although labor mobility across Indian states is likely to be low, entrepreneurs and firms must still decide where to locate, and agglomeration and growth can be determined jointly even when labor is immobile.

Policies can also influence regional TFP through their effect on factor misallocation. For example, Brandt et al. (2013) study distortions within China, and find that most of their estimated within-province distortions are due to the misallocation of capital between the state and the non-state sectors; this misallocation lowers province-level TFP. For China, there has also been some work on policy-induced barriers to trade between provinces; see Young (2000) and Holz (2009) for alternative views on their importance.

4.7.10 Conflict

For some countries, localized conflict can influence the relative prosperity of regions, both in the short run and in ways that unfold over time. India's class conflicts take their most extreme form in the Naxalite or Maoist peasant uprisings, which affect the Red Corridor within eastern states.[40] The affected states are among the poorest in the country, but since economic and political outcomes are jointly determined, identifying the causal effect of conflict is difficult. Comparing Indonesian provinces, Hill et al. (2008) suggest that conflict has been a factor in the slow growth of Maluku and, to a lesser extent, resource-rich Aceh; relevant to the latter, Morelli and Rohner (2010) examine the relationship between the spatial distribution of natural resources and the risk of conflict, including the rise of secessionist movements.

One approach to recovering a causal effect is that of Abadie and Gardeazabal (2003). They use their synthetic control method, reviewed in Section 4.6.6 above, to study the effect of Basque terrorism. They find that it reduced GDP per capita in the Basque Country, relative to a synthetic control region without terrorism, by 10 percentage points.[41]

Another branch of the literature studies the effect of wartime destruction on the spatial distribution of population, or the relative outcomes of affected regions. The aim is not to investigate the overall humanitarian or economic costs of conflict or war, but to see whether past (localized) destruction influences later regional outcomes. Two well-known studies consider the effects of World War II bombing, by Davis and Weinstein (2002) and Brakman et al. (2004b); the former was briefly discussed in Section 4.7.1 above, and both are reviewed in detail in Brakman et al. (2009). Their review concludes that some shocks had permanent effects, consistent with models of agglomeration in which there are multiple equilibria and path dependence.

Miguel and Roland (2011) study the long-term regional effects of the US bombing of Vietnam, noting that it was heavily concentrated in a subset of their 584 sample districts, and hence with scope for differential effects across regions. They find that districts heavily bombed between 1965 and 1975 had moderately lower consumption (compared

[40] Banerjee and Iyer (2005) note that the regions most associated with this conflict are areas where landlord-based systems were implemented under British rule.

[41] As well as the synthetic control method, they also used an event study of the stock prices of firms significantly exposed to the Basque Country, to show that these stock prices outperformed when the 1998-1999 ceasefire became credible, and underperformed at the end of the ceasefire.

to other districts) in 1992–1993, but this effect had disappeared by 2002; nor do they find significant long-run effects on the relative poverty rates, electricity infrastructure, literacy, or population density of the affected areas. A complicating factor is that the Vietnamese government undertook major reconstruction efforts; but otherwise, the findings indicate that long-run patterns of spatial activity are largely independent even of damaging bombing campaigns. Miguel and Roland interpret this as evidence against simple models of regional poverty traps.

4.8. REGIONAL DECLINE

One of the simplest points about regional economics is also one of the most fundamental. The invisible hand is more active at some times and places than others, and once distinct points in space are introduced into economic theory, the conventional arguments that markets can be Pareto-efficient no longer apply. Markets, left to themselves, can establish patterns of regional growth and decline that involve many economic and social costs. One of the best reasons to study regional growth might be to learn how to forestall or reverse regional decline.

In this section, we discuss some of the processes involved in decline. Its analysis is partly the obverse of regional growth; for example, the results of Holmes (1998) tell us not only about the growth of US states with pro-business policies, but also the relative decline of states without them. Similarly, the evolution of location-specific advantages, such as market access, can explain decline as well as growth. Yoon (2013) argues that reductions in local advantages help to explain the decline of the US Rust Belt, compounded by a reversal of agglomeration and a decline in the quality of local public goods. This perhaps hints that decline raises specific issues of its own, which have been under-researched. Our treatment will be relatively discursive and speculative, emphasizing areas for future research rather than drawing heavily on existing work.

What do we mean by decline, and does it matter? Regions could be declining in terms of absolute or (more often) relative living standards and welfare indicators, but also in terms of absolute or relative population, since one response to economic decline will be out-migration and diverted in-migration. This second kind of decline is often a symptom of the first, but has interest in itself, as a distinct process. In the US, a country usually judged to have high labor mobility, it is perhaps not surprising that the primary response to the Dust Bowl was out-migration (Hornbeck, 2012a,b). Similarly, the Rust Belt has seen its share of the US population decline.

With few exceptions, economists generally take a benign view of factor mobility, and see it as a powerful equilibrating force. It is true that out-migration will sometimes benefit both migrants and those who remain behind, but this is not inevitable. In other ways it has the potential to compound the problems of a declining region, as Myrdal (1957) discussed. One complicating factor is selective migration; those who leave a declining region will often be the young and well-educated. Even in the absence of conventional human capital externalities, this form of out-migration could be self-reinforcing, and have social and political consequences for the declining region.[42] This is not to deny, as Myrdal seems to have done, that the logic of a spatial equilibrium will reassert itself. But the process of reaching it may involve significant costs, especially where the decline is absolute rather than relative.

The New Economic Geography literature has investigated decline in terms of the combined effects of changes in inter-regional trade costs and assumptions on labor mobility. In Puga (1999), when trade costs are high, industrial activity is dispersed. If trade costs fall, this promotes the agglomeration of activities with increasing returns. This is compounded by migration, implying the relative (and perhaps absolute) decline of some regions. But if workers do not move across regions, then as trade costs fall further, firms become increasingly sensitive to cost differentials across regions, and industry will spread out once more. Nocco (2005) considers a variant of this model with a role for knowledge spillovers across regions. A natural question is whether agglomeration is optimal; to consider this, Ottaviano and Thisse (2002) study a two-region economy with skilled workers that are mobile, and unskilled workers that are not. Market forces lead to the optimal outcome when trade costs are high or low, but for intermediate levels of trade costs, agglomeration takes place when dispersion is socially desirable.

When regional decline is discussed, an idea often heard is that policy-makers should seek to protect people rather than places. Some economists seem to take the view that, if out-migration is taking place, so be it. But this view risks leaving too much out. Some consequences of a region emptying out are inefficient, and involve multiple externalities. Infrastructure and social overhead capital will be written off or less well utilized, and the local tax base eroded. Movements of population to other areas will require new investment and increase demands on local public goods. Declining regions are likely to become low-trust, high-crime regions. Most of these outcomes will not be internalized by migrants, and it is often hard to see anything creative in the destruction of social capital.

As a practical matter, there is a large literature on regional policy, understood as a response to decline, whether in terms of relative economic position, or sustained out-migration. But one constraint on this literature is that some of the mechanisms underlying decline, such as crime, social unrest, and local political consequences, are complex. There is also a risk that regional policy could be too reactive. Once decline is under way, disadvantages can accumulate, and may be hard to reverse. Anticipation of regional decline

[42] In some ways, points such as these—the limitations and constraints on migration as an equilibrating force—have been better understood in the literature on developing countries. Lipton (1980, p.15) writes that it is "perfectly consistent to claim, as I do, that the migrant on average gains from migration, but the village he leaves behind loses." See also Kanbur and Rapoport (2005).

could even be self-fulfilling, which again suggests the need to consider regional problems in dynamic terms. On the basis that prevention is better than cure, one issue for policy will be a given region's extent of diversification, and hence its robustness to shocks. But, at the risk of laboring the obvious, there is no good reason to expect that decentralized markets will lead to the optimal degree of diversification, not least given the many externalities involved. This in itself could justify some degree of intervention.

To emphasize the lack of diversification of, say, 1960s Detroit, could seem a little too easy, a form of retrospective wisdom. After all, not many are currently calling for Silicon Valley to diversify. But as Glaeser (2011) emphasizes, Detroit's problem was that its fortunes were closely tied not just to a small number of sectors but to a small number of firms, the "big three" of Chrysler, Ford, and General Motors; to an uncomfortable extent, Detroit was a three-company town. The quantitative exercise of Alder et al. (2013) attributes much of the Rust Belt's wider decline to a lack of competition and powerful unions.[43] A related perspective could draw on Gabaix (2011), who argues that idiosyncratic shocks to large firms can account for aggregate business cycle fluctuations to a significant degree. This same idea of the "granularity" of economic activity could also be applied to regional growth and regional decline, and may be especially important for relatively small countries.

Another, closely related, lesson of Detroit might be the potential for path dependence, or regional lock in, a theme of some recent work by geographers. Specialization is an endogenous outcome, the consequences of which unfold over time, and that interact with later shocks. Klepper (2010) argues that the post-war development of Detroit—and that of Silicon Valley—was partly driven by successful spinoffs from high achieving firms, and organizational reproduction. A natural corollary is that, in the long term, success in narrowly defined areas could crowd out other entrepreneurial activities (Glaeser et al. 2010). A city or region could become locked in to particular sectors or lines of activity, bringing the risk of future decline. Martin and Sunley (2006) discuss work on path dependence in more detail, emphasizing that not much is known about why some regional economies lose dynamism, while others evolve and continually reinvent themselves.

Some argue that the solution to regional decline is to promote clusters of firms in particular sectors. The practical importance of clusters, as a source of higher productivity or a response to regional decline, continues to divide opinion.[44] The work of Klepper (2010) implies that understanding specific industrial clusters requires detailed attention to their genealogy. The most famous examples appear to have developed in a largely organic way, rather than through external intervention; there is room for debate over whether pro-cluster policies would be effective, even if desirable.

[43] See also Desmet and Rossi-Hansberg (2013b) for a quantitative account of the declining populations of Rust Belt cities, partly in terms of relatively large local frictions.

[44] See Duranton (2011) for an especially sceptical view.

It might be easier to achieve consensus when the analysis of regional decline focuses on the labor market. Kline and Moretti (2013) discuss the possibility of hiring subsidies that vary across locations, as a candidate place-based policy. An alternative approach emphasizes the potential benefits of local ownership: the argument is partly that locally owned firms are less likely to reduce employment in the face of negative shocks. Kolko and Neumark (2010) investigate this hypothesis for the US, finding that the greatest benefits come not from small independent businesses but corporate headquarters, followed by locally owned chains. A different place-based policy adopted by some countries, including the UK, is to locate public sector offices in depressed regions. Again, this may help to promote stability, although a general equilibrium analysis is needed.

These observations point to the importance of studying regional decline in more depth. In the meantime, reading Adam Smith, or for that matter most contemporary textbooks in economics, would provide little assistance to the citizens of cities and regions that confront decline. Their problems deserve more attention from economists. Until this happens, the mechanisms and costs of decline will be comprehended deeply only by those directly involved, and with much to lose. What the invisible hand gives, it can also take away.

4.9. CONCLUSIONS

The study of regional growth is often thought to be simpler and more straightforward than national growth. We have emphasized, instead, various ways in which it is harder. Regional outcomes are best seen in terms of a spatial equilibrium. Regions are interdependent, and their locations matter. For example, theoretical models predict that market access influences relative prosperity and population density, and these predictions are supported by a variety of evidence. Meanwhile, labor mobility implies that incomes, populations, and living costs are all endogenous, and must be considered jointly. The days when a textbook on regional economics could legitimately base most of its discussion on the neoclassical growth model are gone.

Some of the other inheritances from the cross-country literature have been problematic. Many empirical studies treat the observations on regional units as if they derived from independent entities. But with regional outcomes tightly linked in various ways, it is rarely straightforward to identify causal effects from regional data, or to relate the estimated effects to underlying quantities of interest. For example, in the regression-based studies, it is rarely clear how to interpret the estimated effects of a given variable on productivity or growth. Do these estimates hold constant the spatial distribution of population and economic activity, or do they partly reflect endogenous changes in agglomeration?

This distinction becomes especially important whenever a researcher seeks to draw lessons for national growth, or regional policies. The problem can be seen with an extreme example. If one region gains from a specific policy only by expanding at the expense of another, any analysis which implicitly holds fixed the spatial distributions of population and activity will be misleading about welfare effects. The arithmetic of regional policy is complicated, and what is an addition for one region may be subtraction from another, even when the policy aimed at multiplication. As elsewhere in the study of regional data, the quantitative application of structural models seems the most promising response.

We have also tried to highlight some areas where additional research seems especially needed. When economists consider the possibility of a regional problem, they typically examine disparities in average living standards, and their evolution over time. The empirical literature on this topic is vast. Yet some of the most important regional problems are likely to be those where areas are persistently losing population. We have emphasized that regional decline, conceived in these terms, is likely to be a distinct process, and one that has rarely been studied by economists.

In passing, we have also drawn attention to some of the burdens of adjustment, such as the non-pecuniary externalities that can arise through mobility. Formal models sometimes indicate that even modest economic changes, with similarly modest welfare effects, involve substantial redistributions of population and economic activity. One possible conjecture is that a market economy may sometimes involve "too much" ongoing relocation. Limiting mobility is rarely attractive, however. A more promising avenue would be to investigate forms of economic arrangements that lessen the need for mobility in the first place.

We will not attempt even a short summary of the forces that influence regional growth. Too much remains uncertain, and it is a sign of the current health of the literature that any such survey would quickly become dated. Instead, we have emphasized recent developments in the study of regional growth, both theoretical and empirical. As well as the impetus from the New Economic Geography literature, the increasing availability of (and interest in) regional data sets mean that the field is evolving at a great pace. Methods such as spatial discontinuity designs and the use of natural experiments have shed new light on causal effects, while the quantitative application of structural models is likely to be highly informative. Combined, these developments suggest that regional growth has become a particularly exciting area of economics, rich in data, interesting research questions, new methods, and increasingly sophisticated models. The study of growth has belatedly entered its own Space Age, and there is no going back.

4.10. APPENDIX: DATA AND METHODS

This appendix discusses regional price deflators; criticisms of the beta-convergence approach; and some alternative methods for studying regional growth.

4.10.1 Regional Price Deflators

Within countries, prices for identical goods often differ across locations. Ideally, it would be possible to compare output and income across regions in real terms, in the same way that the Penn World Table allows comparisons of real output across countries. Accurate comparisons of output or productivity across regions require PPP deflators or measures of regional output that aggregate goods and services at a common set of prices (such as producer or "mill" prices). Similarly, for the study of differences in the standard of living across regions, it would be useful to have cost-of-living deflators, partly based on housing costs.

In practice, real comparisons of regional output are rarely possible over long spans of time; only a few countries, including Canada and China, release data which allow for price differences across provinces. Sometimes, deflators may be available for just a few points in time. Aten (2008) and Aten and D'Souza (2008) have undertaken this for the US. The cross-country, cross-section data set of Mitton (2013) adjusts for some differences across regions in the cost of living, by linking regions to data on living costs for particular cities. More generally, measures of regional inequality can adjust for price differences at particular dates, but in principle, a researcher studying regional growth and convergence needs deflators for each date.

This makes it important to consider the main influences on regional price levels, and their variation over time. If labor is homogeneous and mobile, a spatial equilibrium requires real incomes to be equalized across locations; in that case, regions with higher nominal wages must have higher price levels for goods and services, and/or higher housing costs. This result emerges from general equilibrium models, such as those developed in Redding (2012). In his analysis, market access also matters: well-connected (less remote) regions will tend to have relatively low consumer prices for tradable goods. For a migration equilibrium, this must be offset by a higher population that drives up land prices and living costs, and hence equalizes real wages.

It seems likely, at least for developed countries, that national statistical agencies already have some of the raw price data needed to construct regional-level deflators. Deaton and Dupriez (2011) note that the agencies are "strangely reticent" on this topic.[45] For prices to be representative of a region, data on the spatial distribution of population are also needed; but if prices differ across locations, such data are needed in any case, to derive national-level deflators that are representative. For cost-of-living deflators, a major component is likely to be housing costs.

In the absence of official data, an open question is whether empirical researchers could make progress by imputing price levels. One approach to living costs assumes that households with the same budget share of food, but in different locations, have the same level of welfare; a comparison of their nominal expenditure levels then reveals the relative

[45] Nevertheless, work is likely to emerge using disaggregated data on purchases and prices, from other sources; see Handbury and Weinstein (2011).

price levels at the different locations. But the key assumption, that households with the same budget share of food have the same welfare, is strong; see Deaton and Dupriez (2011).

An alternative approach might use simple assumptions about the sensitivity of price levels to development levels or measures of market access, perhaps drawing on theoretical models. These relationships could then be used to map between observable variables and the unobserved true deflators, at least for the purpose of a sensitivity analysis. One question, which could be studied using the currently available data, is whether regional deflators are sometimes stable enough (relative to one another) that growth and convergence studies can give reliable answers even in their absence. A related question is the extent to which cost-of-living deflators can proxy for the price levels of output needed for productivity comparisons. The results of Redding (2012) suggest that this could be risky, not least if market access varies widely across locations.

4.10.2 Beta-Convergence

In their empirical work on convergence, Barro and Sala-i-Martin (1991, 2004) assume that steady-states are similar across regions. This assumption does a lot of work. It means that an explicit theory of steady-state positions is not required. From an econometric point of view, it provides a justification for studies of absolute convergence across regions, of the kind they and other authors have carried out. But in their work, the similarity of steady-states is assumed rather than established. It seems unattractive on theoretical grounds; in a market economy with labor mobility, the average product of labor will necessarily vary across regions, due to composition effects among other forces.

We could still ask whether their approach is informative about the extent of long-run disparities. One perspective on this is to look at the R^2 of an absolute convergence regression. A typical model would have the form:

$$(y_{it} - y_{it-\tau})/\tau = \eta + \beta y_{it-\tau} + \phi_t + \varepsilon_{it}, \tag{4.25}$$

where y_{it} is the logarithm of output per capita for region i at time t. Sala-i-Martin (1996) and Barro and Sala-i-Martin (2004) argue that estimates of β often correspond to a convergence rate of around 2% a year. The regressions that Barro and Sala-i-Martin present for the US states and Japanese prefectures (their Tables 11.2 and 11.2) often have a relatively low R^2 for short subperiods. But over longer spans of time (1880–2000 for the US, 1930–1990 for Japan) their simple regression has an R^2 of 0.92 for both countries. At first glance, this indicates that steady-state positions are similar. But this is misleading: their regression omits fixed effects, which could proxy for time-invariant determinants of relative income levels. In the absence of these fixed effects, it is likely that the parameter estimates are biased, and the high R^2 is misleading.

When the regression (4.25) is discussed in the literature, β is typically regarded as the main parameter of interest. For the study of regional growth, we should be interested in the more general model:

$$(y_{it} - y_{it-\tau})/\tau = \eta_i + \beta y_{it-\tau} + \phi_t + \varepsilon_{it}. \tag{4.26}$$

The variance of the region-specific effects (the η_i) will be denoted σ_η^2, and should be seen as a key parameter of interest. After taking out the common time effects ϕ_t, the model in (4.26) implies that region i is mean-reverting with long-run mean $\mu_i = -\eta_i/\beta$. The cross-section variance of μ_i therefore depends on β^2 and on σ_η^2. But also note that, given continued shocks, each region's output will continue to vary over time.

For the US, using a fixed effects estimator on 10-year subperiods doubles the estimated rate of convergence (results not reported). For a large sample of countries with sub-national data, Gennaioli et al. (2013b) find that including regional fixed effects greatly increases the estimated rate of convergence. The assumption in Barro and Sala-i-Martin that $\sigma_\eta^2 = 0$ seems hard to defend. Barro and Sala-i-Martin provide an alternative justification, which is that $y_{it-\tau}$ may be uncorrelated with η_i. But if the process has been running for any length of time, this alternative assumption is also unattractive, because a mean-reverting process such as (4.26) will necessarily generate a correlation between output per capita and the fixed effects.

There is another reason for querying this approach. In the cross-country literature, the process in (4.26) has a structural justification: it approximates transitional dynamics in the vicinity of a balanced growth path. But for regions, the neoclassical growth model should not be expected to apply, given inter-regional flows of capital and labor. Hence, for regional data, (4.26) is not structural, but only a way of capturing the time-series dependence in the data. Instead, Gennaioli et al. (2013b) suggest the use of a specification in which each region's factor input (perhaps some broad notion of capital) is a Cobb-Douglas function of its endowment of that factor—based on past investment—and the level that would obtain under full mobility. As they acknowledge, this assumption is ad hoc, but it leads to a simple specification which generalizes the standard conditional convergence regression. The extent of the barriers to factor mobility can be estimated from the data, although their estimates indicate higher barriers than might have been expected.

Whichever model is adopted, using a single lag may give a misleading picture. Regional living standards could be influenced by omitted variables which are themselves autocorrelated, and so e_{it} will be serially correlated. For regional data, a natural generalization of 4.26 is:

$$(y_{it} - y_{it-\tau})/\tau = \eta_i + \beta y_{it-\tau} + \phi_t + u_{it} + \varepsilon_{it}, \tag{4.27}$$

$$u_{it} = \rho u_{it-\tau} + v_{it}, \tag{4.28}$$

which implies:

$$(y_{it} - y_{it-\tau})/\tau = \eta_i' + (\beta + \rho/\tau) y_{it-\tau} - (\beta\rho + \rho/\tau) y_{it-2\tau} + \phi_t' + v_{it} + \varepsilon_{it} - \rho\varepsilon_{it-\tau},$$

and makes clear the likely inadequacy of a model with just one lag. There are further reasons that serial correlation is likely. In the cross-country literature, the neoclassical growth model can be used to argue that cross-section and time-series variation in the β parameter should be limited. This seems less plausible for regions, and the heterogeneity will lead to serially correlated errors. Measurement error, partly due to time-varying regional price levels, could also lead to serial correlation. These points suggest that beta-convergence regressions, with or without fixed effects, have significant weaknesses. The remainder of this appendix considers some alternatives.

4.10.3 Time-Series Approaches

Recent studies draw heavily on the implications of convergence for the time-series properties of regional data. Bernard and Durlauf (1996) showed how to relate different concepts of convergence to time-series properties. To fix ideas, we will initially consider how a researcher should proceed in the case of two regions. The choice of the null hypothesis needs thought, and should depend on the claim that a researcher is interested in seeking to falsify. If the two regions are believed to be on parallel growth paths, and a researcher wants to see if this claim can be falsified, a natural approach is to look at their (log) output gap and apply stationarity tests.

When the regions are genuinely following parallel growth paths, a stationarity test such as that of Kwiatkowski et al. (1992) should not reject the null of stationarity. Alternatively, if a researcher wants to examine a claim of divergence, the natural approach is to test whether the output gap contains either a stochastic trend (using a unit root test) or a deterministic linear time trend (as when log incomes in the two regions are trend-stationary processes with different trend growth rates).[46] Note, however, the maintained assumption that long-run steady-states are time-invariant, a point we return to shortly.

Extending these ideas to N regions, output gaps could be defined relative to a particular benchmark region, or a weighted average, as in early work such as Carlino and Mills (1993, 1996). But this approach becomes problematic if one or more regions are diverging from the others. The results will vary with the choice of benchmark, and using a weighted average will indicate non-convergence even when a subset of regions is moving together. In principle, a more attractive approach is to allow each region's growth to be a function of the $N-1$ output gaps with other regions (Carvalho and Harvey, 2005). But a flexible version of this, with separate catch-up coefficients for each ordered pair, implies $N(N-1)$ parameters and hence becomes difficult or impossible to estimate when the number of regions is large.[47]

[46] One potential complication here is that convergence could be present but slow, so that the log output gap is a fractionally integrated process. For a study that includes an application of this idea to data on the contiguous US states, see Mello (2011).

[47] A somewhat related approach is to apply multivariate tests for cointegration, such as Johansen's method, as in the early study of cross-country convergence by Bernard and Durlauf (1995). This approach can

An attractive alternative is that of Pesaran (2007), who develops a test based on all $N(N-1)/2$ pairwise output gaps. Taking the null of interest to be non-convergence, Pesaran shows that under this null, the fraction of pairwise output gaps for which a unit root is rejected should be close to the size of the unit root test that has been applied (e.g. 5%). The fraction of pairwise gaps for which a unit root is rejected can be taken as a measure of the extent of convergence. The detailed data on rejections will divide the sample into groups for which non-convergence is rejected, and regions that are evolving independently. This is more informative than collapsing the issue to a binary opposition between convergence and divergence.

Pesaran's approach has been applied to the contiguous US states by Mello (2011) and to European regions by Le Pen (2011). The fraction of pairwise gaps for which a unit root is rejected is typically low, suggesting non-convergence for the majority of regions, even though other tests provide clear evidence of mean reversion. In response to these findings, Le Pen (2011) argues for the importance of structural breaks—that is, mean shifts in the output gaps. But this highlights a fundamental dilemma for time-series approaches. For the tests to have some power, long spans of data are needed, but then it is harder to maintain the assumption that relative steady-states are time-invariant. If steady-states are evolving over time, this breaks the direct connection between time-series properties and convergence concepts.[48]

The dilemma arises partly from taking a univariate approach to a process influenced by a wider range of variables. In the literature on national growth, the steady-state positions are typically modeled as stable functions of a few variables, as in Mankiw et al. (1992). This is harder to implement for regional data, partly because data on potential control variables are often lacking, and partly because interesting models of regional disparities may not yield simple expressions for steady-states. At least as a way of describing the data, an alternative approach uses the behavior of the cross-section dispersion (or inequality) in regional income to draw conclusions about the underlying statistical processes. Evans (1996) showed that if the units such as regions follow independent random walks, then the (cross-section) log variance will be integrated of order one around an upward quadratic trend. If the regions are instead believed to have converged and to be driven by a common trend, the log variance will be stationary and fluctuate around a constant mean.[49] This notion of convergence does not require the cross-section variance to decline monotonically to zero, an outcome that is unlikely for a collection of stochastic processes. Evans (2000) includes an application of this idea to the contiguous US states.

provide evidence on the number of common stochastic trends likely to be driving the output movements of the N regions. But again, it becomes infeasible when the number of regions is large.

[48] Note that a pairwise output gap process which is stationary, but with mean shifts, is compatible with either catching-up or divergence, depending on whether the mean of the output gap shifts downwards or upwards, respectively. The time-series approach then becomes harder to implement, and interpret.

[49] For some related discussion, see Ng (2008) and Pesaran (2007).

Another route is to develop methods for describing growth paths of economies that are converging, while separating out long-run effects from cyclical components, by using unobserved component models. For studies of US convergence that adopt this approach, see Carvalho and Harvey (2005) and Carvalho et al. (2007).

4.10.4 Distribution Dynamics

A popular approach has been distribution dynamics, developed for cross-country data by Quah (1993) and applied to regional data by Quah (1996). This approach characterizes transitions of income per capita between income classes, using a transition matrix whose elements are the probabilities of moving from one income class to another. There are important ways in which this is more flexible than a panel data model, and more informative about the underlying process. It provides direct information about mobility between income classes, and the stationary distribution implied by a given transition matrix will reveal tendencies latent in observed realizations of income levels (Quah, 1993). Under the strong assumption that the transition probabilities remain stable over time, the stationary distribution provides a long-run forecast of the shape of the distribution of regional income levels.

Kremer et al. (2001) make the useful observation that, when considering income levels for aggregate economic units, banded into wide classes, it is likely that the only non-zero transition probabilities are those between adjacent income classes. More dramatic relative movements are unlikely for countries or regions, at least over short spans of time. In this case, the ratios of the individual elements of the stationary distribution can be derived as ratios of transition probabilities. But since a ratio can be sensitive to a small change in its denominator, the estimated stationary distribution may be sensitive to small changes in the estimated transition probabilities. Hence, at least when the stationary distribution is the main result of interest, one drawback of this approach is a lack of robustness. Kremer et al. suggest an alternative method, which is to iterate the estimated process over a limited number of future periods and study the outcome, rather than emphasizing the stationary distribution.

A further problem arises from the discretization that is often used to construct the transition matrix. An alternative is to treat the state space as continuous and model the joint distribution of outcomes at t and $t + \tau$, as in the cross-country work of Quah (1997) and Johnson (2005), for example. But given the number of regions typically available to a researcher, there is not a great deal of information from which to estimate something as complex as a joint distribution, again implying a lack of robustness.

Despite its problems, an attractive aspect of the distribution dynamics approach is that it can be used to investigate regional polarization. Quah's work on national growth is strongly associated with his "twin peaks" result, the finding that the stationary distribution is bimodal. In the regional growth literature, his methods have been the most popular approach to the study of polarization. Some other methods are available, using various ways of defining polarization (see Anderson et al. 2012; Zhang and Kanbur, 2001).

4.10.5 Multimodality and Mixture Densities

The hypothesis of polarization is a special case of a more general idea, that of convergence clubs. A common intuition is that, over time, regions might sort into distinct groups or clubs, such as rich and poor. Their respective positions could then reflect disparate steady-states, or even the possibility of multiple equilibria. These ideas have been discussed repeatedly in the literature. Using Quah's methods can provide some insight, but approaches have emerged which provide more direct information on the existence of clubs and their membership.

One approach is bump hunting, or the use of formal statistical tests for multimodality, as in Pittau and Zelli (2006). But for many purposes, a more informative approach is to model the cross-section distribution of a regional variable as a mixture distribution. To give an example from regional economics, if regions tend to belong either to an industrialized and services-oriented urban core or to a rural, agricultural periphery, the data might be generated by a mixture distribution with two components. The data for a given region are then drawn from one component distribution with some probability, and the other component with the complementary probability; the idea generalizes readily to mixtures with more than two components. Methods for finite mixtures can be adopted to estimate characteristics of the components, such as means and variances, and provide a probabilistic classification that can be used to assign (fuzzily) any given region to one of the component distributions.

For investigating convergence clubs, alternatives to the mixture density approach address parameter heterogeneity in various ways. Canova (2004) is one of the first contributions along these lines. Other methods for sample-splitting include a regression tree approach as in Johnson and Takeyama (2001), or the methods for inference for threshold estimation developed by Hansen (2000). These approaches typically invoke simple parametric models estimated on subsamples, indicating the extent of parameter heterogeneity. This seems most useful when the hypotheses of interest can be captured by simple regression specifications, but as we have emphasized, general equilibrium models often rule this out.

ACKNOWLEDGMENTS

We thank Paddy Carter, Klaus Desmet, Adeel Malik, Annalisa Marini, Stephen Redding, Esteban Rossi-Hansberg, and Helen Simpson for various helpful contributions. The usual disclaimer applies.

REFERENCES

Abadie, Alberto, Diamond, Alexis, Hainmueller, Jens, 2012. Comparative Politics and the Synthetic Control Method. Manuscript, Harvard, June.

Abadie, Alberto, Gardeazabal, Javier, 2003. The economic costs of conflict: a case study of the Basque Country. American Economic Review 93 (1), 113–132.

Abramovsky, Laura, Harrison, Rupert, Simpson, Helen, 2007. University research and the location of business R&D. Economic Journal 117 (519), C114–C141.

Acemoglu, Daron, Cantoni, Davide, Johnson, Simon, Robinson, James A., 2011. The consequences of radical reform: the French Revolution. American Economic Review 101 (7), 3286–3307.

Acemoglu, Daron, Dell, Melissa, 2010. Productivity differences between and within countries. American Economic Journal: Macroeconomics 2 (1), 169–188.

Acemoglu, Daron, García-Jimeno, Camilo, Robinson, James A., 2012. Finding Eldorado: slavery and long-run development in Colombia. Journal of Comparative Economics 40 (4), 534–564.

Acemoglu, Daron, Hassan, Tarek A., Robinson, James A., 2011. Social structure and development: a legacy of the Holocaust in Russia. Quarterly Journal of Economics 126 (2), 895–946.

Aghion, Philippe, Burgess, Robin, Redding, Stephen J., Zilibotti, Fabrizio, 2008. The unequal effects of liberalization: evidence from dismantling the License Raj in India. American Economic Review 98 (4), 1397–1412.

Aiello, Francesco, Scoppa, Vincenzo, 2000. Uneven regional development in Italy: explaining differences in productivity levels. Giornale degli Economisti e Annali di Economia 59 (2), 270–298.

Alder, Simeon, Lagakos, David, Ohanian, Lee, 2013. The Decline of the US Rust Belt: A Macroeconomic Analysis. Manuscript, January.

Amiti, Mary, Cameron, Lisa, 2007. Economic geography and wages. Review of Economics and Statistics 89 (1), 15–29.

Anas, Alex, Arnott, Richard, Small, Kenneth A., 1998. Urban spatial structure. Journal of Economic Literature 36 (3), 1426–1464.

Anderson, Gordon, Linton, Oliver, Leo, Teng, 2012. A polarization-cohesion perspective on cross-country convergence. Journal of Economic Growth 17 (1), 49–69.

Anselin, Luc, 2001. Spatial econometrics. In: Baltagi, Badi H. (Ed.), A Companion to Theoretical Econometrics. Blackwell, Oxford.

Anselin, Luc (2006). Spatial econometrics. In: Mills, Terence C., Patterson, Kerry (Eds.), Palgrave Handbook of Econometrics, Volume 1: Econometric Theory. Palgrave Macmillan, Basingstoke.

Arkolakis, Costas, Costinot, Arnod, Rodríguez-Clare, Andrés, 2012. New trade models, same old gains? American Economic Review 102 (1), 94–130.

Armstrong, Harvey, Taylor, Jim, 2000. Regional Economics and Policy, third ed. Blackwell, Oxford.

Aten, Bettina H., 2008. Estimates of State and Metropolitan Price Parities for Consumption Goods and Services in the United States, 2005. BEA Paper.

Aten, Bettina H., D'Souza, Roger J., 2008. Regional price parities: comparing price level differences across geographic areas. Survey of Current Business 88, 64–74.

Azzoni, Carlos R., 2001. Economic growth and regional income inequality in Brazil. Annals of Regional Science 35 (1), 133–152.

Bairoch, Paul, 1993. Economics and World History. Harvester Wheatsheaf, Hemel Hempstead.

Baland, Jean-Marie, Robinson, James A., 2008. Land and power: theory and evidence from Chile. American Economic Review 98 (5), 1737–1765.

Baldwin, Richard E., Forslid, Rikard, 2000. The core-periphery model and endogenous growth: stabilizing and destabilizing integration. Economica 67 (267), 307–324.

Baldwin, Richard E., Forslid, Rikard, Martin, Philippe, Ottaviano, Gianmarco I.P., Robert-Nicoud, Frederic, 2003. Economic Geography and Public Policy. Princeton University Press, Princeton.

Baldwin, Richard E., Martin, Philippe, 2004. Agglomeration and regional growth. In: Henderson, J.V., Thisse, J.F. (Eds.), Handbook of Regional and Urban Economics, vol 4. Elsevier, Amsterdam, pp. 2671–2711.

Baldwin, Richard E., Martin, Philippe, Ottaviano, Gianmarco I.P., 2001. Global income divergence, trade, and industrialization: the geography of growth take-offs. Journal of Economic Growth 6 (1), 5–37.

Baldwin, Richard E., Okubo, Toshihiro, 2006. Heterogeneous firms, agglomeration and economic geography: spatial selection and sorting. Journal of Economic Geography 6 (3), 323–346.

Banerjee, Abhijit, Duflo, Esther, Qian, Nancy, 2012. On the Road: Access to Transportation Infrastructure and Economic Growth in China. NBER Working Paper No. 17897.

Banerjee, Abhijit, Iyer, Lakshmi, 2005. History, institutions, and economic performance: the legacy of colonial land tenure. American Economic Review 95 (4), 1190–1213.

Barrios, Salvador, Strobl, Eric, 2009. The dynamics of regional inequalities. Regional Science and Urban Economics 39 (5), 575–591.

Barrios, Thomas, Diamond, Rebecca, Imbens, Guido W., Kolesár, Michal, 2012. Clustering, spatial correlations, and randomization inference. Journal of the American Statistical Association 107 (498), 578–591.

Barro, Robert J., Sala-i-Martin, Xavier, 1991. Convergence across States and Regions. Brookings Papers on Economic Activity 22 (1), 107–182.

Barro, Robert J., Sala-i-Martin, Xavier, 2004. Economic Growth, second ed. MIT Press, Cambridge, MA.

Becker, Sascha O., Boeckh, Katrin, 2011. The Empire is Dead, Long Live the Empire! Long-Run Persistence of Trust and Corruption in the Bureaucracy, CEPR Discussion Paper No 8288.

Becker, Sascha O., Woessmann, Ludger, 2009. Was Weber wrong? A human capital theory of protestant economic history. Quarterly Journal of Economics 124 (2), 531–596.

Beeson, Patricia E., 1991. Amenities and regional differences in returns to worker characteristics. Journal of Urban Economics 30 (2), 224–241.

Behrens, Kristian, Robert-Nicoud, Frederic, 2012. Survival of the Fittest in Cities: Agglomeration, Selection and Polarisation. CIRPÉE Discussion Paper No. 09–19, revised February 2012.

Behrens, Kristian, Duranton, Gilles, Robert-Nicoud, Frederic, 2010. Productive Cities: Sorting, Selection and Agglomeration. CEPR Discussion Paper No. 7922.

Behrens, Kristian, Mion, Giordano, Murata, Yasu, Südekum, Jens, 2011. Spatial Frictions. CEPR Discussion Paper No. 8572.

Belot, Michèle, Ermisch, John, 2009. Friendship ties and geographical mobility: evidence from Great Britain. Journal of the Royal Statistical Society: Series A 172 (2), 427–442.

Bencivenga, Valerie R., Smith, Bruce D., 1997. Unemployment, migration, and growth. Journal of Political Economy 105 (3), 582–608.

Berkowitz, Daniel, Hoekstra, Mark, Schoors, Koen, 2012. Does Finance Cause Growth? Evidence from the Origins of Banking in Russia, NBER Working Paper No 18139.

Bernard, Andrew B., Durlauf, Steven N., 1995. Convergence in international output. Journal of Applied Econometrics 10 (2), 97–108.

Bernard, Andrew B., Durlauf, Steven N., 1996. Interpreting tests of the convergence hypothesis. Journal of Econometrics 71 (1–2), 161–173.

Bernard, Andrew, Eaton, Jonathan, Jensen, J. Bradford, Kortum, Samuel, 2003. Plants and productivity in international trade. American Economic Review 93 (4), 1268–1290.

Bernard, Andrew B., Redding, Stephen J., Schott, Peter K., 2013. Testing for factor price equality with unobserved differences in factor quality or productivity. American Economic Journal: Microeconomics 5 (2), 135–63..

Besley, Timothy, Burgess, Robin, 2000. Land reform, poverty reduction, and growth: evidence from India. Quarterly Journal of Economics 115 (2), 389–430.

Besley, Timothy, Burgess, Robin, 2002. The political economy of government responsiveness: theory and evidence from India. Quarterly Journal of Economics 117 (4), 1415–1451.

Besley, Timothy, Burgess, Robin, 2004. Can labor regulation hinder economic performance? Evidence from India. Quarterly Journal of Economics 119 (1), 91–134.

Black, Duncan, Henderson, J. Vernon, 1999. A theory of urban growth. Journal of Political Economy 107 (2), 252–284.

Bleakley, Hoyt, Lin, Jeffrey, 2012. Portage and path dependence. Quarterly Journal of Economics 127 (2), 587–644.

Bloom, David E., Sachs, Jeffrey D., 1998. Geography, demography, and economic growth in Africa. Brookings Papers on Economic Activity 29 (2), 207–296.

Bosker, Maarten, Garretsen, Harry, 2012. Economic geography and economic development in Sub-Saharan Africa. The World Bank Economic Review 26 (3), 93–136.

Boucekkine, Raouf, Camacho, Carmen, Zou, Benteng, 2009. Bridging the gap between growth theory and the new economic geography: the spatial Ramsey model. Macroeconomic Dynamics 13 (1), 20–45.

Brakman, Steven, Garretsen, Harry, Schramm, Marc, 2004a. The spatial distribution of wages: estimating the Helpman-Hanson model for Germany. Journal of Regional Science 44 (3), 437–466.

Brakman, Steven, Garretsen, Harry, Schramm, Marc, 2004b. The strategic bombing of German cities during World War II and its impact on city growth. Journal of Economic Geography 4 (2), 201–218.

Brakman, Steven, Garretsen, Harry, van Marrewijk, Charles, 2009. The New Introduction to Geographical Economics. Cambridge University Press, Cambridge.

Brandt, Loren, Tombe, Trevor, Zhu, Xiaodong, 2013. Factor market distortions across time, space and sectors in China. Review of Economic Dynamics 16 (1), 39–58.

Breinlich, Holger, 2006. The spatial income structure in the European Union—what role for economic geography? Journal of Economic Geography 6 (5), 593–617.

Brito, Paulo, 2004. The Dynamics of Growth and Distribution in a Spatially Heterogeneous World. Working Papers, Department of Economics, ISEG, WP13/2004/DE/UECE.

Brock, William, Xepapadeas, Anastasios, 2008. Diffusion-induced instability and pattern formation in infinite horizon recursive optimal control. Journal of Economic Dynamics and Control 32 (9), 2745–2787.

Brock, William, Xepapadeas, Anastasios, 2009. General Pattern Formation in Recursive Dynamical Systems Models in Economics. Fondazione Eni Enrico Mattei Working Paper No. 2009.49.

Brülhart, Marius, Carrère, Céline, Trionfetti, Federico, 2012. How wages and employment adjust to trade liberalization: Quasi-experimental evidence from Austria. Journal of International Economics 86 (1), 68–81.

Burgess, Robin, Pande, Rohini, 2005. Do rural banks matter? Evidence from the Indian social banking experiment. American Economic Review 95 (3), 780–795.

Burgess, Robin, Pande, Rohini, Wong, Grace, 2005. Banking for the poor: evidence from India. Journal of the European Economic Association 3 (2–3), 268–278.

Cameron, Gavin, Muellbauer, John, 2000. Earnings biases in the United Kingdom regional accounts: some economic policy and research implications. Economic Journal 110, F412–F429.

Cameron, Gavin, Muellbauer, John, 2001. Earnings, unemployment, and housing in Britain. Journal of Applied Econometrics 16 (3), 203–220.

Canova, Fabio, 2004. Testing for convergence clubs in income per capita: a predictive density approach. International Economic Review 45 (1), 49–77.

Cantoni, Davide, 2010. The Economic Effects of the Protestant Reformation: Testing the Weber Hypothesis in the German Lands. Universitat Pompeu Fabra, December, Manuscript.

Carlino, Gerald, Mills, Leonard O., 1993. Are US regional incomes converging? A time series analysis. Journal of Monetary Economics 32 (2), 335–346.

Carlino, Gerald, Mills, Leonard O., 1996. Are US regional incomes converging? Reply. Journal of Monetary Economics 38 (3), 599–601.

Carvalho, Vasco M., Harvey, Andrew C., 2005. Growth, cycles and convergence in US regional time series. International Journal of Forecasting 21 (4), 667–686.

Carvalho, Vasco M., Harvey, Andrew C., Trimbur, Thomas, 2007. A note on common cycles, common trends, and convergence. Journal of Business & Economic Statistics 25 (1), 12–20.

Caselli, Francesco, 2005. Accounting for cross-country income differences. In: Aghion, Philippe, Durlauf, Steven N. (Eds.), Handbook of Economic Growth, vol. 1A. North-Holland, New York.

Chen, Henry, Gompers, Paul, Kovner, Anna, Lerner, Josh, 2010. Buy local? The geography of venture capital. Journal of Urban Economics 67 (1), 90–102.

Chen, Xi, Nordhaus, William, 2011. Using luminosity data as a proxy for economic statistics. Proceedings of the National Academy of Sciences (US) 108 (21), 8589–8594.

Clark, Gordon L., Gertler, Meric S., Feldman, Maryann P. (Eds.), 2000. The Oxford Handbook of Economic Geography. Oxford University Press, Oxford.

Combes, Pierre-Philippe, 2011. The empirics of economic geography: how to draw policy implications? Review of World Economics 147 (3), 567–592.

Combes, Pierre-Philippe, Duranton, Gilles, Gobillon, Laurent, 2008. Spatial wage disparities: sorting matters! Journal of Urban Economics 63 (2), 723–742.

Combes, Pierre-Philippe, Duranton, Gilles, Gobillon, Laurent, Puga, Diego, Roux, Sébastien, 2012. The productivity advantages of large cities: distinguishing agglomeration from firm selection. Econometrica 80 (6), 2543–2594.

Combes, Pierre-Philippe, Mayer, Thierry, Thisse, Jacques-François, 2008. Economic Geography. Princeton University Press, Princeton NJ.

Conley, Timothy G., 1999. GMM estimation with cross sectional dependence. Journal of Econometrics 92 (1), 1–45.

Corden, W.Max, Findlay, Ronald, 1975. Urban unemployment, intersectoral capital mobility and development policy. Economica 42 (165), 59–78.

Córdoba, Juan Carlos, 2008. On the distribution of city sizes. Journal of Urban Economics 63 (1), 177–197.

Corrado, Luisa, Fingleton, Bernard, 2012. Where is the economics in spatial econometrics? Journal of Regional Science 52 (2), 210–239.

Cowell, Frank, 2011. Measuring Inequality. Oxford University Press, Oxford.

Crespo Cuaresma, Jesús, Feldkircher, Martin, 2013. Spatial filtering, model uncertainty and the speed of income convergence in Europe. Journal of Applied Econometrics 28 (4), 720–741.

Cuñat, Alejandro, Maffezzoli, Marco, 2007. Can comparative advantage explain the growth of US trade? Economic Journal 117 (520), 583–602.

Dahl, Michael S., Sorenson, Olav, 2010. The migration of technical workers. Journal of Urban Economics 67 (1), 33–45.

Davis, Donald R., Dingel, Jonathan I., 2012. A Spatial Knowledge Economy. NBER Working Paper No. 18188.

Davis, Donald, Weinstein, David, 1999. Economic geography and regional production structure: an empirical investigation. European Economic Review 43 (2), 379–407.

Davis, Donald, Weinstein, David, 2002. Bones, bombs, and break points: the geography of economic activity. American Economic Review 92 (5), 1269–1289.

Davis, Donald, Weinstein, David, 2003. Market access, economic geography and comparative advantage: an empirical assessment. Journal of International Economics 59 (1), 1–23.

Davis, Donald, Weinstein, David, 2008. A search for multiple equilibria in urban industrial structure. Journal of Regional Science 48 (1), 29–65.

Deaton, Angus, Dupriez, Olivier, 2011. Spatial Price Differences Within Large Countries. Princeton University, July, Manuscript.

Dell, Melissa, 2010. The persistent effects of Peru's mining mita. Econometrica 78 (6), 1863–1903.

Dell, Melissa, Jones, Benjamin, Olken, Benjamin, 2009. Temperature and incpme: reconciling new cross-sectional and panel estimates. American Economic Review 99 (2), 198–204.

Dell, Melissa, Jones, Benjamin, Olken, Benjamin, 2012. Temperature shocks and economic growth: evidence from the last half century. American Economic Journal: Macroeconomics 4 (3), 66–95.

Démurger, Sylvie, Sachs, Jeffrey D., Woo, Wing Thye, Bao, Shuming, Chang, Gene, Mellinger, Andrew, 2002. Geography, economic policy, and regional development in China, Asian Economic Papers 1 (1), 146–197.

Desmet, Klaus, Ghani, Ejaz, O'Connell, Stephen D., Rossi-Hansberg, Esteban (2012). The Spatial Development of India. World Bank Policy Research Working Paper No. 6060.

Desmet, Klaus, Rossi-Hansberg, Esteban, 2009. Spatial growth and industry age. Journal of Economic Theory 144 (6), 2477–2502.

Desmet, Klaus, Rossi-Hansberg, Esteban, 2010. On spatial dynamics. Journal of Regional Science 50 (1), 43–63.

Desmet, Klaus, Rossi-Hansberg, Esteban, 2012a. Spatial Development. Manuscript, earlier version NBER Working Paper No. 15349.

Desmet, Klaus, Rossi-Hansberg, Esteban, 2012b. Innovation in space. American Economic Review, Papers and Proceedings 102 (3), 447–52.

Desmet, Klaus, Rossi-Hansberg, Esteban, 2013a. Urban Accounting and Welfare. American Economic Review.

Desmet, Klaus, Rossi-Hansberg, Esteban, 2013b. On the spatial economic impact of global warming. Princeton, Manuscript.

Diamond, Jared, Robinson, James A., 2010. Natural Experiments of History. Harvard, Belknap.

Dinkelman, Taryn, 2011. The effects of rural electrification on employment: new evidence from South Africa. American Economic Review 101 (7), 3078–3108.

Dinkelman, Taryn, Schulhofer-Wohl, Sam, 2012. Migration, Congestion Externalities, and the Evaluation of Spatial Investments. CEPR Discussion Paper No. 9126.

Dixit, Avinash K., Stiglitz, Joseph E., 1977. Monopolistic competition and optimum product diversity. American Economic Review 67 (3), 297–308.

Donaldson, Dave, 2010. Railroads of the Raj: Estimating the Impact of Transportation Infrastructure. NBER Working Paper No. 16487. American Economic Review.

Donaldson, Dave, Hornbeck, Richard, 2012. Railroads and American Economic Growth: A "Market Access" Approach. Manuscript, March.

Dorling, Daniel, 2011. Injustice. Policy Press, Bristol.

Drèze, Jean, Sen, Amartya, 1997. Indian Development: Selected Regional Perspectives, WIDER Studies in Development Economics, Clarendon Press, Oxford.

Duclos, Jean-Yves, Esteban, Joan, Ray, Debraj, 2004. Polarization: concepts, measurement, estimation. Econometrica 72 (6), 1737–1772.

Duranton, Gilles, 2007. Urban evolutions: the fast, the slow, and the still. American Economic Review 97 (1), 197–221.

Duranton, Gilles, 2011. "California dreamin": the feeble case for cluster policies. Review of Economic Analysis 3 (1), 3–45.

Duranton, Gilles, Gobillon, Laurent, Overman, Henry G., 2011. Assessing the effects of local taxation using microgeographic data. Economic Journal 121 (555), 1017–1046.

Duranton, Gilles, Monastiriotis, Vassilis, 2002. Mind the gaps: the evolution of regional earnings inequalities in the UK 1982–1997. Journal of Regional Science 42 (2), 219–256.

Duranton, Gilles, Puga, Diego, 2004. Micro-foundations of urban agglomeration economies. In: Henderson, Vernon, Thisse, Jacques-François (Eds.), Handbook of Regional and Urban Economics, vol. 4. North-Holland, Amsterdam.

Duranton, Gilles, Rodríguez-Pose, Andrés, Sandall, Richard, 2009. Family types and the persistence of regional disparities in Europe. Economic Geography 85 (1), 23–47.

Durbin, James, 1960. The fitting of time-series models. Review of the International Statistical Institute 28 (3), 233–244.

Durlauf, Steven N., 2012. Poverty traps and Appalachia. In: Ziliak, James (Ed.), Appalachian Legacy: Economic Opportunity After the War on Poverty. Brookings Institution Press, Washington DC.

Durlauf, Steven N., Johnson, Paul A., Temple, Jonathan R.W., 2005. Growth econometrics. In: Aghion, P., Durlauf, S.N. (Eds.), Handbook of Economic Growth, vol. 1A. North-Holland, Amsterdam, pp. 555–677.

Durlauf, Steven N., Johnson, Paul A., Temple, Jonathan R.W., 2009. The econometrics of convergence. In: Mills, Terence C., Patterson, Kerry (Eds.), Palgrave Handbook of Econometrics, vol. 2: Applied Econometrics. Palgrave Macmillan.

Eaton, Jonathan, Eckstein, Zvi, 1997. Cities and growth: theory and evidence from France and Japan. Regional Science and Urban Economics 27 (4–5), 443–474.

Eaton, Jonathan, Kortum, Samuel, 1999. International technology diffusion: theory and measurement. International Economic Review 40 (3), 537–570.

Eaton, Jonathan, Kortum, Samuel, 2002. Technology, geography, and trade. Econometrica 70 (5), 1741–1779.

Eberhardt, Markus, Teal, Francis, 2011. Econometrics for grumblers: a new look at the literature on cross-country growth empirics. Journal of Economic Surveys 25 (1), 109–155.

Eeckhout, Jan, 2004. Gibrat's law for (all) cities. American Economic Review 94 (5), 1429–1451.

Ellison, Glenn, Glaeser, Edward, 1999. The geographic concentration of industry: does natural advantage explain agglomeration? American Economic Review: Papers and Proceedings 89 (3), 311–316.

Enflo, Kerstin Sofia, 2010. Productivity and employment—is there a trade-off? Comparing Western European regions and American states 1950–2000. Annals of Regional Science 45 (2), 401–421.

Evans, Paul, 1996. Using cross-country variances to evaluate growth theories. Journal of Economic Dynamics and Control 20 (6–7), 1027–1049.

Evans, Paul, 2000. Income dynamics in regions and countries. In: Hess, Gregory D., van Wincoop, Eric (Eds.), Intranational Macroeconomics. Cambridge University Press, Cambridge.

Evans, Paul, Karras, Georgios, 1996. Do economies converge? Evidence from a panel of US states. Review of Economics and Statistics 78 (3), 384–388.

Fafchamps, Marcel, Schündeln, Matthias, 2013. Local financial development and firm performance: evidence from Morocco. Journal of Development Economics 103, 15–28.

Fally, Thibault, Paillacar, Rodrigo, Terra, Cristina, 2010. Economic geography and wages in Brazil: evidence from micro-data. Journal of Development Economics 91 (2), 155–168.

Fedorov, Leonid, 2002. Regional inequality and regional polarization in Russia, 1990–99. World Development 30 (3), 443–456.

Feldman, Maryann P., 1994. The Geography of Innovation. Kluwer Academic Publishers, Dordrecht.

Felkner, John S., Townsend, Robert M., 2011. The geographic concentration of enterprise in developing countries. Quarterly Journal of Economics 126 (4), 2005–2061.

Florida, Richard, 2002. The Rise of the Creative Class. Basic Books, New York.

Fogel, Robert W., 1964. Railroads and American Economic Growth: Essays in Economic History. Johns Hopkins University Press, Baltimore.

Fujita, Masahisa, Krugman, Paul R., Venables, Anthony J., 1999. The Spatial Economy: Cities, Regions and International Trade. MIT Press, Cambridge MA.

Fujita, Masahisa, Ogawa, Hideaki, 1982. Multiple equilibria and structural transition of non-monocentric urban configurations. Regional Science and Urban Economics 12 (2), 161–196.

Fujita, Masahisa, Thisse, Jacques-François, 2002. Economics of Agglomeration. Cambridge University Press, Cambridge.

Fujita, Masahisa, Thisse, Jacques-François, 2003. Does geographical agglomeration foster economic growth? And who gains and loses from it? Japanese Economic Review 54 (2), 121–145.

Gabaix, Xavier, 1999. Zipf's law for cities: an explanation. Quarterly Journal of Economics 114 (3), 739–767.

Gabaix, Xavier, 2011. The granular origins of aggregate fluctuations. Econometrica 79 (3), 733–772.

Gabriel, Stuart A., Mattey, Joe P., Wascher, William L., 2003. Compensating differentials and evolution in the quality-of-life among U.S. states. Regional Science and Urban Economics 33 (5), 619–649.

Gajwani, Kiran, Kanbur, Ravi, Zhang, Xiaobo, 2006. Comparing the Evolution of Spatial Inequality in China and India: A Fifty-Year Perspective. DSGD Discussion Papers 44, International Food Policy Research Institute (IFPRI).

Gallup, John Luke, Sachs, Jeffrey D., Mellinger, Andrew D., 1999. Geography and economic development. International Regional Science Review 22 (2), 179–232.

Ganong, Peter, Shoag, Daniel, 2013. Why has Regional Income Convergence in the US Declined? Manuscript, Harvard.

Gennaioli, Nicola, La Porta, Rafael, Lopez-de-Silanes, Florencio, and Shleifer, Andrei, 2013a. Human capital and regional development. Quarterly Journal of Economics 128 (1), 105–164.

Gennaioli, Nicola, La Porta, Rafael, Lopez-de-Silanes, Florencio, and Shleifer, Andrei (2013b). Growth in Regions. NBER Working Paper No. 18937.

Gibbons, Stephen, Overman, Henry, 2012. Mostly pointless spatial econometrics? Journal of Regional Science 52 (2), 172–191.

Gibson, Edward L., 2005. Boundary control: subnational authoritarianism in democratic countries. World Politics 58 (1), 101–132.

Gingerich, Daniel W., 2013. Governance indicators and the level of analysis problem: empirical findings from South America. British Journal of Political Science 2, 1–38.

Glaeser, Edward L., 2011. Triumph of the City. Macmillan, London.

Glaeser, Edward L., Gottlieb, Joshua D., 2009. The wealth of cities: agglomeration economies and spatial equilibrium in the United States. Journal of Economic Literature 47 (4), 983–1028.

Glaeser, Edward L., Kerr, Sari Pekkala, Kerr, William R., 2012. Entrepreneurship and Urban Growth: An Empirical Assessment with Historical Mines. NBER Working Paper No. 18333.

Glaeser, Edward L., Rosenthal, Stuart S., Strange, William C., 2010. Urban economics and entrepreneurship. Journal of Urban Economics 67 (1), 1–14.

Gollin, Douglas, Parente, Stephen L., Rogerson, Richard, 2004. Farm work, home work, and international productivity differences. Review of Economic Dynamics 7 (4), 827-850.

Graham, Bryan S., Temple, Jonathan R.W., 2006. Rich nations, poor nations: how much can multiple equilibria explain? Journal of Economic Growth 11 (1), 5–41.

Grosfeld, Irena, Zhuravskaya, Ekaterina, 2012. Persistent Effects of Empires: Evidence from the Partitions of Poland. Discussion Paper, SSRN.

Grossman, Gene, Helpman, Elhanan, 1991. Innovation and Growth in the Global Economy. MIT Press, Cambridge MA.

Guiso, Luigi, Sapienza, Paola, Zingales, Luigi, 2004. Does local financial development matter? Quarterly Journal of Economics 119 (3), 929–969.

Guriev, Sergei, Vakulenko, Elena, 2012. Convergence Between Russian regions. CEFIR/NES Working Paper No. 180.

Handbury, Jessie, Weinstein, David E., 2011. Is New Economic Geography Right? Evidence from Price Data. NBER Working Paper No. 17067, May.

Hanson, Gordon, 1996. Localization economies, vertical organization, and trade. American Economic Review 86 (5), 1266–1278.

Hanson, Gordon, 1997. Increasing returns, trade, and the regional structure of wages. Economic Journal 107 (440), 113–133.

Hanson, Gordon, 2005. Market potential, increasing returns, and geographic concentration. Journal of International Economics 67 (1), 1–24.

Hanson, Gordon, Xiang, Chong, 2004. The home market effect and bilateral trade patterns. American Economic Review 94 (4), 1108–1129.

Harari, Mariaflavia, La Ferrara, Eliana, 2013. Conflict, Climate and Cells: A Disaggregated Analysis. CEPR Discussion Paper No. 9277.

Harris, Chauncy, 1954. The market as a factor in the localization of industry in the United States. Annals of the Association of American Geographers 44 (2), 315–348.

Harris, Richard, 2010. Models of regional growth: past, present and future. Journal of Economic Surveys 25 (5), 913–951.

Harris, John R., Todaro, Michael P. 1970. Migration, unemployment and development: A two-sector analysis. American Economic Review 60 (1), 126–142.

Hassler, John, Krusell, Per, 2012. Economics and Climate Change: Integrated Assessment in a Multi-Region World. Manuscript, IIES, Stockholm.

Head, Keith, Mayer, Thierry, 2004. Empirics of agglomeration and trade. In: Henderson, Vernon, Thisse, Jacques-François (Eds.), Handbook of Regional and Urban Economics, vol. 4. North-Holland, Amsterdam.

Head, Keith, Mayer, Thierry, 2006. Regional wage and employment responses to market potential in the EU. Regional Science and Urban Economics 36 (5), 573–595.

Head, Keith, Mayer, Thierry, 2010. Gravity, market potential and economic development. Journal of Economic Geography 10 (1), 1–14.

Head, Keith, Ries, John, 2001. Increasing returns versus national product differentiation as an explanation for the pattern of US-Canada trade. American Economic Review 91 (4), 858–876.

Helpman, Elhanan, 1998. The size of regions. In: Pines, D., Sadka, E., Zilcha, I. (Eds.), Topics in Public Economics: Theoretical and Applied Analysis. Cambridge University Press, Cambridge, pp. 33–54.

Helpman, Elhanan, Krugman, Paul R., 1985. Market Structure and International Trade. MIT Press, Cambridge MA.

Henderson, J. Vernon, Storeygard, Adam, Weil, David N., 2012. Measuring economic growth from outer space. American Economic Review 102 (2), 994–1028.

Henderson, J. Vernon, Wang, Hyoung Gun, 2005. Aspects of the rural-urban transformation of countries. Journal of Economic Geography 5 (1), 23–42.

Herbst, Jeffrey, 2000. States and Power in Africa. Princeton University Press, Princeton.

Hering, Laura, Poncet, Sandra, 2009. The impact of economic geography on wages: disentangling the channels of influence. China Economic Review 20 (1), 1–14.

Hering, Laura, Poncet, Sandra, 2010. Market access impact on individual wages: evidence from China. Review of Economics and Statistics 92 (1), 145–159.

Herrendorf, Berthold, Schmitz Jr., James A., Teixeira, Arilton (2012). The role of transportation in U.S. economic development: 1840–1860. International Economic Review, 53(3), 693–715.

Hill, Hal, Resosudarmo, Budy P., Vidyattama, Yogi, 2008. Indonesia's changing economic geography. Bulletin of Indonesian Economic Studies 44 (3), 407–435.

Hirschman, Albert O., 1958. The Strategy of Economic Development. Yale University Press, New Haven.

Hobsbawm, Eric J., 1962. The Age of Capital. Weidenfeld and Nicolson, London.

Hodler, Roland, Raschky, Paul A., 2010. Foreign Aid and Enlightened Leaders. Monash Discussion Paper No. 54/10.

Holmes, Thomas J., 1998. The effect of state policies on the location of manufacturing: evidence from state borders. Journal of Political Economy 106 (4), 667–705.

Holmes, Thomas J., 2010. Structural, experimentalist, and descriptive approaches to empirical work in regional economics. Journal of Regional Science 50 (1), 5–22.

Holz, Carsten A., 2009. No Razor's Edge: Reexamining Alwyn Young's evidence for increasing interprovincial trade barriers in China. Review of Economics and Statistics 91 (3), 599–616.

Hornbeck, Richard, 2012a. The enduring impact of the American Dust Bowl: short- and long-run adjustments to the environmental catastrophe. American Economic Review 102 (4), 1477–1507.

Hornbeck, Richard, 2012b. Nature versus nurture: the environment's persistent influence through the modernization of American agriculture. American Economic Review: Papers and Proceedings 102 (3), 245–249.

Hsieh, Chang-Tai, Klenow, Peter, 2010. Development accounting. American Economic Journal: Macroeconomics 2 (1), 207–223.

Huang, Rocco R., 2008. Evaluating the real effect of bank branching deregulation: comparing contiguous counties across US state borders. Journal of Financial Economics 87 (3), 678–705.

Isard, Walter, Azis, Iwan J., Drennan, Matthew P., Miller, Ronald E., Saltzman, Sidney, Thorbecke, Erik, 1998. Methods of Interregional and Regional Analysis. Ashgate, Aldershot.

Iyer, Lakshmi, 2010. Direct versus indirect colonial rule in India: long-term consequences. Review of Economics and Statistics 92 (4), 693–713.

Jayaratne, Jith, Strahan, Philip E., 1996. The finance-growth nexus: evidence from bank branch deregulation. Quarterly Journal of Economics 111 (3), 639–670.

Johnson, Paul A., 2000. A nonparametric analysis of income convergence across the US states. Economics Letters 69 (2), 219–223.

Johnson, Paul A., 2005. A continuous state space approach to "Convergence by Parts". Economics Letters 86 (3), 317–321.

Johnson, Paul A., Takeyama, Lisa N., 2001. Initial conditions and economic growth in the US states. European Economic Review 45 (4–6), 919–927.

Judt, Tony, 1996. A Grand Illusion? An Essay on Europe, Hill and Wang.

Kaldor, Nicholas, 1970. The case for regional policies. Scottish Journal of Political Economy 17, 337–348. (Reprinted in Targetti, F., Thirlwall, A.P. (Eds.), 1989. The Essential Kaldor. Duckworth, London).

Kanbur, Ravi, Rapoport, Hillel, 2005. Migration selectivity and the evolution of spatial inequality. Journal of Economic Geography 5 (1), 43–57.

Kanbur, Ravi, Venables, Anthony J., 2005. Spatial Inequality and Development. UNU-WIDER and Oxford University Press, Oxford.

Kaplan, Robert D., 2012. The Revenge of Geography. Random House, New York.

Kelejian, Harry H., Prucha, Ingmar R., 2007. HAC estimation in a spatial framework. Journal of Econometrics 140 (1), 131–154.

Klepper, Steven, 2010. The origin and growth of industry clusters: the making of Silicon Valley and Detroit. Journal of Urban Economics 67 (1), 15–32.

Klepper, Steven, 2011. Nano-economics, spinoffs, and the wealth of regions. Small Business Economics 37 (2), 141–154.

Kline, Patrick, Moretti, Enrico, 2013. Place based policies with unemployment. American Economic Review 103 (3), 238–243.

Knight, John B., Gunatilaka, Ramani, 2011. Does economic growth raise happiness in China? Oxford Development Studies 39 (1), 1–24.

Kolko, Jed, Neumark, David, 2010. Does local business ownership insulate cities from economic shocks? Journal of Urban Economics 67 (1), 103–115.

Kongsamut, Piyabha, Rebelo, Sergio, Xie, Danyang, 2001. Beyond balanced growth. Review of Economic Studies 68 (4), 869–882.

Kremer, Michael, Onatski, Alexei, Stock, James, 2001. Searching for prosperity. Carnegie-Rochester Conference Series on Public Policy 55 (1), 275–303.

Krugman, Paul, 1991. Increasing returns and economic geography. Journal of Political Economy 99 (3), 483–499.

Krugman, Paul, 1995. Development, Geography, and Economic Theory. MIT Press, Cambridge, MA.

Krugman, Paul R., Venables, Anthony J., 1995. Globalization and the inequality of nations. Quarterly Journal of Economics 110 (4), 857–880.

Krusell, Per, Smith, Anthony A., 2009. Macroeconomics and Global Climate Change: Transition for a Many-Region Economy. Manuscript, IIES, Stockholm.

Kwiatkowski, Denis, Phillips, Peter C.B., Schmidt, Peter, Shin, Yongcheol, 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? Journal of Econometrics 54 (1–3), 159–178.

Lagakos, David, Waugh, Michael, 2013. Selection, agriculture, and cross-country productivity differences. American Economic Review 103 (2), 948–980.

Le Pen, Yannick, 2011. A pair-wise approach to output convergence between European regions. Economic Modelling 28 (3), 955–964.

LeSage, James P., Fischer, Manfred, 2008. Spatial growth regressions: model specification. estimation and interpretation. Spatial Economic Analysis 3 (3), 275–304.

Lessmann, Christian, 2011. Spatial Inequality and Development—Is There an Inverted-U Relationship? CESifo Working Paper No. 3622.

Lessmann, Christian, 2013. Foreign direct investment and regional inequality: a panel data analysis. China Economic Review 24, 129–149.

Lipscomb, Molly, Mobarak, A. Mushfiq, Barham, Tania, 2013. Development effects of electrification: evidence from the topographic placement of hydropower plants in Brazil. American Economic Journal: Applied Economics 5 (2), 200–231.

Lipton, Michael, 1980. Migration from rural areas of poor countries: the impact on rural productivity and income distribution. World Development 8 (1), 1–24.

Lucas, Robert E. Jr., 1978. On the size distribution of business firms. Bell Journal of Economics 9 (2), 508–523.

Lucas, Robert E. Jr., Rossi-Hansberg, Esteban, 2002. On the internal structure of cities. Econometrica 70 (4), 1445–1476.

Magrini, Stefano, 2004. Regional (di)convergence. In: Henderson, J. Vernon, Thisse, Jacques-François (Eds.), Handbook of Regional and Urban Economics, vol. 4. North-Holland, Amsterdam, pp. 2741–2796.

Mankiw, N.Gregory, Romer, David, Weil, David N., 1992. A contribution to the empirics of economic growth. Quarterly Journal of Economics 107 (2), 407–437.

Marshall, Alfred, 1920. Principles of Economics, eighth ed. Macmillan, London.

Martin, Philippe, Ottaviano, Gianmarco I.P., 1999. Growing locations: industry location in a model of endogenous growth. European Economic Review 43 (2), 281–302.

Martin, Philippe, Ottaviano, Gianmarco I.P., 2001. Growth and agglomeration. International Economic Review 42 (4), 947–968.

Martin, Ron, Sunley, Peter, 2006. Path dependence and regional economic evolution. Journal of Economic Geography 6 (4), 395–437.

Melitz, Marc, 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. Econometrica 71 (6), 1695–1725.

Melitz, Marc, Ottaviano, Gianmarco I.P., 2008. Market size, trade, and productivity. Review of Economic Studies 75 (1), 295–316.

Melitz, Marc, Redding, Stephen, 2013. Firm Heterogeneity and Aggregate Welfare. NBER Working Paper No. 18919.

Mello, Marcelo, 2011. Stochastic convergence across US states. Macroeconomic Dynamics 15 (2), 160–183.

Michaels, Guy, 2008. The effect of trade on the demand for skill: evidence from the interstate highway system. Review of Economics and Statistics 90 (4), 683–701.

Michaels, Guy, Rauch, Ferdinand, Redding, Stephen J., 2012. Urbanization and structural transformation. Quarterly Journal of Economics 127 (2), 535–586.

Michalopoulos, Stelios, Papaioannou, Elias, 2013. Pre-colonial ethnic institutions and contemporary African development. Econometrica 81 (1), 113–152.

Miguel, Edward, Roland, Gérard, 2011. The long-run impact of bombing Vietnam. Journal of Development Economics 96 (1), 1–15.

Milanovic, Branko, 2005a. Worlds Apart. Princeton University Press, Princeton.

Milanovic, Branko, 2005b. Half a world: regional inequality in five great federations. Journal of the Asia Pacific Economy 10 (4), 408–445.

Minerva, G. Alfredo and Ottaviano, Gianmarco I.P. (2009). Endogenous growth theories: agglomeration benefits and transportation costs. In: Capello, Roberta, Nijkamp, Peter (Eds.), Handbook of Regional Growth and Development Theories. Edward Elgar, Cheltenham.

Mion, Giordano, 2004. Spatial externalities and empirical analysis: the case of Italy. Journal of Urban Economics 56 (1), 97–118.

Mitchener, Kris James, McLean, Ian W., 1999. U.S. regional growth and convergence, 1880–1980. Journal of Economic History 59 (4), 1016–1042.

Mitchener, Kris James, McLean, Ian W., 2003. The productivity of US states since 1880. Journal of Economic Growth 8 (1), 73–114.

Mitton, Todd, 2013. The Wealth of Subnations: Geography, Institutions and Within-Country Development. Manuscript, Brigham Young University.

Moene, Karl Ove, 1988. A reformulation of the Harris-Todaro mechanism with endogenous wages. Economics Letters 27 (4), 387–390.

Morelli, Massimo, Rohner, Dominic, 2010. Natural Resource Distribution and Multiple Forms of Civil War. University of Zurich Working Paper No. 498.

Moretti, Enrico, 2004. Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data. Journal of Econometrics 121 (1–2), 175–212.

Moretti, Enrico, 2010. Local multipliers. American Economic Review: Papers and Proceedings 100 (2), 373–377.

Moretti, Enrico, 2012. The New Geography of Jobs. Houghton Mifflin Harcourt, New York.

Murata, Yasusada, 2008. Engel's law, Petty's law and agglomeration. Journal of Development Economics 87 (1), 161–177.

Myrdal, Gunnar, 1957. Economic Theory and Under-Developed Regions. Duckworth, London.

Naritomi, Joana, Soares, Rodrigo R., Assunção, Juliano J., 2012. Institutional development and colonial heritage within Brazil. Journal of Economic History 72 (2), 393–422.

Neary, J. Peter, 2001. Of hype and hyperbolas: introducing the new economic geography. Journal of Economic Literature 39 (2), 536–561.

Ngai, L.Rachel, Pissarides, Christopher A., 2007. Structural change in a multisector model of growth. American Economic Review 97 (1), 429–443.

Ng, Serena, 2008. A simple test for nonstationarity in mixed panels. Journal of Business and Economic Statistics 26 (1), 113–127.

Nguyen, Binh T., Albrecht, James W., Vroman, Susan B., Westbrook, M.Daniel, 2007. A quantile regression decomposition of urban-rural inequality in Vietnam. Journal of Development Economics 83 (2), 466–490.

Nocco, Antonella, 2005. The rise and fall of regional inequalities with technological differences and knowledge spillovers. Regional Science and Urban Economics 35 (5), 542–569.

Nocke, Volker, 2006. A gap for me: Entrepreneurs and entry. Journal of the European Economic Association 4 (5), 929–956.

Nordhaus, William, 2006. Geography and macroeconomics: new data and new findings. Proceedings of the National Academy of Sciences 103 (10), 3510–3517.

Nunn, Nathan, 2007. Relationship-specificity, incomplete contracts and the pattern of trade. Quarterly Journal of Economics 122 (2), 569–600.

Nunn, Nathan, Puga, Diego, 2012. Ruggedness: the blessing of bad geography in Africa. Review of Economics and Statistics 94 (1), 20–36.

Okubo, Toshihiro, Picard, Pierre, Thisse, Jacques-François, 2010. The spatial selection of heterogeneous firms. Journal of International Economics 82 (2), 230–237.

Olivetti, Claudia, Paserman, M. Daniele, 2013. In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850–1930. CEPR Discussion Paper No. 9372.

Oswald, Andrew J., Wu, Stephen, 2010. Objective confirmation of subjective measures of human well-being: evidence from the U.S.A. Science 327 (5965), 576–579.

Oswald, Andrew J., Wu, Stephen, 2011. Well-being across America. Review of Economics and Statistics 93 (4), 1118–1134.

Ottaviano, Gianmarco I.P., 2012. Agglomeration, trade and selection. Regional Science and Urban Economics 42 (6), 905–1068.

Ottaviano, Gianmarco I.P., Thisse, Jacques-François, 2001. On economic geography in economic theory: increasing returns and pecuniary externalities. Journal of Economic Geography 1 (2), 153–179.

Ottaviano, Gianmarco I.P., Thisse, Jacques-François, 2002. Integration, agglomeration and the political economics of factor mobility. Journal of Public Economics 83 (3), 429–456.

Ottaviano, Gianmarco I.P., Thisse, Jacques-François, 2004. Agglomeration and economic geography. In: Henderson, J. Vernon, Thisse, Jacques-François (Eds.), Handbook of Regional and Urban Economics, vol. 4. North-Holland, Amsterdam.

Ottaviano, Gianmarco I.P., Thisse, Jacques-François, 2005. New economic geography: what about the N? Environment and Planning A 37 (10), 1707–1725.

Ottaviano, Gianmarco I.P., Tabuchi, Takatoshi, Thisse, Jacques-François, 2002. Agglomeration and trade revisited. International Economic Review 43 (2), 409–435.

Overman, Henry G., 2010. "Gis a job": what use geographical information systems in spatial economics? Journal of Regional Science 50 (1), 165–180.

Overman, Henry G., Puga, Diego, 2002. Regional unemployment clusters. Economic Policy 34, 115–143.

Overman, Henry G., Rice, Patricia, Venables, Anthony J., 2010. Economic linkages across space. Regional Studies 44 (1), 17–33.

Pesaran, M., Hashem, 2007. A pair-wise approach to testing for output and growth convergence. Journal of Econometrics 138 (1), 312-355.

Pesaran, M.Hashem, Shin, Yongcheol, Smith, Ron P., 1999. Pooled mean group estimation of dynamic heterogeneous panels. Journal of the American Statistical Association 94 (446), 621–634.

Pesaran, M.Hashem, Smith, Ron, 1995. Estimating long-run relationships from dynamic heterogeneous panels. Journal of Econometrics 68 (1), 79–113.

Phan, Diep, Coxhead, Ian, 2010. Inter-provincial migration and inequality during Vietnam's transition. Journal of Development Economics 91 (1), 100–112.

Picard, Pierre M., Okubo, Toshihiro, 2012. Firms' locations under demand heterogeneity. Regional Science and Urban Economics 42 (6), 961–974.

Pittau, Maria Grazia, 2005. Fitting regional income distributions in the European Union. Oxford Bulletin of Economics and Statistics 67 (2), 135–161.

Pittau, Maria Grazia, Zelli, Roberto, 2006. Empirical evidence of income dynamics across EU regions. Journal of Applied Econometrics 21 (5), 605–628.

Pittau, Maria Grazia, Zelli, Roberto, Gelman, Andrew, 2010. Economic disparities and life satisfaction in European regions. Social Indicators Research 96 (2), 339–361.

Prager, Jean-Claude, Thisse, Jacques-François, 2012. Economic Geography and the Unequal Development of Regions. Routledge, London.

Puga, Diego, 1999. The rise and fall of regional inequalities. European Economic Review 43 (2), 303–334.

Putnam, Robert D. et al., 1993. Making Democracy Work: Civic Traditions in Modern Italy. Princeton University Press.

Pyke, Frank, Becattini, Bruno, Sengenberger, Werner, 1990. Industrial Districts and Inter-firm Cooperation in Italy. International Institute for Labour Studies, Geneva.

Quah, Danny, 1993. Empirical cross-section dynamics in economic growth. European Economic Review 37 (2–3), 426–434.

Quah, Danny T., 1996. Regional convergence clusters across Europe. European Economic Review 40 (3–5), 951–958.

Quah, Danny T., 1997. Empirics for growth and distribution: stratification, polarization, and convergence clubs. Journal of Economic Growth 2 (1), 27–59.

Rappaport, Jordan, 2007. Moving to nice weather. Regional Science and Urban Economics 37 (3), 375–398.

Rappaport, Jordan, Sachs, Jeffrey, 2003. The United States as a coastal nation. Journal of Economic Growth 8 (1), 5–46.

Redding, Stephen, 2010. The empirics of New Economic Geography. Journal of Regional Science 50 (1), 297–311.

Redding, Stephen J., 2012. Goods Trade, Factor Mobility and Welfare. NBER Working Paper No. 18008.

Redding, Stephen, Schott, Peter, 2003. Distance, skill deepening and development: will peripheral countries ever get rich? Journal of Development Economics 72 (2), 515–541.

Redding, Stephen, Sturm, Daniel, 2008. The costs of remoteness: evidence from German division and reunification. American Economic Review 98 (5), 1766–1797.

Redding, Stephen, Venables, Anthony, 2004. Economic geography and international inequality. Journal of International Economics 62 (1), 53–82.

Rice, Patricia, Venables, Anthony J., 2003. Equilibrium regional disparities: theory and British evidence. Regional Studies 37 (6), 675–686.

Rice, Patricia, Venables, Anthony J., Pattachini, Eleonora, 2006. Spatial determinants of productivity: analysis for the regions of Great Britain. Regional Science and Urban Economics 36 (6), 727–752.

Roback, Jennifer, 1982. Wages, rents, and the quality of life. Journal of Political Economy 90 (6), 1257–1278.

Rosenthal, Stuart S., Strange, William C., 2004. Evidence on the nature and sources of agglomeration economies. In: Henderson, J. Vernon, Thisse, Jacques-François (Eds.), Handbook of Regional and Urban Economics, vol. 4. North-Holland, Amsterdam, pp. 2119–2171.

Rossi-Hansberg, Esteban, 2005. A spatial theory of trade. American Economic Review 95 (5), 1464–1491.

Rossi-Hansberg, Esteban, Wright, Mark, 2007. Urban structure and growth. Review of Economic Studies 74 (2), 597–624.

Rud, Juan Pablo, 2012. Electricity provision and industrial development: evidence from India. Journal of Development Economics 97 (2), 352–367.

Sachs, Jeffrey D., Malaney, Pia, 2002. The economic and social burden of malaria. Nature 415 (6872), 680–685.

Sala-i-Martin, Xavier, 1996. Regional cohesion: evidence and theories of regional growth and convergence. European Economic Review 40 (6), 1325–1352.

Sarafidis, Vasilis, Wansbeek, Tom, 2012. Cross-sectional dependence in panel data analysis. Econometric Reviews 31 (5), 483–531.

Satchi, Mathan, Temple, Jonathan R.W., 2009. Labor markets and productivity in developing countries. Review of Economic Dynamics 12 (1), 183–204.

Scoppa, Vincenzo, 2007. Quality of human and physical capital and technological gaps across Italian regions. Regional Studies 41 (5), 585–599.

Scotchmer, Suzanne, Thisse, Jacques-François, 1992. Space and competition: a puzzle. Annals of Regional Science 26 (3), 269–286.

Scott, James, 2009. The Art of Not Being Governed: An Anarchist History of Upland Southeast Asia. Yale University Press, New Haven.

Sen, Amartya, 1980. Equality of what? In: McMurrin, S. (Ed.), Tanner Lectures on Human Values, vol. 1. Cambridge University Press, Cambridge.

Sennett, Richard, 1998. The Corrosion of Character. Norton, New York.

Serra, Maria Isabel, Fernanda Pazmino, Maria, Lindow, Genevieve, Sutton, Bennett, Ramirez, Gustavo, 2006. Regional Convergence in Latin America. IMF Working Paper No. 06/125.

Seya, Hajime, Tsutsumi, Morito, Yamagata, Yoshiki, 2012. Income convergence in Japan: a Bayesian spatial Durbin model approach. Economic Modelling 29 (1), 60–71.

Shioji, Etsuro, 2001. Public capital and economic growth: a convergence approach. Journal of Economic Growth 6 (3), 205–227.

Skoufias, Emmanuel, Katayama, Roy S., 2011. Sources of welfare disparities between and within regions of Brazil: evidence from the 2002–2003 household budget survey (POF). Journal of Economic Geography 11 (5), 897–918.

Solnit, Rebecca, 2013. Diary. London Review of Books 35 (3), 34–35.

Starrett, David, 1978. Market allocations of location choice in a model with free mobility. Journal of Economic Theory 17 (1), 21–37.

Tabellini, Guido, 2010. Culture and institutions: economic development in the regions of Europe. Journal of the European Economic Association 8 (4), 677–716.

Tamura, Robert, 2012. Development Accounting and Convergence for US States. Manuscript, Clemson University.

Temple, Jonathan, 1999. The new growth evidence. Journal of Economic Literature 37 (1), 112–156.

Temple, Jonathan R.W., 2005. Dual economy models: a primer for growth economists. The Manchester School 73 (4), 435–478.

Temple, Jonathan, Woessmann, Ludger, 2006. Dualism and cross-country growth regressions. Journal of Economic Growth 11 (3), 187–228.

Trivedi, Kamakshya, 2006. Educational human capital and levels of income: evidence from states in India, 1965–92. Journal of Development Studies 42 (8), 1350–1378.

Turner, Chad, Tamura, Robert, Mulholland, Sean E., Baier, Scott, 2007. Education and income of the states of the United States: 1840–2000. Journal of Economic Growth 12 (2), 101–158.

Van Nieuwerburgh, Stijn, Weill, Pierre-Olivier, 2010. Why has house price dispersion gone up? Review of Economic Studies 77 (4), 1567–1606.

Venables, Anthony J., 1996. Equilibrium locations of vertically linked industries. International Economic Review 37 (2), 341–359.

Venables, Anthony J., 2005. Spatial disparities in developing countries: cities, regions, and international trade. Journal of Economic Geography 5 (1), 3–21.

Ventura, Jaume, 1997. Growth and interdependence. Quarterly Journal of Economics 112 (1), 57–84.

Walz, Uwe, 1996. Transport costs, intermediate goods, and localized growth. Regional Science and Urban Economics 26 (6), 671–695.

Warner, Andrew (2002). Institutions, Geography, Regions, Countries and the Mobility Bias. CID Working Paper No. 91, Harvard.

Weeks, Melvyn, Yao, James Yudong, 2003. Provincial conditional income convergence in China, 1953–1997: a panel data approach. Econometric Reviews 22 (1), 59–77.

Wei, Kailei, Yao, Shujie, Liu, Aying, 2009. Foreign direct investment and regional inequality in China. Review of Development Economics 13 (4), 778–791.

Williamson, Jeffrey G., 1965. Regional inequality and the process of national development: a description of the patterns. Economic Development and Cultural Change 13 (4), 1–84.

Williamson, Jeffrey G., 1974. Late Nineteenth-Century American Development: A General Equilibrium History. Cambridge University Press, Cambridge.

Williamson, Jeffrey G., 2006. Globalization and the Poor Periphery Before 1950. MIT Press, Cambridge, MA.

Yang, Dennis Tao, Zhu, Xiaodong, 2013. Modernization of agriculture and long-term growth. Journal of Monetary Economics 60 (3), 367–382.

Yoon, Chamna, 2013. The Decline of the Rust Belt: A Dynamic Spatial Equilibrium Analysis. Manuscript, University of Pennsylvania.

Young, Alwyn, 1991. Learning by doing and the dynamic effects of international trade. Quarterly Journal of Economics 106 (2), 369–405.

Young, Alwyn, 2000. The razor's edge: distortions and incremental reform in the People's Republic of China. Quarterly Journal of Economics 115 (4), 1091–1135.

Young, Alwyn, 2013. Inequality, the Urban-Rural Gap and Migration. Manuscript, LSE.

Zhang, Xiaobo, Kanbur, Ravi, 2001. What difference do polarisation measures make? An application to China. Journal of Development Studies 37 (3), 85–98.

CHAPTER FIVE

The Growth of Cities

Gilles Duranton[*,‡] and Diego Puga[†,‡]

[*]Wharton School, University of Pennsylvania, 3620 Locust Walk, Philadelphia, PA, 19104, USA
[†]Centro de Estudios Monetarios y Financieros (CEMFI), Casado del Alisal 5, 28014 Madrid, Spain
[‡]Centre for Economic Policy Research, London

Abstract

Why do cities grow in population, surface area, and income per person? Which cities grow faster and why? To these questions, the urban growth literature has offered a variety of answers. Within an integrated framework, this chapter reviews key theories with implications for urban growth. It then relates these theories to empirical evidence on the main drivers of city growth, drawn primarily from the United States and other developed countries. Consistent with the monocentric city model, fewer roads and restrictions on housing supply hinder urban growth. The fact that housing is durable also has important effects on the evolution of cities. In recent decades, cities with better amenities have grown faster. Agglomeration economies and human capital are also important drivers of city growth. Although more human capital, smaller firms, and a greater diversity in production foster urban growth, the exact channels through which those effects percolate are not clearly identified. Finally, shocks also determine the fate of cities. Structural changes affecting the broader economy have left a big footprint on the urban landscape. Small city-specific shocks also appear to matter, consistent with the recent wave of random growth models.

Keywords

Urban growth, Agglomeration economies, Land use, Transportation, Amenities

JEL Classification Codes

C52, R12, D24

5.1. INTRODUCTION

In 2010, the mean population size of the 366 US metropolitan areas was 707,000, with a range from 18.3 million to just over 50,000. Between 2000 and 2010, these cities grew on average by 10.7%. The first decade of the 21st century was not exceptional for US urban growth. US metropolitan areas grew on average by 17.9% per decade since 1920, the earliest year for which consistent data is available.[1] This figure of 17.9% exceeds

[1] The computations for the United States are based on the 2009 definition of metropolitan areas. Using the earliest definition of metropolitan areas that can be applied to county population data, the 1950 Standard Metropolitan Statistical Areas, we observe a mean growth of 7.3% between 2000 and 2010 and

ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53540-5.00005-7

aggregate population growth by 5.3% points even though a growing population could be accommodated in more cities instead of larger cities. When it comes to urban growth, the United States is not an exceptional country. In Spain, urban areas grew on average by 17.5% between 2000 and 2010, and by 18.1% per decade on average between 1920 and 2010, exceeding aggregate population growth in Spain by 9.2% points. In France, metropolitan areas grew on average by 4% between 1999 and 2007, and by 7.7% per decade on average between 1936 and 2007, exceeding aggregate population growth in France by 2% points.

Although cities tend to grow over time, they do not grow uniformly at the same rate. The standard deviation of the growth rate of US metropolitan areas between 2000 and 2010 is slightly larger than its corresponding mean. Observing individual city growth rates over a decade with means and standard deviations of about the same magnitude is typical. This is the case for the 1920–2010 period in the United States, in Spain, and in France. These figures about the mean and standard deviation of the growth rates of cities naturally lead to asking why cities keep growing even after countries are already highly urbanized, and why some cities grow faster than others.

Being able to answer these questions is important for at least three reasons. The first is that the population growth of cities is economically important in itself. Extremely large investments in building new housing and infrastructure must be made to accommodate the demographic growth of cities. For instance, American households spend about a third of their income on housing, according to the Consumer Expenditure Survey. For their part, various levels of the US government spend more than $200 billion every year to maintain and expand the road infrastructure. Given that most of these investments are extremely durable, it is important to plan them properly and, for this, we need to understand why and how cities grow.

Second, urban economics has proposed a number of theories to explain the population size of cities. Following Alonso (1964), Mills (1967), and Muth (1969), a large literature has focused on the importance of location within the city and its impact on commuting costs as a key determinant of land use and housing development in cities. In turn, the ease of commuting, the availability of housing, and earnings determine the population size of cities. Following Rosen (1979) and Roback (1982), urban economists have also paid great attention to the role of amenities in attracting people to cities. Recognizing that earnings and productivity are themselves systematically related to the population size of cities, much work has been devoted to modeling the productive advantages of cities or

15.8% by decade on average between 1920 and 2010. These lower figures probably understate the true population growth of US cities which, to some extent, grew through the expansion of their suburban areas that were not taken into account by the 1950 definition. On the other hand, the figures based on 2009 definitions probably overstate the true growth of US cities since they partly reflect the selection of the fastest growing cities that became the largest and form the existing set of metropolitan areas.

agglomeration economies explicitly (e.g. Fujita, 1988; Helsley and Strange, 1990; Glaeser, 1999; Duranton and Puga, 2001). The trade-off between agglomeration economies and urban costs, at the core of systems of cities models building on Henderson (1974), is widely accepted as the key explanation behind the existence of cities and provides some important implications for their population growth. Finally, the existence of some regularities in the size distribution of cities and in the patterns of urban growth has motivated alternative approaches which emphasize the importance of random shocks in urban growth (e.g. Gabaix, 1999a).

These theories offer useful guidance to conduct empirical work on urban growth by providing us with specifications and by highlighting a number of identification pitfalls. Conversely, an evaluation of the key drivers of urban growth is also an evaluation of the predictions of the core approaches to the economics of cities.

A third reason to study urban growth is that cities offer an interesting window through which to study the process of economic growth. How cities grow and why may hold important lessons for how and why economies grow. Existing theories of economic growth emphasize the importance of direct interactions. Such interactions often involve direct physical proximity between individuals and are thus naturally studied within cities. Taking the advice of Lucas (1988) seriously, it may be in cities that economic growth is best studied.

We also note that the population growth of cities may be easier and simpler to study than the process of growth of entire countries. The large cross-country growth literature which builds on Barro's (1991) work is afflicted by fundamental data and country heterogeneity problems that are much less important in the context of cities within a country. Furthermore, cross-country growth regressions are plagued by endogeneity problems that are often extremely hard to deal with in a cross-country setting (Durlauf et al. 2005). As we show in this review, looking at cross-sections of cities within countries offers more hope of finding solutions to these identification problems.

To finish this introduction, we would like to delineate more precisely what this chapter does and what it does not do. First, we focus mostly on cities in developed economies. Most of the empirical evidence we discuss below originates from there, the United States in particular. Consistent with this, the theories we discuss consider implicitly mature cities between which workers move. Rural-urban migrations, urbanization, and the role of cities in developing countries are not examined here. We refer the reader instead to Henderson (2005) in a previous volume of this handbook. Second, we discuss and attempt to unify work that has taken place within one discipline–economics. We are aware that other social scientists in geography, planning, or sociology, have taken an interest in urban growth. We leave the bigger task of integrating cross-disciplinary perspectives to others (see Storper and Scott, 2009 for references and one such attempt).

5.2. LAND USE AND TRANSPORTATION

Urban scholars have long recognized that transportation costs are a fundamental determinant of both the population size of cities and their patterns of land use.[2] To understand more precisely the articulation between transportation, land use, and city population, we start with a simple monocentric urban model in the spirit of Alonso (1964), Mills (1967), and Muth (1969).[3] We then use the predictions of this model to structure our examination of the empirical literature on cities and transportation. In subsequent sections, we also enrich this model to account for other features such as amenities and agglomeration economies.

5.2.1 The Monocentric City Model

Consider a linear monocentric city. Land covered by the city is endogenously determined and can be represented by a segment on the positive real line. Production and consumption of a numéraire good take place at a single point $x = 0$, the Central Business District (CBD). Preferences can be represented by a utility function $U(A, u(h, z))$ written in terms of the common amenity level enjoyed by everyone in the city, A, and a sub-utility $u(h, z)$ derived from individual consumption of housing, h, and of the numéraire, z. Commuting costs increase linearly with distance to the CBD, so that a worker living at distance x incurs a commuting cost τx. This leaves $w - \tau x$ for expenditure on housing and the numéraire.[4] Denoting by $P(x)$ the rental price of housing at a distance x from the CBD, we can use a dual representation of the sub-utility derived from housing and the numéraire, and represent preferences with:

$$U(A, v(P(x), w - \tau x)), \tag{5.1}$$

where $\frac{\partial U}{\partial A} > 0$, $\frac{\partial U}{\partial v} > 0$, $\frac{\partial v}{\partial P(x)} < 0$, and $\frac{\partial v}{\partial (w-\tau x)} > 0$.

All residents in the city are identical in income and preferences, enjoy a common amenity level, and are freely mobile within the city. At the residential equilibrium, residents

[2] For early cities, urban historians insist on the difficulty of supplying their residents with food. See for instance Duby (1981–1983), De Vries (1984), or Bairoch (1988). For modern cities, the same scholars point at the cost of moving residents within cities as the being key impediment on urban growth. On that they agree with observers of contemporary cities such as Glaeser and Kahn (2004) who often mention the automobile as the single most important driver of urban change. LeRoy and Sonstelie (1983) and Glaeser et al. (2008), among others, argue that the transportation technologies and their relative costs are also a major driver of where rich and poor residents live within cities. We do not address this last set of issues here.

[3] These models derive from a common ancestor, Thünen's (1826) *Isolated State*, who applied a similar logic to understand the spatial organization of crops in large farms. A detailed presentation of the monocentric model can be found in Fujita (1989).

[4] We generalize this specification below in several ways, including allowing commuting costs to be non-linear and endogenizing wages.

must derive the same sub-utility from housing consumption and the numéraire:

$$v(P(x), w - \tau x) = \overline{v}. \tag{5.2}$$

Totally differentiating Equation (5.2) with respect to x yields:

$$\frac{\partial v(P(x), w - \tau x)}{\partial P(x)} \frac{\mathrm{d}P(x)}{\mathrm{d}x} - \tau \frac{\partial v(P(x), w - \tau x)}{\partial (w - \tau x)} = 0, \tag{5.3}$$

which implies:

$$\frac{\mathrm{d}P(x)}{\mathrm{d}x} = - \frac{\tau}{-\frac{\partial v(P(x),w-\tau x)}{\partial P(x)} / \frac{\partial v(P(x),w-\tau x)}{\partial (w-\tau x)}} = - \frac{\tau}{h(x)} < 0, \tag{5.4}$$

where the simplification follows from Roy's identity. Equation (5.4) is often referred to as the Alonso-Muth condition. It states that, at the residential equilibrium, if a resident moves marginally away from the CBD, the cost of her current housing consumption falls just as much as her commuting costs increase. Thus, the price of housing decreases with distance to the CBD. Then, residents react to this lower price by consuming more housing (larger residences) the farther they live from the CBD. To see this, simply differentiate the Hicksian demand for housing with respect to x:

$$\frac{\partial h(P(x), \overline{v})}{\partial x} = \frac{\partial h(P(x), \overline{v})}{\partial P(x)} \frac{\mathrm{d}P(x)}{\mathrm{d}x} \geqslant 0. \tag{5.5}$$

Note, this is a pure substitution effect, since utility is being held constant at $\overline{v}$. This also implies that the price of housing is convex in distance to the CBD; house prices do not need to fall as fast as commuting costs increase with distance to the CBD to keep city residents indifferent, since they enjoy having a larger house.

To supply housing, a perfectly competitive construction industry uses land and capital under constant returns to scale, to produce an amount $f(x)$ of housing floorspace per unit of land at a distance x from the CBD. The rental price of land, denoted $R(x)$, varies across the city. The rental price of capital is constant and exogenously given, so we omit it as an argument of the unit cost function in construction $c(R(x))$. The zero-profit condition for the construction sector can then be written as:

$$P(x) = c(R(x)). \tag{5.6}$$

Totally differentiating Equation (5.6) with respect to x yields:

$$\frac{\mathrm{d}P(x)}{\mathrm{d}x} = \frac{\partial c(R(x))}{\partial R(x)} \frac{\mathrm{d}R(x)}{\mathrm{d}x}, \tag{5.7}$$

which implies:

$$\frac{\mathrm{d}R(x)}{\mathrm{d}x} = \frac{\mathrm{d}P(x)}{\mathrm{d}x} \frac{1}{\frac{\partial c(R(x))}{\partial R(x)}} = \frac{\mathrm{d}P(x)}{\mathrm{d}x} f(x) < 0, \tag{5.8}$$

where the simplification follows from the envelope theorem. Thus, the reduction in the price of housing as one moves away from the CBD gets reflected in a reduction in the price of land. The construction industry then reacts to lower land prices by building with a lower capital to land ratio (fewer stories and larger gardens) further away from the CBD.

Land is built if the rent $R(x)$ it can fetch in residential use is at least as high as the rent $\underline{R}$ it can fetch in the best alternative use (e.g. agriculture). The edge of the city is thus located at a distance $\overline{x}$ from the CBD such that $R(\overline{x}) = \underline{R}$. The physical extent of the city must also be sufficient to hold its population N:

$$N = \int_0^{\overline{x}} d(x)\mathrm{d}x, \tag{5.9}$$

where $d(x)$ denotes population density at a distance x from the CBD. Using Equations (5.4) and (5.8), we can express population density as:

$$d(x) = \frac{f(x)}{h(x)} = \frac{\frac{\mathrm{d}R(x)}{\mathrm{d}x} / \frac{\mathrm{d}P(x)}{\mathrm{d}x}}{-\tau / \frac{\mathrm{d}P(x)}{\mathrm{d}x}} = -\frac{1}{\tau}\frac{\mathrm{d}R(x)}{\mathrm{d}x}. \tag{5.10}$$

Substituting this expression for $d(x)$ into Equation (5.9), solving the integral, and using $R(\overline{x}) = \underline{R}$ yields $N = \frac{R(0)-\underline{R}}{\tau}$. This implies a very simple expression for land rent at the CBD ($x = 0$):

$$R(0) = \underline{R} + \tau N. \tag{5.11}$$

Valuing Equation (5.6) at $x = 0$ and using (5.11), we can write the price of housing at the CBD as $P(0) = c(\underline{R} + \tau N)$. Equation (5.2) holds for any location in the city, so valuing it at an arbitrary x and at $x = 0$, and using the previous expression for $P(0)$ yields:

$$\begin{aligned} v(P(x), w - \tau x) &= \overline{v} = v(P(0), w) \\ &= v(c(\underline{R} + \tau N), w). \end{aligned} \tag{5.12}$$

This can be inverted to solve for house prices $P(x)$ as a function of x, N, w, τ, and $\underline{R}$. That is the "closed city" version of the monocentric city model, which treats population N as a parameter. The "open city" version allows N to be endogenously determined by migration across cities to attain a common utility level $\overline{U}$. If the amenity level A is common to all cities, we only need to consider the sub-utility derived from housing and the numéraire and can write the condition of utility equalization across cities as:

$$v(c(\underline{R} + \tau N), w) = \overline{v}. \tag{5.13}$$

This spatial equilibrium condition can be inverted to solve for N as a function of $\overline{v}, w, \tau$, and $\underline{R}$.[5]

[5] These models also deliver a number of proportionality results between urban aggregates such as total differential land rent and total commuting costs. We do not develop them here given our focus on population size. The interested reader can refer to Arnott and Stiglitz (1981) and Fujita (1989).

Before going any further, it is worth asking whether the monocentric city model provides reasonable guidance for empirical work. The main issue with the monocentric city model is that it imposes a particular geography for employment.[6] Observation suggests that the geography of most cities is far less extreme than postulated by the monocentric city model, where all employment is concentrated in a single location. In 1996, only about 25% employees in US metropolitan areas worked within 5 km of their CBD (Glaeser and Kahn, 2001). There is a tendency for employment to diffuse away from centers and for metropolitan areas to develop secondary centers (Anas et al. 1998; McMillen, 2001). Despite these clear limitations, the monocentric model remains useful for a number of reasons. First, there is strong empirical support for the existence of declining gradients of land and housing prices; population density; and intensity of construction as predicted by the monocentric city model (see McMillen, 2006, for an introduction to the voluminous literature on this topic). In addition, the monocentric city model has comparative static properties relevant for urban growth that carry through to models with a richer spatial structure, including polycentric cities; and to models without an explicit modeling of space within the city. Simple models that capture essential elements of reality, such as the monocentric city model, are useful because they provide a solid base to specify regressions, help us with identification, and facilitate the interpretation of results. However, we must also be careful not to give a narrow structural interpretation to parameters estimated using the monocentric model as motivation.

5.2.2 Commuting Infrastructure and Population Growth

Local transportation improvements are often justified on the basis that they promote city growth. The monocentric city model sustains this claim. Consider a local improvement in transportation that lowers τ in one particular city within a large urban system. It follows immediately from Equation (5.13) that a reduction in commuting costs increases this city's population with unit elasticity.

The intuition for this result is straightforward and is illustrated in Figure 5.1. The figure plots land rent as a function of distance to the CBD before (solid, downward-sloping curve) and after (dashed, downward-sloping curve) a fall in τ. (Ignore for now the dashed lines to the left of the vertical axis, which will be used in Section 5.2.3.) Since any change in this particular city is too small to affect the large urban system, the level of utility of every resident in the city must remain unchanged to satisfy the spatial equilibrium condition (5.13). Someone living at the CBD does not commute to work and is thus, not directly affected by the fall in τ. The spatial equilibrium condition then implies that

[6] Following Fujita and Ogawa (1982) and, more recently, Lucas and Rossi-Hansberg (2002), economists have attempted to endogenize the location of employment in cities. These models deliver very useful insights and, in some cases, plausible narratives about observed changes in urban forms. However, these models are too complex for their comparative statics results to be easily tested except in some specific dimensions like the number of subcenters (McMillen and Smith, 2003).

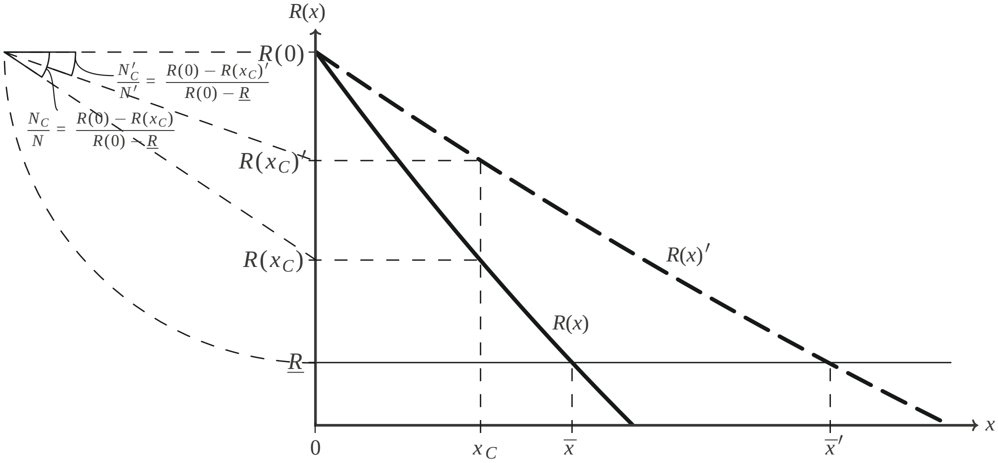

$R(x)$
$R(0)$
$\frac{N'_C}{N'} = \frac{R(0) - R(x_C)'}{R(0) - \underline{R}}$
$\frac{N_C}{N} = \frac{R(0) - R(x_C)}{R(0) - \underline{R}}$
$R(x_C)'$
$R(x_C)$
$R(x)'$
$R(x)$
$\underline{R}$
x
0
x_C
$\bar{x}$
$\bar{x}'$

Figure 5.1 Residential and agricultural land rent against distance to the CBD.

residential land rent at the CBD, $R(0) = \underline{R} + \tau N$, must remain unchanged, which requires population N to increase in the same proportion as τ falls. Everywhere beyond the CBD, residents benefit from the reduction in τ, but land and house prices increase as a result of immigration, offsetting the utility gain from lower commuting costs. The shift in land rents pushes outwards the edge of the city, given by the intersection of $R(x)$ with $\underline{R}$, from $\bar{x}$ to $\bar{x}'$. The larger population is housed through a combination of this increase in the spatial size of the city and rising densities everywhere (as people react to rising house prices by reducing their housing consumption, and the construction industry reacts to rising land prices by building more floorspace per unit of land).

This prediction of a unit elasticity of city population with respect to commuting costs maps directly into the following regression:

$$\Delta_{t+1,t} \log N_i = \beta_0 - \beta_1 \Delta_{t+1,t} \log \tau_i + \epsilon_{it}, \tag{5.14}$$

where i indexes cities, $\Delta_{t+1;t}$ is a time-differencing operator between period t and period $t+1$, β_1 is the elasticity of interest (predicted to be unity); and ϵ_{it} is an error term which, for the time being, we can interpret as a random disturbance.

To begin, we note that testing whether the coefficient β_1 estimated in regression (5.14) differs from unity would be more than a test of the core mechanism of the monocentric city model. It would be a joint test of several assumptions in that model, including the linearity of commuting costs and free labor mobility. Because we do not expect all the conditions leading a unit population elasticity to hold, being able to reject that the estimated value of β_1 is exactly one, is of secondary importance. Instead, we are primarily interested in knowing whether commuting costs affect the population of cities and how important this factor might be both in absolute terms and relatively to other drivers of urban growth. The advice of Leamer and Levinsohn (1995), "estimate, don't test," is particularly relevant here.

Equation (5.14) belongs to a much broader class of regressions where the growth of cities is regressed on a number of explanatory variables. Hence, the estimation issues raised by this regression also occur in most urban growth regressions. We discuss these general issues at length here and avoid repeating them when discussing similar regressions below.

A first key issue is the speed of adjustment. Rather than assume free labor mobility in a static sense, one could think of the equilibrium population N_i^* that satisfies Equation (5.13) as a steady state toward which the city converges. New housing takes time to build and we cannot expect an immediate adjustment of city population after a change in commuting costs. We might instead posit the following myopic adjustment process where $N_{it+1} = N_i^{*\lambda} N_{it}^{1-\lambda}$. The parameter λ can be interpreted as a rate of convergence. We have $\lambda = 0$ if residents cannot change city, and $\lambda = 1$ if they fully adjust between any two periods.[7] Taking logs of this adjustment process equation implies $\Delta_{t+1,t} \log N_i = \lambda(\log N_i^* - \log N_{it})$. The spatial equilibrium condition (5.13) implies that τN_i^* should be constant in steady state, i.e. $\log N_i^* = \beta_0 - \beta_1 \log \tau_{it}$, with β_1 predicted to be unity. Combining these two equations leads to the following regression:

$$\Delta_{t+1,t} \log N_i = \lambda\beta_0 - \lambda \log N_{it} - \lambda\beta_1 \log \tau_{it} + \epsilon_{it}. \tag{5.15}$$

The choice between a "changes-on-changes" regression like (5.14) and a "changes-on-levels" regression like (5.15) matters because these two regressions use very different sources of variation in the data and, as a result, suffer from different identification problems. Ideally, this choice of specification should be driven by informed priors about how population adjusts. An advantage of Equation (5.15) is that the speed of convergence λ is estimated together with the parameter of interest, β_1.

The recognition that transport improvements may take time to affect city growth and that other factors will influence this process creates additional identification concerns. As a first step, we should control for other factors that may simultaneously affect city growth. However, it is not possible to control for all such factors. If there are omitted variables that drive urban growth and are correlated with transportation costs, then ordinary least square (OLS) estimates of the effects of transport costs on city growth will be biased. A second, related, concern is possible reverse causation, where transport infrastructure is assigned on the basis of expected growth. Even in the absence of forward-looking infrastructure assignment, transport costs and future growth can be correlated. This is because we expect any measure of local transport costs to be serially correlated and, as discussed below, there is also persistence in urban growth.

[7] Baldwin (2001) shows how this ad hoc migration specification can be consistent with forward-looking behavior when migration across cities generates congestion frictions.

As made clear by the rest of this chapter, these concerns of correlated omitted variables and reverse causation or endogeneity plague all city growth regressions. This is unsurprising. Regressions like (5.15) strongly resemble cross-country growth regressions in the line of Barro (1991).[8] It is well known that these regressions are afflicted by serious problems of correlated omitted variables and endogeneity (Durlauf et al. 2005). In a cross-country setting, these problems are extremely hard to deal with and solutions are very few. Looking at cities within countries offers more hope regarding identification.[9]

One might be tempted to tackle the problem of correlated omitted variables by building a panel of cities and estimating a regression based on Equations (5.14) or (5.15) with city fixed-effects. However, if transport infrastructure is allocated on the basis of the economic fortunes of cities the correlation between changes in the transport infrastructure and the growth residual (i.e. the error term) in the regression will be much stronger than the correlation between the level of infrastructure and the error. That is, fixed-effect and first-difference estimations can suffer from worse biases than simple cross-sectional estimations.

Duranton and Turner (2012) tackle correlated omitted variables and reverse causation using instrumental variables to estimate a regression that is very close to Equation (5.15).[10] They use as dependent variable the change in log employment between 1983 and 2003 for US metropolitan areas. As a proxy for τ they use lane kilometers of interstate highways within metropolitan areas in 1983 (although interstate highways represent only a small proportion of the roadway, they carry a disproportionate amount of traffic). Duranton and Turner (2012) instrument interstate highways using three historical measures of roads: the 1947 highway plan that was the template for the modern US interstate highway system, a map of 1898 railroads, and a map of old exploration routes of the continent dating back to 1528.

As usual with this type of strategy, it relies on the instruments being relevant, i.e. on their ability to predict roads conditional on the control variables being used. Denoting the instruments Z_i:

$$\mathbb{C}\text{ov}(\log \tau_i, Z_i|.) \neq 0. \tag{5.16}$$

This condition can be formally assessed (see Angrist and Pischke, 2008, for details). In the case of the three instruments used by Duranton and Turner (2012) the relevance condition

[8] In the neoclassical models of growth that underpin cross-country growth regressions, the changes-on-levels specification for the regressions arises from the slow adjustment of capital in the process of convergence toward steady state. In our case, this is driven by the slow adjustment of labor.

[9] Cross-country growth regressions are also afflicted by fundamental problems of data and cross-country heterogeneity which are much less important in the context of metropolitan areas within countries.

[10] Duranton and Turner (2012) derive their specification from a model where, unlike in the monocentric model, relative locations within cities do not matter even though residents have a demand for transportation within the city. As noted above, it is reassuring that better transportation is predicted to be a driver of urban growth in a class of models broader than the monocentric city model.

is satisfied even when using a demanding set of control variables. This is because the 1947 highway map was, by and large, implemented; old railroads, were turned into roads, or highways were built alongside them, and many pathways discovered a long time ago through exploration, are still pathways today.

The validity of the instruments also relies on them being exogenous, i.e. on them being correlated with population growth only through the roadway so that they are orthogonal with the error term:

$$\mathbb{C}\text{ov}(\epsilon_i, Z_i|.) = 0. \tag{5.17}$$

Establishing exogeneity is much harder than establishing relevance. The first step for the defense of any set of instruments is to show that they are not directly linked to the dependent variable. In the case at hand, the 1947 highway planners were interested in linking US cities together but were not concerned with future commuting patterns. Railroad builders in 1898 were interested in shipping grain, cattle, lumber, and passengers across the continent. Early explorers were interested in finding a wide variety of things, from the fountain of youth to pathways to the Pacific. This first step is necessary but not sufficient. The exogeneity condition (5.17) fails when an instrument is correlated with a missing variable that also affects the dependent variable. For instance, cities in more densely populated parts of the country in 1947 received more kilometers of planned highways. Those cities might also have grown less between 1980 and 2000. The second step is thus to use further controls, and in particular population controls, in the instrumental variable (IV) estimation to preclude such correlations with missing variables as much as possible.

Overidentification tests are the next element of any IV strategy. They can be conducted when there are more instruments than (endogenous) parameters to estimate. However, we expect very similar instruments to lead to very similar estimates and thus pass overidentification tests. This should not be taken as a proof of instruments validity. Overidentification tests are more meaningful when the instruments rely on very different sources of variation in the data.

Finally, a difference between OLS and IV estimates can be indicative of an OLS bias. However, with invalid instruments, the bias on the IV estimate could be even worse. Thus, whenever there are significant differences between OLS and IV estimates it is important to provide out-of-sample evidence for the channels through which the OLS bias percolates.

In conclusion, any reasonable IV strategy needs to (i) establish the strength of its instruments, (ii) provide a plausible argument that the instruments are independent from the dependent variable, (iii) preclude alternative indirect channels of correlation between the instruments and the dependent variable, (iv) show that different instruments provide the same answer, and (v) provide out-of-sample evidence explaining differences between OLS and IV estimates. This said, no IV strategy can be entirely fool proof since instrument validity relies on the absence of a correlation with an unobserved term, as shown by Equation (5.17). Despite their limitations, IV strategies are likely to remain an important part of the toolkit for the analysts of the growth of cities. Natural experiments and discontinuities are scarce and the context in which they take place is often very specific.[11]

Turning to the results of Duranton and Turner (2012), they find that a 10% increase in a city's stock of interstate highways in 1983 causes the city's employment to increase by about 1.5% over the course of the following 20 years when using IV, compared with about 0.6% when using OLS. The higher coefficient on the roadway with IV is consistent with the institutional context in which interstate highways are built in the United States. There is a funding formula that equalizes funding per capita and thus gives fewer roads to denser and fast-growing places where land is more expensive. In addition, this formula is not universally applied and many road projects are make-work subsidies for poorly performing places. Duranton and Turner (2012) provide evidence to that effect.

Note also that the estimated 0.15 elasticity of city employment with respect to the roadway is not directly comparable to the unit elasticity of city population with respect to transportation costs. This is because there is no proportional relationship between highways and commuting costs. The chief reason is that more roads beget more traffic, as shown by Duranton and Turner (2011) in a companion paper. As a result, the speed of travel declines only a little when more roads are provided. In turn, this suggests that the proper estimation of Equation (5.15) requires knowing more about the relationship between roads, traffic, and speed of travel. This also calls for a more detailed modeling of the commuting technology. In addition, Duranton and Turner (2012) estimate that the adjustment of population to increased road provision is slow at the metropolitan level.

A more meaningful comparison is with other drivers of city growth. The elasticity of city growth with respect to roads estimated by Duranton and Turner (2012) implies that a one standard deviation of 1983 interstate highways translates into two-thirds of a standard deviation in city growth. This is comparable to the effect of one standard deviation in January temperatures found by Rappaport (2007) in his analysis of population displacement in the United States toward nicer weather. It is also slightly larger than the effect of one standard deviation in the initial stock of university graduates found by Glaeser and Saiz (2004). We discuss the role of amenities and human capital at greater length below.

5.2.3 Commuting Infrastructure and Land Use

We have just seen that, following a decline in unit commuting costs, cities should experience an influx of population. To accommodate this larger population, cities physically expand outwards and experience rising densities. Of these two channels, outwards expansion is more important. To see this, consider any arbitrary point x_C, and think of the

[11] Greenstone et al. (2010) and Holmes (1998) are key examples of the use of, respectively, quasi-experimental evidence and discontinuities in this area of research, although neither focuses on the effect of transport improvements that we discuss in this section.

segment of the city between the CBD and x_C as the historical central city, and the segment between x_C and the city edge $\overline{x}$ as the suburbs. Let $N_C = \int_0^{x_C} d(x)\mathrm{d}x$ denote the (endogenous) population of the central city. Then, using Equations (5.10) and (5.11), we can calculate the share of population in the central city as:

$$\frac{N_C}{N} = \frac{R(0) - R(x_C)}{R(0) - \underline{R}}. \tag{5.18}$$

A reduction in τ increases land rent at any given point beyond the CBD including x_C, but it does not affect land rent $R(0)$ at the CBD (where there is no need to commute and migration keeps utility unchanged) nor land rent at the city edge, which is fixed at $\underline{R}$. Then, Equation (5.18) implies that the share of population in the central city falls when commuting costs are reduced. This reduction in $\frac{N_C}{N}$ is shown graphically in Figure 5.1 to the left of the vertical axis, based on Equation (5.18). This has important implications for the analysis of suburbanization, since it implies that improvements in local transportation foster the suburbanization of population.

The positive relationship between roads and suburbanization implied by the monocentric city model is explored in Baum-Snow's (2007) pioneering work on US cities.[12] His main specification is of the following form:

$$\Delta \log N_{C(i)} = \beta_0 - \beta_1 \Delta\tau_i + \beta_2 x_{C(i)} + \beta_3 \Delta \log N_i + X_i \beta_4' + \epsilon_i, \tag{5.19}$$

where the dependent variable is the change in log central city population between 1950 and 1990. His measure of commuting, $\Delta\tau_i$, is the change in the number of rays of interstate highways that converge toward the central city. The specification controls for the change in log population for the entire metropolitan area $\Delta \log N_i$ and the radius of the central city $x_{C(i)}$.

The key identification challenge is that rays of interstate highways going to the central city may not have caused suburbanization but instead accompanied it. Baum-Snow's (2007) innovative identification strategy relies on using the 1947 map of planned interstate highways. Planned rays of interstate highways are a strong predictor of rays that were actually built. As already argued, the 1947 highway plan was not developed with suburbanization in mind but aimed instead at linking cities between them. Finally, Baum-Snow (2007) also controls for a number variables such as changes in log income or changes in the distribution of income which could drive suburbanization and be associated with the assignment of interstate highways.

[12] Baum-Snow (2007) motivates his specification verbally with a closed-city (i.e. constant population) version of the monocentric city model. With constant population in the city, when a fall in commuting costs flattens the land and house price gradients, each resident consumes more housing and land. This expands the city boundary outwards and (unlike in the open-city version of the model with endogenous population) reduces density. Suburbanization then follows from the relocation of some former central city residents to the suburbs.

The main finding of Baum-Snow (2007) is that an extra ray of interstate highways leads to a decline in central city population of about 9%. This IV estimate is larger than its OLS counterpart, perhaps because more highways were built in cities that suburbanized less. This finding is confirmed when estimating the effect of highways using a panel of shorter first differences and city fixed effects.

More puzzling in light of the monocentric model is the fact that central cities experienced not only a relative decline but also an absolute decline in their population. Over 1950–1990, the population of central cities fell by an average 17% while total metropolitan area population rose by 72%. This evolution could be explained by a concomitant increase in incomes in the United States leading residents to consume more housing. In the monocentric city model, it follows from Equation (5.12) that an increase in the wage w that affects all cities equally leaves their populations unchanged. By Equation (5.11), land rent at the CBD is also unchanged. The land rent at the city edge must still equal the rent in the best alternative use, $\underline{R}$. If housing is a normal good, the economy-wide increase in w then simply makes the house-price gradient flatter. Differentiating Equation (5.4) with respect to w, yields:

$$\frac{\partial^2 P(x)}{\partial w \partial x} = \frac{\tau}{(h(x))^2} \frac{\partial h(x)}{\partial w} > 0. \tag{5.20}$$

Residents each consume more housing and this leads to a reduction in central city population (population in $x \in [0, x_c]$).

Other explanations for the decline of central cities in the United States have focused on a variety of social and material ills that have afflicted central cities such as crime (Cullen and Levitt, 1999), the degradation of the housing stock (Brueckner and Rosenthal, 2009), racial preferences (Boustan, 2010), and related changes in the school system (Baum-Snow and Lutz, 2011).[13]

The suburbanization of population is one of several phenomena that has been associated with urban sprawl. Another key dimension of sprawl is the scatteredness of development, i.e. how much undeveloped land is left between buildings. Burchfield et al. (2006) merge data based on high-altitude photographs from around 1976 with data based on satellite images from 1992 to track development on a grid of 8.7 billion 30 × 30 m cells covering the United States. For each metropolitan area, they compute an index of sprawl measuring the percentage of undeveloped land in the square kilometer surrounding the average residential development. Burchfield et al. (2006) show that US metropolitan areas differ widely in terms of how scattered development is in each one of them, but for most individual areas the scatteredness of development has been very persistent over time.

[13] Existing evidence points at black in-migration followed by white flight and crime as being the two main factors. The race explanation is specific to the United States, and this may explain why it has experienced greater central city decline than other developed countries. These factors are, of course, in addition to the uniquely important role played by the car in the United States.

Among various factors that could potentially affect sprawl, they look at transportation. They find that a denser road network in the suburbs is not associated with more scattered development. At the same time, the car-friendliness of the city center does matter. Cities that were originally built around public transportation (proxied by streetcar passengers per capita ca. 1900) tend to be substantially more compact, even in terms of their recent development, than cities built from the start around the automobile. Other factors that lead to more scattered urban development include ground water availability, temperate climate, rugged terrain, specialization in spatially decentralized sectors, a high-variance over time in decade-to-decade local population growth, having large parts of the suburbs not incorporated into municipalities, and financing a lower fraction of local public services through local taxes.

5.3. HOUSING

Our modeling of housing so far misses two key features that matter enormously in reality: the supply of housing is only imperfectly elastic, and housing is durable. In themselves, these two characteristics do not determine whether a city will grow or decline. They will however determine how cities will react to positive and negative shocks.

To model the effects of imperfectly elastic housing supply, and housing durability, we first need to enrich our model by incorporating an elastic demand for labor that helps determine the wage. We can then study how imperfectly elastic housing supply affects a city's population, wages, and house prices following a labor demand shock. Note that this extension to our model is of independent interest since it also allows us to study the effects of changes in labor demand on city growth. We return to this later in this chapter.

5.3.1 Housing Supply Restrictions

Suppose labor demand in each city depends negatively on wages w and positively on a local productivity shifter B_i, with i used to index cities. With a constant unit labor supply per worker, local labor supply is simply given by the local labor force N_i. We can then characterize the labor market equilibrium by a wage function:

$$w_i = w(B_i, N_i), \tag{5.21}$$

with $\frac{\partial w_i}{\partial B_i} > 0$ and $\frac{\partial w_i}{\partial N_i} < 0$. Consider a positive shock to local productivity in a city, i.e. an exogenous increase in B_i in some city i. In the short run (where we take city workforce, and hence labor supply, to be fixed), such an increase in the demand for labor leads to higher wages since $\frac{\partial w_i}{\partial B_i} > 0$. The long-run consequences, however, depend on land and housing supply, which help determine the evolution of N_i.

In the standard monocentric city model, as developed above, the construction industry can develop as much land as necessary at the price of land $\underline{R}$ determined by its best alternative use. The construction industry can also redevelop already developed areas by increasing or decreasing the density of development.

Then, when a positive productivity shock increases the wage, this makes the city relatively more attractive and causes its population to grow. Substituting Equation (5.21) into the spatial equilibrium condition of Equation (5.13) and applying the implicit function theorem directly implies:

$$\frac{\mathrm{d}N_i}{\mathrm{d}B_i} = -\frac{\frac{\partial v}{\partial w_i}\frac{\partial w_i}{\partial B_i}}{\frac{\partial v}{\partial w_i}\frac{\partial w_i}{\partial N_i} + \frac{\partial v}{\partial P(x)}\frac{\partial c(R(x))}{\partial R(x)}\tau} > 0. \tag{5.22}$$

To sign this derivative, recall from the statement of Equation (5.1) that $\frac{\partial v}{\partial P(x)} < 0$ and $\frac{\partial v}{\partial w_i} > 0$; recall also that the envelope theorem, as used to simplify Equation (5.8), implies that $\frac{\partial c}{\partial R(x)} = \frac{1}{f(x)} > 0$; and we have just stated that $\frac{\partial w_i}{\partial B_i} > 0$ and $\frac{\partial w_i}{\partial N_i} < 0$. This implies $\frac{\mathrm{d}N_i}{\mathrm{d}B_i} > 0$.

To house this larger population, new dwellings must be built, which requires an increase in land and house prices everywhere to make it worthwhile for the construction industry to outbid alternative uses such as agriculture at the expanded urban fringe:

$$\begin{aligned}\frac{\mathrm{d}P(x)}{\mathrm{d}B_i} &= -\frac{\frac{\partial v}{\partial w_i}\left(\frac{\partial w_i}{\partial B_i} + \frac{\partial w_i}{\partial N_i}\frac{\mathrm{d}N_i}{\mathrm{d}B_i}\right)}{\frac{\partial v}{\partial P(x)}} \\ &= \tau\frac{\partial c(R(x))}{\partial R(x)}\frac{\mathrm{d}N_i}{\mathrm{d}B_i} > 0.\end{aligned} \tag{5.23}$$

The first line of Equation (5.23) follows from substituting Equation (5.21) into Equation (5.2) and applying the implicit function theorem. Note that the short-run wage rise resulting from a positive productivity shock $\left(\frac{\partial w_i}{\partial B_i}\right)$ is mitigated by the population growth that it triggers $\left(\frac{\partial w_i}{\partial N_i}\right)$. This also dampens the increase in house prices. However, since population in the city grows following the positive productivity shock, the overall effect on house prices must still be an increase. This is the implication of the second line of Equation (5.23), which is obtained by substituting Equation (5.21) into (5.13), totally differentiating with respect to B_i, and using the resulting equation to substitute the right-hand side expression on the first line of (5.23). Local inhabitants react to higher house prices by choosing to live in smaller dwellings at any given distance from the CBD.[14]

[14] A cross-sectional implication of these comparative statics is that high-productivity cities will tend to be larger in population, have higher density, pay higher wages, and have more expensive houses.

In reality, the supply of land is not completely elastic, as assumed so far. It is limited both by geographical constraints and by land-use regulations.[15] This has important implications for the growth of cities. As a benchmark, consider the case where the supply of land is completely inelastic. For instance, a city could reach a green belt at its edge and the edge of the city would become fixed at $\overline{x}$. The spatial equilibrium condition of Equation (5.13) is then replaced by:

$$v(c(R(\overline{x}) + \tau N), w) = \overline{v}, \tag{5.24}$$

where land rent at the city edge is now strictly greater than the agricultural land rent: $R(\overline{x}) > \underline{R}$. A positive productivity shock still increases the wage and makes the city relatively more attractive. However, with an inelastic land supply the only way to house a larger population is through an increase in density. Compared with the case of an elastic supply of land, the green belt causes land rents to be higher everywhere in the city, from the fixed edge $\overline{x}$ to the CBD, which makes population grow by less. This comparison shows that land-use regulations affect the extent to which a positive shock that makes a city more attractive translates into higher house prices or more population.

The above comparative statics, by showing that the effect of B_i on city growth is mediated by restrictions on the supply of developable land, provide useful guidance for empirical work. They highlight that measures of the stringency and restrictiveness of land-use regulations cannot be used directly as explanatory variables in a city growth regression. Instead, the stringency of land-use regulations should be interacted with predictors of city growth. In cities that are predicted to grow, we expect strong population growth when land-use regulations are lax, and strong wage and housing price growth when they are stringent. In their analysis of US metropolitan areas between 1980 and 2000, Glaeser et al. (2006) use two robust predictors of city growth to demonstrate this process. Since human capital is strongly correlated with city growth during this period, they use the initial share of the local population with a bachelor's degree as their first predictor. The second predictor is an index that exploits the idea that the sectoral composition of cities in an important determinant of the evolution of their labor demand, and sectors expand and contract differently as a result of largely national factors. As first suggested by Bartik (1991), cities with a high share of employment in sectors with high growth nationally are thus expected to grow faster in population. For a given predicted employment growth, Glaeser et al. (2006) show that highly regulated cities experienced lower population growth rate, higher income growth, and higher growth in housing prices.

[15] In theory, cities could use land-use regulations to increase the supply of housing, for instance through densification schemes. In practice, land-use regulations in developed countries and many developing countries are, in most cases, geared toward restricting housing supply through, among others, minimum lot size regulations, maximum building height, green belts, and lengthy and cumbersome approval processes.

This type of analysis raises again an inference problem due to the potential endogeneity of land-use regulation.[16] When trying to explain population growth in cities, we expect land-use regulations to be more stringent in what would otherwise be fast-growing cities because current landowners may lobby hard for stricter regulations when they expect housing prices to appreciate.[17] When trying to explain wage growth in cities, we also expect wages to rise faster in more regulated cities since only households with high wages may be able to afford more expensive houses in these cities. All this casts doubts on the direction of causality in the findings reported above.

While a complete disentangling of the stringency of land-use regulations, population growth, and wage growth in cities has escaped the literature so far, Saiz (2010) offers interesting findings about the exogeneity of land-use regulations.[18] Most importantly, land-use regulations are more stringent in cities where there is less usable land. Usable land is defined as all land not covered by water or with a steep slope. By that measure, cities like San Francisco and Miami have very little usable land whereas cities like Atlanta and Columbus are largely unconstrained in their development. Saiz (2010) shows a strong link between these natural constraints and stringent land-use regulations. This suggests that, ultimately, the limits on city growth imposed by land-use regulations are geographical limits magnified by human interventions.

5.3.2 Housing Durability

The durability of housing has important implications for city growth since people can move out of a city whereas houses cannot.

When a city experiences a positive shock, as we saw in Section 5.3.1, more workers are attracted to it and additional housing is built. On the other hand, when a city experiences a negative shock and some workers leave, existing housing is not destroyed. More specifically, if housing is durable, its supply will be kinked—with a steep slope below its current equilibrium level and a flatter slope above this level. This suggests an interesting

[16] There is also an important issue of how to measure regulation. There are three main approaches. The first is to estimate a wedge between property prices and construction costs, as, among many others, Glaeser et al. (2005). This method has the obvious drawback of doing no more than putting a name on a residual. The second is to precisely document some key regulations at a high level of geographical resolution like Glaeser and Ward (2009). The difficulty of this exercise makes it difficult to go much beyond one metropolitan area which the analyst knows extremely well. The third approach tracks land-use regulations across a wide range of jurisdictions and aggregates them into one aggregate index such as the Wharton Residential Land Use Regulatory Index (Gyourko et al. 2008). The drawbacks here are the possible heterogeneity in the data (e.g. sanctuarized greenbelt in one city may not be the same thing as a slow-moving greenbelt in another) and aggregation biases.

[17] On the other hand, employers may lobby for laxer regulations when they expect their activities to expand.

[18] Given the difficulty of the exercise, a better identification of zoning issues is likely to use good restrictions coming from plausible theories of housing regulations. See Fischel (2000), Ortalo-Magné and Prat (2010), or Hilber and Robert-Nicoud (2013) for recent contributions.

asymmetry between city growth and city decline. When cities grow, they experience moderate house price increases and large population changes. When cities contract, they experience large house price drops and small population changes.

Glaeser and Gyourko (2005) document this asymmetry between urban growth and urban decline using data for 321 US cities for each decade between 1970 and 2000. This asymmetry also holds for the more recent past. In the United States, according to the US census, 17 metropolitan areas, including Las Vegas (NV) and Raleigh (NC), enjoyed a population growth more than 20% points above the mean of 10.7%, between 2000 and 2010. On the other hand, New Orleans was the only city which declined by more than 10% during the same period.[19] Even Youngstown (OH) and Johnstown (PA)—which come just before New Orleans at the bottom of the growth ranking for 2000–2010—did not decline by more than 6% over a decade.

Housing is durable but not permanent. It depreciates slowly over time. This suggests another step to the argument above. After a negative shock, some households leave, and housing prices decline, which induces many to stay. Then, over time, the housing stock depreciates and housing supply declines. Since house prices, that is, the market values of properties, may be well below their construction costs, houses that depreciate are not likely to be refurbished. Households will thus slowly leave the city as the housing stock slowly depreciates. Put differently, housing decline is expected to be persistent. Indeed, urban decline one decade is a strong predictor of urban decline the following decade; whereas city growth one decade is a less strong predictor of city growth for the following decade (Glaeser and Gyourko, 2005).

Glaeser and Gyourko (2005) also argue that those who stay in declining cities because of low housing prices are likely to be those with the lowest labor market opportunities in case of out-migration. They provide evidence that declines in population are associated with declines in human capital in their sample of US cities.

5.4. URBAN AMENITIES

Following the work of Rosen (1979) and Roback (1982), urban economists have paid great attention to the role of amenities in attracting people to cities. If cities differ in terms of their amenity level, the spatial equilibrium condition (5.13) must be taken up one level. Substituting Equation (5.12) into the initial utility function (5.1), and indexing cities by i, we can write this more general version of the spatial equilibrium condition as:

$$U(A_i, v(c(\underline{R} + \tau N_i), w)) = \overline{U}. \tag{5.25}$$

Recall from the presentation of Equation (5.1) that $\frac{\partial U}{\partial A_i} > 0$ and $\frac{\partial U}{\partial v} > 0$. Recall also from the derivation of Equation (5.22) that $\frac{\partial v}{\partial N_i} = \frac{\partial v}{\partial P(x)} \frac{\partial c(R(x))}{\partial R(x)} \tau < 0$. Applying the implicit

[19] Adding to this, the decline of New Orleans was due to an extremely rare weather event that caused the sudden and massive destruction of its housing stock.

function theorem to (5.25) directly implies:

$$\frac{\mathrm{d}N_i}{\mathrm{d}A_i} = -\frac{\frac{\partial U}{\partial A_i}}{\frac{\partial U}{\partial v}\frac{\partial v}{\partial N_i}} > 0. \tag{5.26}$$

This suggests a first obvious channel through which amenities can affect urban growth: cities where amenities improve become relatively more attractive and grow in population, while cities where amenities deteriorate lose population. These changes in the supply of amenities are sometimes the result of local improvements. One can think of the cleaning and rejuvenation of old historical downtowns, particularly in Europe. Other instances of local changes in the supply of amenities are the result of some economy-wide shock that affects cities heterogeneously. For example, the invention of air-conditioning has reduced the disamenity of extremely hot summer weather in cities in the southern United States.

There are two other possibilities which are less well understood but, possibly, at least as relevant empirically. First, some demographic changes might be at play. Cities with nice downtowns may be particularly appealing to childless educated workers in their twenties or early thirties whereas cities with mild winters may be particularly attractive to pensioners. These two groups have grown substantially in size and so have these two types of cities.

Second, aggregate economic growth increases wages. If amenities complement other goods, higher wages lead to an increased appeal of high-amenity cities. In this context, migration to high amenity cities is a consequence of economic growth raising the demand for amenities, not of changes in the supply of amenities.

To understand this argument in greater depth, let us return to the spatial equilibrium condition described by Equation (5.25). Since we are now considering an economy-wide increase in the wage, we cannot treat the common utility level $\overline{U}$ as a constant; instead, it will change equally in all cities. Totally differentiating (5.25) with respect to w yields:

$$\frac{\partial U}{\partial v}\left(\frac{\partial v}{\partial N_i}\frac{\mathrm{d}N_i}{\mathrm{d}w} + \frac{\partial v}{\partial w}\right) = \frac{\mathrm{d}\overline{U}}{\mathrm{d}w}. \tag{5.27}$$

Totally differentiating (5.27) with respect to A_i results in:

$$\frac{\partial^2 U}{\partial A_i \partial v}\left(\frac{\partial v}{\partial N_i}\frac{\mathrm{d}N_i}{\mathrm{d}w} + \frac{\partial v}{\partial w}\right) + \frac{\partial U}{\partial v}\frac{\partial v}{\partial N_i}\frac{\partial^2 N_i}{\partial A_i \partial w} = 0. \tag{5.28}$$

Rearranging implies:

$$\frac{\partial^2 N_i}{\partial A_i \partial w} = -\frac{\partial^2 U}{\partial A_i \partial v}\frac{\frac{\partial v}{\partial N_i}\frac{\mathrm{d}N_i}{\mathrm{d}w} + \frac{\partial v}{\partial w_i}}{\frac{\partial U}{\partial v}\frac{\partial v}{\partial N_i}}. \tag{5.29}$$

To sign this expression, note that the numerator of the fraction on the right, $\frac{\partial v}{\partial N_i}\frac{\mathrm{d}N_i}{\mathrm{d}w} + \frac{\partial v}{\partial w} = \frac{\mathrm{d}v}{\mathrm{d}w}$, must be positive. This is because when rising wages cause a movement of population across cities, some cities must lose population while others gain population. In cities that lose population, $\frac{\mathrm{d}v}{\mathrm{d}w} > 0$, and by Equation (5.28) the same must be true in every city to maintain a spatial equilibrium. Stated differently, an economy-wide increase in wages must cause utility to rise everywhere. Since $\frac{\partial U}{\partial v} > 0$ and $\frac{\partial v}{\partial N_i} < 0$, it follows that $\frac{\partial^2 N_i}{\partial A_i \partial w}$ has the same sign as $\frac{\partial^2 U}{\partial A_i \partial v}$. Hence, if utility is supermodular in the level of amenities and the sub-utility derived from housing and other goods ($\frac{\partial^2 U}{\partial A_i \partial v} > 0$), an economy-wide increase in income makes cities with greater amenities grow in population relative to other cities. Intuitively, if the value that consumers place on additional amenities increases as they are able to afford a better bundle of housing and other goods, aggregate economic growth makes high-amenity cities relatively more attractive.

Both supply and demand channels suggest a link between amenities and urban growth. Taken literally, the supply explanation described by Equation (5.26) suggests regressing changes in population on changes in amenities:

$$\Delta_{t+1,t} \log N_i = \beta_0 + \beta_1 \Delta_{t+1,t} A_i + \epsilon_{it}. \tag{5.30}$$

This regression mirrors regression (5.14), with the only difference that now the level of amenities replaces (log) commuting costs as the driver of city growth. By the same argument that was applied to regression (5.14), if the adjustment of population is sluggish after a change in amenities, one is naturally led to estimate instead:

$$\Delta_{t+1,t} \log N_i = \beta_0 - \lambda \log N_{it} + \beta_1 A_{it} + \epsilon_{it}. \tag{5.31}$$

This is the counterpart to the transportation regression (5.15) and, equivalently to that case, the coefficient of interest, β_1, measures the effect of amenities on population in cities and λ the speed of population adjustment.

Turning to the demand-for-amenities explanation, the comparative statics of Equation (5.29) suggest regressing local population changes on local amenities interacted with national wage growth. In practice, interacting national wage growth and amenities is likely to be problematic for several reasons. First, sluggish population adjustment is likely to make extremely difficult the identification of faster population growth for cities with higher amenities during periods when national wage growth is higher. For instance, cyclical downturns, which imply both lower wages and less mobility, are likely to act as a confounding factor. Second, cyclical behavior and sluggish adjustment also suggest measuring population growth over periods of five or ten years, which limits the potential length of a panel of city growth. To avoid these problems, one may prefer to rely on cross-sectional variation rather than longitudinal variation and check whether, against the background of rising incomes nationwide, high-amenity cities have attracted more people. Note that this leads to regression (5.31) again.

Since both demand and supply explanations can be used as motivations for the specification of Equation (5.31), estimating such a regression will help identify the overall effect of amenities on urban growth but will not assist us much in disentangling demand and supply explanations. While we return to this issue later, at this stage the most pressing issue is how to measure amenities.

From the spatial equilibrium we can define the shadow price of amenities as:

$$Q = h(0)\frac{\mathrm{d}P(0)}{\mathrm{d}A_i}, \tag{5.32}$$

which is the extra housing cost that a resident is willing to pay to live in a city with higher amenities. Note that this is valued at the CBD ($x = 0$) so that commuting costs do not have to be considered separately (utility equalization within the city implies that the same shadow price applies to other locations within the city with higher commuting costs and lower housing prices). In Equation (5.32), the shadow price of amenities only depends on housing variables (h and P) and the level of amenities (A). In a more general setting where amenities and land enter production as well as consumption, expression (5.32) also contains a wage term as in Roback (1982) and subsequent literature:

$$Q = h(0)\frac{\mathrm{d}P(0)}{\mathrm{d}A_i} - \frac{\mathrm{d}w}{\mathrm{d}A_i}. \tag{5.33}$$

This wage term reflects that amenities can affect not just land prices but also wages.[20]

There is a long tradition of empirical research motivated by Equation (5.33) that attempts to value amenities by separately regressing housing expenditures and wages in cities on a set of broadly defined amenities (e.g. Blomquist et al. 1988). The amenities considered range from the availability of good restaurants to nice architecture to low crime or richly endowed public libraries. The coefficients on each of these amenities in the housing regression and in the wage regression are used in place of, respectively, $h(0)\frac{\mathrm{d}P(0)}{\mathrm{d}A_i}$ and $\frac{\mathrm{d}w}{\mathrm{d}A_i}$ in Equation (5.33) to compute a shadow price of each individual amenity. The overall value of amenities in each location can then be assessed by aggregating its bundle of amenities valued at its estimated shadow price. Using this approach to estimate the impact of amenities on city growth is problematic. Most of the amenities that are usually considered are likely to be endogenous to city growth. For instance, whether good restaurants cause city growth or result from it is unclear. Then, aggregating an arbitrary number of poorly identified coefficients is unlikely to be informative about the effects

[20] There are two different channels. First, amenities may impact productivity directly. For instance, a coastal location may lower trade costs while being enjoyed as a consumption amenity. Second, consumption amenities can affect wages indirectly when land is a factor of production, since higher amenities imply higher land prices. Then, firms substitute away from land in production, which lowers the marginal product of labor. In addition, Moretti (2011) shows that with imperfectly mobile workers, the expression that values amenities should also contain a term to reflect imperfect mobility. The estimation of this mobility term is an open challenge.

of amenities on city growth. In fact, this approach often provides quality-of-life rankings that seem hard to reconcile with commonly accepted notions of attractiveness.[21]

The main advantage of the standard approach building on Roback (1982) is that it allows us to see whether the main effect of amenities is to raise the utility of residents or to provide a productivity advantage to firms. If amenities mainly raise the utility of residents, these will be willing to accept higher rents or lower wages in order to enjoy them ($h(0)\frac{\mathrm{d}P(0)}{\mathrm{d}A_i} > 0$, and $\frac{\mathrm{d}w}{\mathrm{d}A_i} < 0$ in Equation (5.33). If amenities mainly provide a productivity advantage, firms will be willing to incur higher building costs or higher wage costs to locate where they can benefit from them ($h(0)\frac{\mathrm{d}P(0)}{\mathrm{d}A_i} > 0$ and $\frac{\mathrm{d}w}{\mathrm{d}A_i} > 0$).

This approach is useful to study not just amenities but also other city characteristics. For instance, Ottaviano and Peri (2006) study the extent to which a greater diversity of countries of origin among residents of US cities is associated with productivity advantages or consumption amenity advantages. Following Altonji and Card (1991) and Card (2001), they instrument the diversity of each city by combining historical stocks of immigrants by origin at the local level with immigration flows by origin at the national level (under the assumption that recent immigrant flows sort across cities proportionately to historical stocks of the same origin). They find that cities with a greater diversity of countries of origin have both higher wages and higher land rents and conclude that it is the higher productivity effect of diversity that dominates. Given the discussion above, this approach is most useful when applied to study a single well-defined amenity or city characteristic that is either exogenous or appropriately instrumented.

To solve the problems of mixing heterogenous amenities, many of which are endogenous, much of recent research on urban amenities has focused on the weather. That weather variables should be valued highly by consumers is needed for them to play an important role in location decisions and more specifically in the growth of cities. Reassuringly, the literature that values amenities usually estimates high shadow prices for climate-related variables.[22] The weather is also often deemed to be exogenous. Although most manifestations of the weather are not a consequence of city growth, some caution here is nonetheless needed since most measures of weather are likely to be correlated to other determinants of urban growth. This suggests enriching regression (5.31) with a number of control variables and assessing the robustness of the estimated weather coefficients against the inclusion of these controls.

As argued by Glaeser et al. (2001), weather—as measured by January and July temperatures—is one of the most reliable predictors of city growth in recent US history. Warmer temperatures in January and cooler temperatures in July are both strongly associated with city growth. These findings are confirmed and greatly extended in

[21] For instance, Blomquist et al. (1988) rank Pueblo (CO), a county in Macon (GA), and one in Binghamton (NY) as three of the most desirable places to live in the United States whereas New York City is close to the bottom.

[22] This is true of Blomquist et al. (1988), Albouy (2008), and many others.

Rappaport's (2007) comprehensive study. His main conclusion is that a pleasant climate (in the form of mild winters, and summers that are not too hot) is a major engine of population growth for US counties between 1970 and 2000. More specifically, he shows that a standard deviation in January temperature is associated with a 0.6 standard deviation in population growth. For July temperature, one standard deviation is associated with a 0.2 standard deviation in population growth. For European countries, Cheshire and Magrini (2006) also reach similar conclusions.

This said, in the United States the correlation between summer and winter temperatures and city growth reflects to a large extent the rise of Southern cities. As pointed by Glaeser and Tobio (2008), Southern cities which offer milder winters and warmer summers also differ from other US cities in the evolution of their wages and housing costs. These are potentially two important missing variables in regression (5.31), since both housing costs and wages affect the spatial equilibrium condition (5.25). Importantly, housing in Southern cities appears to have become relatively cheaper. This obviously raises some doubts about the importance of the weather as a key driver of city growth.[23]

In this respect, Rappaport (2007) makes an important observation that the effects of nice weather in the United States are also observed outside of the south for areas with mild summers. In addition, for these areas the development of air-conditioning made little difference, if any. This result is more immediately consistent with explanations that rely on a rising demand for amenities than those that highlight supply changes.[24]

Another direction taken by recent research is to look for a summary variable that would proxy for the entire bundle of amenities in a city. The first possibility, suggested by Glaeser et al. (2001), is to estimate the aggregate value of amenities in a city relative to another city or to the average city as the sum of the difference in housing costs minus the difference in wages. This builds again on the spatial equilibrium condition for workers, which implies that differences in real wages (i.e. nominal wages corrected of housing costs) should be offset by differences in amenities. Albouy (2008) implements this strategy empirically, while also correcting for differences in non-labor income, in federal taxes, and in the price of goods other than housing. He obtains an aggregate amenity value for each city that better corresponds to perceived notions of attractiveness. He then regresses this aggregate value on a number of individual amenity variables to study the relative importance of each. This approach seems promising.

[23] Pushing the logic of the Roback (1982) model, Glaeser and Tobio (2008) find that the relative decline in the costs of housing and rising productivity in the US south imply no increase in the willingness to pay for southern amenities (actually these two features imply a decline).

[24] Matters are actually even more complicated than this because amenities and land-use regulations appear to interact in some interesting fashion. Gyourko et al. (2013) show that some cities with good amenities such as San Francisco, Santa Cruz, or Boston have imposed ever more restrictive zoning regulations. As a result, population growth has been limited but property price appreciation has been extremely strong. This has also led to the sorting of high-income workers in these "superstar cities."

As Carlino and Saiz (2008) note, however, it still is subject to the concern that current property prices also partly reflect expectations about future population growth. Then studying the effects of amenities on urban growth by regressing population growth on an "amenity index" that contains expectations of population growth is potentially problematic.

As another summary variable capturing a large set of amenities, Carlino and Saiz (2008) propose using the number of leisure visits to each city. They first show that leisure visits, as collected by a consultancy in the tourism industry, correlate well with alternative measures of amenities and quality of life, including Albouy's (2008). Second, they regress population growth between 1990 and 2000 for US metropolitan areas on leisure visits and find that the elasticity of population with respect to leisure visits is about 2% over this 10-year period. This coefficient is robust to the inclusion of many other control variables. This said, we can again imagine a number of ways leisure visits might be correlated with city growth without having a causal effect on the latter. Tourism is itself a strong growth industry. However, the correlations are robust to the exclusion of the likes of Las Vegas and Orlando. In addition, fast-growing cities receive a greater inflow of newcomers who, in turn, may receive more visits from family and friends. We can again use an instrumental variable approach to circumvent this simultaneity problem. Carlino and Saiz (2008) use two exogenous determinants of leisure visits: the number of historic places and the coastal share within a 10 kilometer radius of the central city. This instrumental variable approach leads to an even higher elasticity of city population with respect to amenities of 4%.

5.5. AGGLOMERATION ECONOMIES

The monocentric city model, by focusing on the trade-off between commuting costs and house prices within a single city, highlights the costs of bigger cities. To study meaningfully multiple cities within an urban system, we need to consider also the productive benefits of bigger cities. For simplicity, let us abstract from amenity differences and refer back to the spatial equilibrium condition of Equation (5.13). If we treat the wage w_i as a parameter independent of a city's population, then $\frac{\mathrm{d}v}{\mathrm{d}N_i} < 0$, so that any individual prefers to live alone than to live in a city of any size. Even if we endogenise the wage but have $\frac{\mathrm{d}w}{\mathrm{d}N_i} < 0$, as in Section 5.3.1, it remains that $\frac{\mathrm{d}v}{\mathrm{d}N_i} < 0$. Stated differently, if new potential sites for cities are available, then agglomeration economies are essential to understand why cities exist at all.

A simple way to incorporate agglomeration economies into the monocentric city model is to recognize that the wage in each city i depends positively on its population: $w_i = w(N_i)$ with $\frac{\mathrm{d}w_i}{\mathrm{d}N_i} > 0$. Many urban models have this feature and there is broad empirical evidence supporting it, as discussed below. Now, as a city's population increases there is both the negative effect on residents' utility of rising urban costs $\left(\frac{\mathrm{d}v}{\mathrm{d}P(0)}\frac{\mathrm{d}P(0)}{\mathrm{d}N_i} < 0\right)$ and the positive effect of stronger agglomeration economies $\left(\frac{\mathrm{d}v}{\mathrm{d}w_i}\frac{\mathrm{d}w_i}{\mathrm{d}N_i} > 0\right)$. The fact that in reality there is no one extremely large city but instead multiple cities of finite population size suggests that v is a concave and non-monotonic function of N_i. Initially, agglomeration economies dominate and utility increases with city size. Eventually, higher costs of housing and commuting dominate and utility decreases with city size.

5.5.1 City Formation and Urban Systems

We now develop a simple model of a system of cities in the tradition of Henderson (1974). Before turning to city creation, we begin by deriving an expression for $w(N_i)$ built on explicit micro-foundations, following Abdel-Rahman and Fujita (1990). Suppose there are multiple perfectly competitive final sectors, identified by superindex j, each of which produces a homogenous final good that is freely tradable across cities. Final production technology features a constant elasticity of substitution across intermediate inputs that are sector-specific and non-tradable across cities, so that output in sector j in city i is:

$$Y_i^j = B^j\left\{\int_0^{m_i^j}\left[y_i^j(h)\right]^{\frac{1}{1+\sigma^j}}\mathrm{d}h\right\}^{1+\sigma^j}, \tag{5.34}$$

where h indexes intermediate varieties, $y_i^j(h)$ denotes intermediate input quantities, m_i^j denotes the endogenous mass of intermediates available in sector j in city i, B^j is a measure of technological development that will be useful for comparative statics, and $0 < \sigma_j < 1$. As in Ethier (1982), intermediates are produced by monopolistically competitive firms à la Dixit and Stiglitz (1977) with technology:

$$y_i^j(h) = \beta^j l_i^j(h) - \alpha^j, \tag{5.35}$$

where $l_i^j(h)$ is firm-level employment. Workers are freely mobile across sectors as well as across locations.

In equilibrium, all firms in any given city and sector set the same profit-maximizing price $q_i^j = w_i^j(1+\sigma^j)/\beta^j$. Free entry drives intermediate profits to zero: $q_i^j y_i^j - w_i^j l_i^j = 0$. Using Equation (5.35) and the pricing rule to expand this expression and solving for y_i^j shows there is a fixed level of intermediate output consistent with zero profits in each sector:

$$y_i^j = \frac{\alpha^j}{\sigma^j}. \tag{5.36}$$

Equating (5.35) and (5.36) allows solving for the constant workforce of each intermediate supplier: $l_i^j = \alpha^j(1+\sigma^j)/(\beta^j\sigma^j)$. Hence, the equilibrium mass of intermediate producers in sector j of city i is:

$$m_i^j = \frac{N_i^j}{l_i^j} = \frac{\beta^j\sigma^j}{\alpha^j(1+\sigma^j)}N_i^j. \tag{5.37}$$

By choice of units for intermediate output, we can set $\beta^j = (1+\sigma^j)(\alpha^j/\sigma^j)^{\sigma^j/(1+\sigma^j)}$. Substituting Equations (5.36) and (5.37) into (5.34) yields aggregate production in sector j of city i as:

$$Y_i^j = B^j \left[\left(y_i^j \right)^{\frac{1}{1+\sigma^j}} m_i^j \right]^{1+\sigma^j} = B^j \left(N_i^j \right)^{1+\sigma^j}. \tag{5.38}$$

Zero profits in final production imply that $w_i^j N_i^j = P^j Y_i^j$. Thus, wages are given by:

$$w_i^j = P^j B^j \left(N_i^j \right)^{\sigma^j}. \tag{5.39}$$

Note that aggregate production is subject to increasing returns at the sector and city level. As the size of a sector in a city increases, it supports a wider range of shared intermediate suppliers. Gains from variety in final production then imply that it is possible to increase output more than proportionately relative to the increase in employment producing intermediates for this sector. The literature has explored many alternative agglomeration mechanisms with similar implications, both theoretically and empirically (see Duranton and Puga, 2004; Rosenthal and Strange, 2004; Puga, 2010, for reviews).

In equilibrium, all cities are specialized in a single sector. To see this, note that any equilibrium must be such that wages are equalized across sectors in a city. Consider now a small perturbation in the distribution of employment across sectors within a city keeping its total population constant. Since $\partial w_i^j / \partial N_i^j > 0$, sectors that see employment rise begin paying higher wages and attract more workers, whereas sectors that see employment fall begin paying lower wages and lose more workers. The only equilibrium that is stable with respect to perturbations in the distribution of local employment across sectors has all local workers employed by the same sector. Two assumptions drive full urban specialization in this simple model. First, since intermediates are sector-specific, agglomeration economies arise within sector only. Thus, mixing multiple sectors in a single city would increase house prices and commuting costs without bringing any benefits relative to having sectors operate in separate cities. Second, final goods are assumed freely tradable, which eliminates the proximity-concentration trade-off that would otherwise arise. Cross-sector externalities and trade costs would provide static motives for a diversity of sectors in cities. See Duranton and Puga (2000) for a review of such static extensions. Below we also discuss a dynamic alternative proposed by Duranton and Puga (2001).

Next we model the internal structure of cities using a version of the monocentric city model that both generalizes and simplifies the version presented in Section 5.2. In particular, let us generalize the specification for commuting costs so that they are not necessarily linear but instead have an elasticity γ with respect to distance. Further, we have so far had commuting costs incurred in units of the single consumption good in the economy. Since we are now considering multiple consumption goods, and each city is specialized in the production of just one such good, let us have commuting costs in each city incurred in terms of the locally produced good j. We can then write commuting costs for a resident living at distance x from the CBD as $P^j \frac{1+\gamma}{\gamma} \tau x^\gamma$, where the normalization constant $\frac{1+\gamma}{\gamma}$ is just meant to simplify notation below.

At the same time, to obtain closed-form solutions without the need to specify a functional form for utility, let us make the simplifying assumption that all residences have the same size and are built with a constant capital to land ratio. Thus, every individual consumes one unit of floorspace built on one unit of land with a fixed amount of capital. Relative to the version of the monocentric city model seen before, this implies the restriction $h(x) = f(x) = 1$. Hence, the physical extent of the city is the same as its population: $\overline{x} = N_i$.

Then, totally differentiating the spatial equilibrium condition with respect to x yields the land and house price gradients: $\frac{\mathrm{d}R(x)}{\mathrm{d}x} = \frac{\mathrm{d}P(x)}{\mathrm{d}x} = -P^j(1+\gamma)\tau x^{\gamma-1}$. Without further loss of generality, let us set the cost of capital per residence equal to zero (so that house and land prices are equal instead of differing by a constant) and the rental price of land when not in urban also equal to zero (otherwise, land prices would be higher everywhere by the value of that rent). Integrating the land price gradient $\frac{\mathrm{d}R(x)}{\mathrm{d}x}$ and using $R(\overline{x}) = \underline{R} = 0$ and $\overline{x} = N_i$ to obtain the integration constant, we can express house and land prices as:

$$P(x) = R(x) = P^j \frac{1+\gamma}{\gamma} \tau \left(N_i^\gamma - x^\gamma\right). \tag{5.40}$$

Note that with fixed housing consumption, utility equalization within the city implies that the sum of commuting costs and housing expenditure is the same for every resident, and equal to the house price at the CBD: $P^j \frac{1+\gamma}{\gamma} \tau x^\gamma + P(x) = P(0) = P^j \frac{1+\gamma}{\gamma} \tau N_i^\gamma$. Integrating $R(x)$, as given by Equation (5.40), over the physical extent of the city yields total land rents:

$$R_i = \int_0^{N_i} R(x)\mathrm{d}x = P^j \tau N_i^{1+\gamma}. \tag{5.41}$$

We now consider endogenous city formation in this framework. Following Becker and Henderson (2000) we study three alternative mechanisms: self-organization, land developers, and active local governments.

Let us begin with self-organization, so that cities arise as the result of the uncoordinated decisions of individual agents. Since free trade in final goods equalizes their prices across locations and since housing consumption is fixed, utility only depends on income net of housing costs and commuting costs. To compute income, we need to consider what happens to land rents. Under self-organization the simplest assumption is that land rents are shared by all residents in the city, each getting $\frac{R_i}{N_i}$. Recall that the sum of housing costs and commuting costs for every resident is equal to the price of housing at the CBD, $P(0)$. Using c_i to denote per-capita income net of housing costs and commuting costs, we can

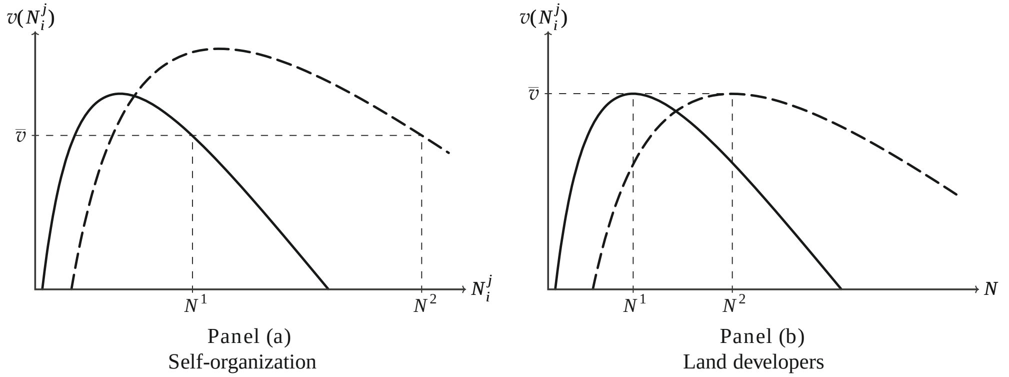


Figure 5.2 City sizes and utility.

write this as $c_i = w(N_i) + \frac{R_i}{N_i} - P(0)$. Substituting Equations (5.39)–(5.41) this becomes:

$$c_i = P^j \left(B^j N_i^{\sigma^j} - \frac{\tau}{\gamma} N_i^{\gamma} \right). \tag{5.42}$$

We consider a continuum of cities. Under self-organization, the number (mass) of cities in each sector is given, since no agent is large enough to create a new city on their own. An equilibrium distribution of population across cities simply requires utility equalization and stability with respect to small perturbations. Panel (a) of Figure 5.2 plots utility as a function of population size, as given by Equation (5.42), for two cities specializing in different sectors.[25] Provided that $\gamma > \sigma^j$, for each city type, utility is a concave function of population. Initially, agglomeration economies dominate and utility increases with city size but eventually higher costs of housing and commuting dominate and utility decreases with city size. The difference in specialization affects the relationship between population and net income through differences in agglomeration economies (σ^j), in the productivity shifter (B^j), and in final goods prices (P^j).

Consider first the solid curve. For any given level of utility, there are at most two population sizes that provide this utility level, one above and one below the efficient size that maximizes utility in a city of that specialization. However, cities below the efficient size cannot survive small perturbations in the distribution of workers. This is because those that gain population get closer to the efficient size and attract even more workers while those that lose population get further away from the efficient size and lose even more workers. If cities specializing in sector 1 (with utility plotted as a solid curve) have the population marked by N^1 in panel (a) of Figure 5.2, then cities specializing in sector 2 (with utility plotted as a dashed curve) have the population marked by N^2 to ensure workers have no incentive to migrate across cities. Thus, under self-organization, all cities

[25] The figure is plotted for $\gamma = \tau = 0.045, \sigma^1 = 0.038, \sigma^2 = 0.040, B^1 = B^2 = 1, P^1 = 1$, and $P^2 = 1.43$.

of the same specialization have the same population size and this size cannot be smaller than the efficient city size for each sector.

The reason why cities tend to be too large under self-organization is a lack of coordination: each worker would prefer more cities of a smaller size each in their sector but a worker alone cannot create a new city. Following Henderson (1974), suppose all land at each potential location for a city is controlled by a land developer who collects land rents. There is free entry and perfect competition among land developers. Each of them announces a population size and specialization for their city as well as a level of transfers T_i^j that they are willing to provide to workers locating in their city. When active, each land developer seeks to maximize land rents $R_i = P^j \tau N_i^{1+\gamma}$, net of any transfers:

$$\max_{\{T_i, N_i\}} \Pi_i = P^j \tau N_i^{1+\gamma} - T_i N_i, \tag{5.43}$$

subject to the participation constraint for workers. This constraint results from incorporating transfers into workers' income and ensuring that they achieve the same per-capita income net of housing and commuting costs $\bar{c}$ as in the best alternative location:

$$P^j B^j N_i^{\sigma^j} + T_i - P^j \frac{1+\gamma}{\gamma} \tau N_i^{\gamma} = \bar{c}, \tag{5.44}$$

Further, population must be positive: $N_i \geqslant 0$.

Solving for T_i^j in Equation (5.44) and substituting this into (5.43) yields an equivalent program:

$$\max_{\{N_i\}} \Pi_i = P^j B^j N_i^{1+\sigma^j} - P^j \frac{\tau}{\gamma} N_i^{1+\gamma} - \bar{c} N_i. \tag{5.45}$$

The equivalence between the programs (5.43) and (5.45) shows that, in maximizing land rents net of transfers, developers behave as if running a factory-town, in which they hired workers at their going net compensation levels ($\bar{v}$) and sold in national markets all output produced in the city ($P^j Y_i = P^j B^j N_i^{1+\sigma^j}$) net of commuting cost expenditure ($P^j \frac{\tau}{\gamma} N_i^{1+\gamma}$), keeping any residual as a profit.

The first-order condition for (5.45) is:

$$\bar{v} = P^j \left((1+\sigma^j) B^j N_i^{\sigma^j} - \frac{1+\gamma}{\gamma} \tau N_i^{\gamma} \right). \tag{5.46}$$

Substituting this first-order condition into Equation (5.45) yields maximized profits for the developer:

$$\Pi_i = P^j \tau N_i^{1+\gamma} - \sigma^j P^j B^j N_i^{1+\sigma^j}. \tag{5.47}$$

Equating this expression for Π_i with the expression in the original developer's program of Equation (5.43) shows that, in maximizing their profits, developers offer each worker the following transfer:

$$T_i = \sigma^j P^j B^j N_i^{\sigma^j}. \tag{5.48}$$

This transfer covers the gap between the market wage $w_i^j = P^j B^j N_i^{\sigma^j}$ and the city-level marginal product of labor $(1+\sigma^j)P^j B^j N_i^{\sigma^j}$. Thus, in maximizing their profits, developers internalize the city-level externality created by agglomeration economies.

Free entry and perfect competition among land developers exhausts their profits. Using $\Pi_i = 0$ in Equation (5.47) and solving for N_i, we obtain equilibrium city size in the presence of land developers:[26]

$$N_i = \left(B^j \frac{\sigma^j}{\tau}\right)^{\frac{1}{\gamma-\sigma^j}}. \tag{5.49}$$

This is the optimal city size. To see this, consider a situation where land at each site, instead of being owned by a developer, is shared by residents who elect a local government that runs the city so as to maximize their welfare. The utility of each resident in the city is given by Equation (5.42). This utility is maximized for N_i given by Equation (5.49).

Panel (b) of Figure 5.2 plots utility as a function of population size for two cities specializing in different sectors in the presence of competitive land developers.[27] There are two differences with respect to the equilibrium under self-organization of panel (a). First, taking final goods prices as given, land developers are induced through competition to create cities of the efficient size for their sector. If this was not the case, then another developer could enter and make a profit by offering a more efficient city and capturing as profit, by means of lower transfers, the difference in utility relative to the best alternative city. Optimal city size is obtained because, in equilibrium, developers must make transfers that cover the gap between private and social returns opened by agglomeration economies. With zero profits for developers, total land rents equal total transfers, and thus, are just enough to cover that gap.[28] The second difference is that if, as shown by the dashed curve for sector 2 in panel (a), cities in a certain sector offer higher utility at the peak, this will attract more developers to this sector. Entry increases economy-wide output in that sector and lowers its general equilibrium final good price until all cities offer the same utility at the efficient size for their sector. This is what shifts downwards the utility for sector 2 in panel (b) relative to panel (a).

Equilibrium city sizes, as given by Equation (5.49), are the result of what Fujita and Thisse (2002) call the fundamental trade-off of urban economics: between agglomeration

[26] Note that the second-order condition for profit maximization requires $\gamma > \sigma^j$.

[27] The curves are plotted with developer profits driven to zero, in which case utility is still given by Equation (5.42), like under self-organization. We use the same parameters as in panel (a), except P^2, which now adjusts through free entry by land developers until utility is equalized across cities.

[28] Using $\Pi_i = 0$ in Equation (5.43) implies $T_i N_i = P^j \tau N_i^{(1+\gamma)} = R_i$, while multiplying both sides of Equation (5.48) by N_i implies $T_i N_i = \sigma^j P^j B^j N_i^{(1+\sigma^j)}$. This is a classic result in urban economics known as the Henry George Theorem (Serck-Hanssen, 1969; Starrett, 1974;Vickrey, 1977). Its best-known version is associated with local public goods (Flatters et al. 1974; Stiglitz, 1977; Arnott and Stiglitz, 1979).

economies, which make wages and productivity increase with city size, and crowding diseconomies, which makes commuting and housing costs increase with city size.[29] The lower the magnitude of commuting costs, as captured by τ, the larger is city size. This confirms that the implication of the monocentric city model that improvements in commuting infrastructure foster urban growth is robust to the introduction of agglomeration economies and endogenous city creation. A decrease in the elasticity of commuting costs with respect to distance, γ, similarly leads to larger cities. Turning to the other side of the trade-off, stronger agglomeration economies, as measured by σ^j, make cities larger. Since crowding costs are unlikely to be very different for workers engaged in different activities but agglomeration economies will be stronger in some sectors than in others, there is an important link between sectoral specialization and city size. In particular, cities specializing in sectors with higher agglomeration economies (high σ^j) will be larger in size. Black and Henderson (2003) show that US cities can be classified into groups with similar specialization and size.

In the model in this section, sectors are defined as groups of firms using similar bundles of inputs. In a more general setting, the combinations of activities present in cities of different sizes depend on more complicated links, some of which form through agglomeration economies extending across sectors, and others through links between various parts of the production process within a firm. Duranton and Puga (2005) model such a multi-stage production process within each firm in a general equilibrium model of an urban system. They show that as technological developments and transport improvements facilitate the spatial fragmentation of activities within the firm, management and business service provision will tend to concentrate in larger cities whereas actual production will concentrate in smaller cities. They also show that such a process has occurred in the United States since the 1950s. In effect, this implies an increasing specialization by functions and occupations instead of traditional sectoral divisions.

The technological parameter in the model, B^j, allows us to study the effects on city growth of sectoral shocks and of aggregate growth. To study sectoral shocks, consider a continuum of sectors, so that a shock to just one of these sectors does not affect the entire economy. Then, by Equation (5.49), a positive shock to B^j makes cities specializing in this sector grow in size. Higher supply lowers output prices in this sector, which induces some developers to move away from it until utility equalization, which implies

[29] Equation (5.49) reflects city sizes with developers. Under self-organization, we must be in the downward-sloping portion of the size-utility relationship depicted in Figure 5.2 instead of at the optimum. For given goods prices, a fall in τ or γ or an increase in σ^j or B^j pushes the size-utility curve outwards and makes city size increase, resulting in the same qualitative comparative statics as with developers. However, while with efficient city sizes the envelope theorem implies no additional effects operating through goods prices, without developers we must consider such general equilibrium price effects which, if sufficiently strong, could in principle offset the standard comparative statics. In Section 5.6.4 we discuss further the importance of considering the elasticity of demand and price effects when looking at the impact of productivity changes on cities.

that the value of output per worker net of commuting costs must stay constant, is restored (see Equation (5.42)). At the new equilibrium, there will be fewer cities specializing in the sector experiencing the positive technology shock and each of them will be larger in population.

Turning to aggregate technical change, consider a situation where $B^j = B$ is common to all sectors and experiences an increase. Given that this will affect every city, we can no longer treat utility as constant. Instead, it will change equally everywhere. By Equation (5.42) and the envelope theorem:

$$\frac{B}{\bar{c}}\frac{\mathrm{d}\bar{c}}{\mathrm{d}B} = \frac{BN_i^{\sigma^j}}{BN_i^{\sigma^j} - \frac{\tau}{\gamma}N_i^{\gamma}} > 1. \tag{5.50}$$

Thus, cities amplify the growth effects of technical progress; per-capita income net of commuting costs increases more than proportionately with aggregate technical change.

By Equation (5.49) aggregate technical also makes cities grow in population, with initially larger cities tending to grow more. Henderson and Wang (2007) suggest that over the last few decades this tendency of aggregate technical change to increase the relative size of the then largest cities has been offset by the tendency of increased democratization around the world to lower urban concentration, thus helping keep city size distributions roughly stable. This effect of democratization can also be linked to the systems of cities model presented in this section. As we have seen, city developers or local governments can help cities get closer to their efficient size, while without them cities tend to be too few and too large. For this to be the case, Henderson and Wang (2007) argue that local governments need to be able to set up new cities, to finance new infrastructure so that existing towns can expand into cities, and to enable land development in well-functioning land markets where regulation is transparent and land ownership is clearly defined. Arguably, all these are characteristics that are closely related to more democratic regimes.

The operation of city developers and local governments in the model of this section is purely static. As a result, changes such as the sectoral shocks we have examined or aggregate population growth cause swings in population sizes. Henderson and Venables (2009) develop a dynamic model of city formation where housing and urban infrastructure are durable. Then population changes smoothly and it is instead the price of housing that is subject to swings. In this dynamic version of the model, cities are created sequentially and city developers borrow to finance development. The subsidies paid by developers are no longer as in Equation (5.48) and as such the total value of the subsidy equals the total value of the externality created by agglomeration economies. Instead the subsidy to the marginal migrant covers the marginal externality he or she creates. Cuberes (2011) provides empirical evidence of such sequential city growth: in many countries the largest cities grow more initially, but over time their growth tends to settle and smaller cities start growing faster.

Following Bartik (1991), the link between specialization and city size has often been used to predict city growth in multiple contexts. Applications include studying the interactions between land-use regulations and urban growth as described in Section 5.3.1 above. For each city, the Bartik predictor takes employment growth by industry at the national level (excluding the city at hand) and averages it across industries using initial local employment shares as weights. This measure is sometimes alleged to provide a measure of city growth that is clean from city-specific shocks.[30]

While the Bartik predictor may be plausibly used as a measure of local labor demand shocks affecting a city, like in Glaeser and Gyourko (2005), using it as an instrument may be more problematic. To see this, consider regressing changes in local output or in local wages on changes in local employment to estimate the agglomeration elasticity σ as suggested by Equations (5.38) or (5.39). If there are sectoral shocks with some unobserved component, for instance affecting B^j, these will become part of the error term in the regression. By construction, those sectoral shocks will also be part of the Bartik predictor. Thus, the Bartik predictor will violate the exogeneity requirement to be used as an instrument for changes in local employment.

Finally, note that the model above assumes free trade between cities. As noted above, cities fully specialize in equilibrium because of the combination of free trade in final goods and within-sector agglomeration economies. Introducing trade costs brings in the proximity-concentration trade-off that is familiar from international trade (Brainard, 1997). A diversity of sectors in a city reduces the strength of agglomeration economies but saves transport costs when supplying a mixed bundle of goods to local consumers. The prediction that lower transportation costs between cities should lead to greater urban specialization has received mixed empirical support. With the secular decline in transport costs, one would expect urban specialization to increase. Instead, sectoral specialization in US cities has declined since at least the 1970s while functional specialization has increased during the same period (Duranton and Puga, 2005). Allowing for transportation costs to respond differently to changes in infrastructure generates a richer set of predictions. In particular, Duranton et al. (2013) develop a framework where more highways to enter or exit a city make it cheaper to export heavier goods which are more sensitive to the provision of roadway. For US cities they find that cities with more highways tend to be more specialized in the production of heavier goods and export them more.

Redding and Sturm (2008) consider a model similar to the one developed above but with only one sector for which differentiated varieties directly enter the utility function, as in Helpman (1998). There are still both agglomeration economies and crowding costs

[30] It is worth noting that endogenous changes in the number and specialization of cities can alter the link between changes in national sectoral employment and changes in city sizes. For instance, as seen above when discussing the comparative statics on the productivity shifter B^j, sector-specific shocks that increase equilibrium city sizes typically lead to a consolidation of the sector's employment in fewer cities of larger size, so that cities with similar sectoral composition may experience very different changes during the adjustment. Alternatively, a positive demand shock may lead to a sector being present in more cities without significant changes in the size of cities that initially hosted the sector.

related to city size. In addition, the introduction of transport costs creates additional concentration and dispersion forces related to the relative location of cities in space. The interaction between transportation costs and increasing returns in production creates a home market effect where firms want to concentrate their production in cities with good access to large markets (Krugman, 1980). Counteracting this is the fact that firms close to large markets face a larger number of competitors. An important prediction of this framework is that cities with a better market access should be larger. Redding and Sturm (2008) successfully test this prediction using the division of Germany after the Second World War as a natural experiment. They show that West German cities located close to the Iron Curtain lost significant market access and declined in population relative to other West German cities.

5.5.2 Empirical Magnitude of Urban Benefits and Costs

We have seen that agglomeration economies are essential to understand why cities exist at all, and their magnitude fundamentally affects city sizes and patterns of firm and worker location. Thus, quantifying agglomeration economies has been a key aim of the empirical literature in urban economics, especially in recent years.

Agglomeration economies imply that firms located in larger cities are able to produce more output with the same inputs. Thus, perhaps the most natural and direct way to quantify agglomeration economies is to estimate the elasticity of some measure of average productivity with respect to some measure of local scale, such as employment density or total population. This elasticity corresponds to parameter σ in the model just presented. In early work, Sveikauskas (1975) regressed log output per worker in a cross-section of city-industries on log city population and found an elasticity of about 0.06. More recent studies have obtained estimates of around 0.02–0.05, after dealing with three key potential problems in the original approach.

The first problem is that measuring productivity with output per worker will tend to provide upwardly biased estimates of σ, since capital is likely to be used more intensively in large cities. To address this concern, recent contributions focus on total factor productivity, calculated at the aggregate level for each area being considered or, more recently, at the plant level. A particularly influential contribution using this approach is that of Henderson (2003), who estimates total factor productivity using plant-level data in high-tech and machinery sectors for the United States.

A second concern when estimating agglomeration economies is that productivity and city size are simultaneously determined. If a location has an underlying productive advantage, then it will tend to attract more firms and workers and become larger as a result. Following Ciccone and Hall (1996), the standard way to tackle this issue is to instrument for the current size or density of an area. The usual instruments are historical population data for cities and characteristics that are thought to have affected the location of population in the past but that are mostly unrelated to productivity today. The logic behind these instruments is that there is substantial persistence in the spatial distribution of population (which provides relevance), but the drivers of high productivity today greatly differ from those in the distant past (which helps satisfy the exclusion restriction). Most studies find that reverse causality is only a minor issue in this context and that estimates of σ are not substantially affected by instrumenting (Ciccone and Hall, 1996; Combes et al. 2010). An alternative strategy to deal with a potential endogeneity bias is to use panel data and include city-time fixed effects when estimating plant-level productivity, to capture any unobserved attributes that may have attracted more entrepreneurs to a given city (Henderson, 2003).[31] Finally, Greenstone et al. (2010) follow an ingenious quasi-experimental approach. They identify US counties that attracted large new plants involving investments above one million dollars as well as runner-up counties that were being considered as an alternative location by the firm. They find that, after the new plant opening, incumbent plants in chosen counties experience a sharp increase in total factor productivity relative to incumbent plants in runner-up counties.

A third concern with productivity-based estimates is that agglomeration economies are not the only reason why average productivity may be higher in larger cities. As in Melitz and Ottaviano (2008) or Syverson (2004), the large number of firms in larger cities may make competition tougher, reducing markups and inducing less productive firms to exit. In this case, higher average productivity in larger cities could result from firm selection eliminating the least productive firms rather than from agglomeration economies boosting the productivity of all firms. Combes et al. (2012b) develop a framework to distinguish between agglomeration and firm selection. They nest a generalized version of the firm selection model of Melitz and Ottaviano (2008) and a simple model of agglomeration in the spirit of Fujita and Ogawa (1982) and Lucas and Rossi-Hansberg (2002). This nested model enables them to parameterize the relative importance of agglomeration and selection. The main prediction of their model is that, while selection and agglomeration effects both make average firm log productivity higher in larger cities, they have different predictions for how the shape of the log productivity distribution varies with city size. More specifically, stronger selection effects in larger cities, by excluding the least productive firms, should lead to a greater left truncation of the distribution of firm log productivities in larger cities. Stronger agglomeration effects, by making all firms more productive, should lead instead to a greater rightwards shift of the distribution of firm log productivities in larger cities. If firms that are more productive are also better at reaping the benefits of agglomeration, then agglomeration should lead not only to a rightwards shift but also to an increased dilation of the distribution of firm log productivities in larger cities.

[31] There are some clear limitations to this strategy. Changes in sectoral productivity are potentially determined simultaneously with changes in employment in the same sector. One may perhaps argue that employment adjusts only slowly after productivity shocks. This then calls for using high-frequency data but serial correlation is likely to be a major issue in this case.

Using a quantile approach that allows estimating a relative change in left truncation, shift, and dilation between two distributions and establishment-level data for France, Combes et al. (2012b) conclude that productivity differences across urban areas in France are mostly explained by agglomeration. They compare locations with above median employment density against those with below-median density (results are almost identical when comparing cities with population above or below 200,000). The distribution of firm log productivity in areas with above-median density is shifted to the right and dilated relative to areas below median density. On the other hand, they find no difference between denser and less dense areas in terms of left truncation of the log productivity distribution, indicating that firm selection is of similar importance in cities of different sizes. Their results show that firms in denser areas are thus on average about 9.7% more productive than in less dense areas. Put in terms of σ, this implies an elasticity of 0.032. However, the productivity boost of larger cities is greater for more productive firms, so the productivity gain is 14.4% for firms at the top quartile and only of 4.8% for firms at the bottom quartile.

For estimating the empirical magnitude of σ, an alternative to comparing establishments' productivity across cities is to compare workers' wages instead. As shown in Equation (5.42), from the point of view of workers, higher wages in larger cities are offset by higher house prices. Looking at the spatial equilibrium from the point of view of firms, Equation (5.39) shows that for firms to be willing to pay higher wages to produce in larger cities, there must be productive advantages that offset the higher costs. Thus, comparing wages across cities of different sizes also allows us to quantify the magnitude of agglomeration economies. This approach is used by Glaeser and Maré (2001), Combes et al. (2008), Combes et al. (2010) and De la Roca and Puga (2012), among others. A key concern when interpreting the existence of an earnings premium for workers with similar observable characteristics in larger cities is that there may be unobserved differences in worker ability across cities. Following Glaeser and Maré (2001), a standard way to tackle this concern is to use panel data for individual workers and introduce worker fixed effects. Compared with a simple pooled OLS regression, a fixed-effects regression reduces the estimate of σ by about one-half (Combes et al. 2010). This drop in the estimated elasticity when worker fixed-effects are introduced is sometimes interpreted as evidence of more productive workers sorting into bigger cities. However, De la Roca and Puga (2012) argue that the drop is mostly due to the existence of important learning advantages of larger cities. A pooled OLS regression mixes the static advantages from locating in a larger city, with the learning effects that build up over time as workers in larger cities are able to accumulate more valuable experience, with any possible sorting. Introducing worker fixed-effects makes the estimation of agglomeration economies be based exclusively on migrants, and captures the change in earnings they experience when they change location. This implies that an earnings regression with worker-fixed effects likely is expected to provide an accurate estimate of σ, capturing the static productive advantages of larger cities. Recent studies find the estimated value of σ thus estimated to be around 0.025 (Combes et al. 2010; De la Roca and Puga, 2012). At the same time, to more fully capture the benefits of larger cities, we should also study learning effects. We return to these below.

As we have seen, equilibrium and efficient city sizes are the result of a trade-off between agglomeration economies, as measured by σ, and urban crowding costs, as measured by γ. While there is now a large literature estimating the value of σ, the elasticity of urban productivity advantages with respect to city size, much less is known about γ, the elasticity of crowding costs with respect to city size. Combes et al. (2012a) develop a methodology to estimate this and apply it to French data. As highlighted by the monocentric city model studied in Section 5.2, house prices within each city vary with distance to the city center offsetting commuting costs. House prices at the city center capture the combined cost of housing and commuting in each city, so they are a relevant summary of urban costs. Combes et al. (2012a) use information about the location of parcels in each city and other parcel characteristics from recorded transactions of land parcels to estimate unit land prices at the center of each city. They then regress these estimated (log) prices at the center of each city on log city population to obtain an estimate of the elasticity of unit land prices at the center of each city with respect to city population: 0.72. Multiplying this by the share of land in housing (0.25) and then by the share of housing in expenditure (0.23), yields an elasticity of urban crowding costs with respect to population of 0.041.

Hence, existing empirical estimates suggest that the difference between the crowding costs elasticity γ and the agglomeration elasticity σ is small, perhaps 0.02 or less.[32] This has some interesting implications. On the one hand, optimal city sizes as given by Equation (5.49) should be highly sensitive to changes in agglomeration economies and productivity. On the other hand, mild deviations from optimal city sizes as described by Equation (5.49) should have only a small economic cost. This in turn means that it may be important to better account for migration costs when studying cities: with free mobility small productivity shocks may have large consequences for city sizes, whereas if mobility costs are important migration may only weakly respond to shocks, since the net effect from changes in agglomeration benefits and crowding costs achieved by moving may be small.

[32] Unfortunately, the empirical literature only provides estimates for an average agglomeration elasticity for all cities, not for city-specific agglomeration elasticities. There are sector-specific agglomeration elasticities available from the literature (e.g. Henderson, 2003) but they are subject to more serious identification concerns than agglomeration elasticities estimated at the city level since there is no good instrument for sectoral employment in cities. It is also unclear how elasticities for sectoral employment map into city-specific agglomeration elasticities given that most cities are far from being fully specialized.

5.6. HUMAN CAPITAL AND ENTREPRENEURSHIP

The models of cities considered so far are static. We have used comparative static results from those models to provide predictions about the effects of some manifestations of economic growth, such as better transportation or higher incomes, on the population and structure of cities. This unidirectional approach is valid if aggregate growth is not affected by the drivers of urban growth, as is arguably the case for urban amenities. However, the lack of feedback from cities to aggregate growth is questionable for the drivers of urban growth that we examine in this section: human capital and entrepreneurship. As discussed below, a good case can be made that human capital and entrepreneurship affect the growth of cities. Human capital and entrepreneurship are also arguably at the heart of the process of aggregate growth (Lucas, 1988; Aghion and Howitt, 1992). To explore two-way interactions between urban population growth and aggregate economic growth, dynamic models are needed.

As stressed in Section 5.5, a complete modeling of cities must include some form of agglomeration benefits. It is possible that agglomeration economies are static (i.e. take place in production) and affect the dynamics of aggregate growth only indirectly. It is also possible that agglomeration benefits are dynamic (i.e. take place in the accumulation of factors) and affect the dynamics of growth directly. In this section, we first explore a model in which agglomeration benefits are static but have dynamic implications before turning to dynamic benefits from agglomeration.

5.6.1 Human Capital and Urban Growth: Static Externalities

The model that follows draws from Duranton and Puga (2013) and captures key elements from Black and Henderson (1999). There are N_{it} workers in city i at time t. The output of each of these workers is:

$$y_{it} = BH_{it}^{\sigma} h_{it}^{\alpha} l_{it}^{1-\alpha}. \tag{5.51}$$

This production process offers constant returns at the individual level in the worker's human capital, h_{it}, and labor, l_{it}, but it is subject to a city-level externality in aggregate human capital, H_{it}^{σ}. Duranton and Puga (2013) develop micro-foundations for this production function in which the human capital externality arises by fostering entrepreneurship. Aggregate human capital in each city is the sum of the individual capital of its workers: $H_{it} = h_{it}N_{it}$. Each worker devotes a share δ of the unit of time that he or she has every period to accumulating human capital and a share $1 - \delta$ to working.[33] As a result of this investment, human capital evolves according to the following accumulation equation:

$$h_{it} - h_{it-1} = b\delta h_{it-1}. \tag{5.52}$$

[33] In Black and Henderson (1999), the share of time devoted to human capital accumulation is endogenous. As in much of the endogenous growth literature, it ends up being constant in steady state following intertemporal utility maximization by consumers with log-linear intertemporal preferences.

The parameter b measures the marginal return to the time devoted to human capital accumulation: $d(h_t/h_{t-1})/d\delta = b$. Note that human capital at time t needs to be a linear function of human capital at time $t-1$, as in Equation (5.52), for self-sustained but non-explosive growth to be possible.

We identify the accumulation factor h_{it} with human capital and model its accumulation accordingly in Equation (5.52) through a time investment made by individuals. Like Romer (1986), we could have labeled the accumulation factor physical capital instead. This would have made no difference to our modeling of production in Equation (5.51) but would have required a different accumulation process to replace Equation (5.52), since investment in physical capital is more appropriately modeled as foregone consumption measured in output, rather than foregone time spent learning. We prefer to focus on human capital given the rich literature providing evidence about human capital externalities in cities.[34]

Another possibility would be to identify the accumulation factor with knowledge, following Romer (1990). The accumulation Equation (5.52) would then be more appropriately modeled by describing firms conducting research and development. Successful innovators are rewarded with patents, while their innovation also increases a common stock of knowledge available to all, which in turn facilitates further innovations. While knowledge arguably plays an important role in long-run aggregate growth, using knowledge as accumulation factor in an urban context would force us to model its diffusion across cities to get non-trivial interactions between cities and aggregate growth. We return to this issue in the next section.

We model cities as in Section 5.5.1. This implies that the consumption of a worker living in city i is $c_{it} = y_{it} - \frac{\tau}{\gamma} N_i^{\gamma}$. Substituting Equation (5.51), $H_{it} = h_{it} N_{it}$ and $l_{it} = (1-\delta)$ into this expression, we can write per-capita consumption as:

$$c_{it} = B(1-\delta)^{1-\alpha} h_{it}^{\alpha+\sigma} N_{it}^{\sigma} - \frac{\tau}{\gamma} N_i^{\gamma}. \tag{5.53}$$

Since returns to human capital investments are the same everywhere, with perfect mobility across cities workers choose their city of residence at each period to maximize their present consumption. With profit-maximizing land developers, as in Section 5.5.1, equilibrium city sizes are optimal and are given by:

$$N_{it} = \left(B(1-\delta)^{1-\alpha} h_{it}^{\alpha+\sigma} \frac{\sigma}{\tau} \right)^{\frac{1}{\gamma-\sigma}}. \tag{5.54}$$

[34] With physical capital instead of human capital and a standard investment function where capital in t is equal to capital in $t-1$ minus depreciation plus foregone consumption, the production externality needs to be such that $\sigma = \alpha$ for self-sustained growth to be possible (Romer, 1986; Duranton and Puga, 2004). On the other hand, the accumulation equation no longer requires the linearity assumed in Equation (5.52). In any case, the results obtained from both sets of assumptions are qualitatively the same.

Note this expression, which maximizes c_{it} in Equation (5.53), is the same as Equation (5.49) from Section 5.5.1, with the productivity shifter B replaced by $B(1-\delta)^{1-\alpha}h_{it}^{\alpha+\sigma}$. In Section 5.5.1, we treated the productivity shifter B as an exogenous parameter to see how aggregate or sectoral shocks would affect cities. The term $B(1-\delta)^{1-\alpha}h_{it}^{\alpha+\sigma}$ is instead endogenous and driven by human capital accumulation. As workers become more productive through their accumulation of human capital, they find it worthwhile to agglomerate in larger cities. Hence, when economic growth takes the form of human capital accumulation, it leads to growing city sizes $\left(\frac{\mathrm{d}N_{it}}{\mathrm{d}h_{it}} > 0\right)$.

The relationship between human capital and growth does not stop here. The growth of cities, through agglomeration economies, amplifies the effects of human capital accumulation for aggregate growth. Following Duranton and Puga (2013), we can write the evolution of output per worker as:

$$\begin{aligned}\frac{y_{it}}{y_{it-1}} &= \left(\frac{h_{it}}{h_{it-1}}\right)^{\alpha+\sigma}\left(\frac{N_{it}}{N_{it-1}}\right)^{\sigma}\\ &= (1+b\delta)^{(\alpha+\sigma)\left(1+\frac{\sigma}{\gamma-\sigma}\right)}\\ &\approx 1+b\delta\frac{\gamma(\alpha+\sigma)}{\gamma-\sigma},\end{aligned} \tag{5.55}$$

where the first line of Equation (5.55) is obtained from Equation (5.51); the second line makes use of Equations (5.52) and (5.54); and the third provides a simple linear approximation when $b\delta$ is small. The last line of Equation (5.55) shows that in the absence of agglomeration economies ($\sigma = 0$) the growth rate of output is $b\delta\alpha$. With positive agglomeration economies ($\sigma > 0$), the growth rate of output per person is higher at $b\delta\frac{\gamma(\alpha+\sigma)}{\gamma-\sigma}$. We can compute the contribution of urban agglomeration to economic growth as:

$$\frac{b\delta\frac{\gamma(\alpha+\sigma)}{\gamma-\sigma} - b\delta\alpha}{b\delta\frac{\gamma(\alpha+\sigma)}{\gamma-\sigma}} = \frac{(\alpha+\gamma)\sigma}{(\alpha+\sigma)\gamma}. \tag{5.56}$$

This expression represents the increase in the growth rate as the result of urban agglomeration economies ($\sigma > 0$) relative to the total growth rate that appears in Equation (5.55). Empirically, recall from the discussion in Section 5.5.2 that estimates in the literature of σ, the agglomeration coefficient; and γ, the urban costs coefficient, are small. If we use our preferred estimates of $\sigma = 0.025$ and $\gamma = 0.04$, Equation (5.56) implies that cities account for 64% of aggregate growth.[35]

[35] The computation also requires assigning a value to α. The 64% figure is obtained from $\alpha = 0.5$, following the finding by Mankiw et al. (1992) of equal shares for labor and human capital in production. However, our results are not at all sensitive to this choice. With $\alpha = 0.7$, the contribution of urban agglomeration to aggregate growth is still 64%, with $\alpha = 0.3$ it is 65%. To a first approximation, the contribution of urban agglomeration to growth is σ/γ.

While this is a large number, we should keep in mind that we only consider growth from human capital accumulation and ignore other sources of growth such as physical capital accumulation and knowledge accumulation.[36] This nonetheless suggests that Lucas (1988) made an important point when he suggested looking at cities to understand the effects of human capital externalities. The large contribution of agglomeration to aggregate growth is also consistent with results from the human capital literature, which typically finds that external returns to human capital in cities are of about the same magnitude as private returns (e.g. Moretti, 2004a).

5.6.2 Human Capital and Urban Growth: Dynamic Externalities

We now turn to the modeling of dynamic agglomeration effects. As suggested by Alfred Marshall long ago: "The mysteries of trade become no mysteries; but they are as it were in the air, children learn many of them unconsciously. Good work is rightly appreciated, inventions and improvements in machinery, in process and the general organization of the business have their merits promptly discussed: if one man starts a new idea, it is taken up by others and combined with suggestions of their own; and thus becomes the source of further new ideas" (Marshall, 1890: iv.x.3). Several approaches have been developed to model these ideas. In an approach related to Black and Henderson (1999), Eaton and Eckstein (1997) adapt the Lucas (1988) model of human capital and growth to an urban context. To discuss their framework, let us start with a simple production function with no agglomeration effect. The output of a worker in city i is:

$$y_{it} = Bh_{it}^{\alpha} l_{it}^{1-\alpha}, \tag{5.57}$$

where, again, each worker devotes a share δ of her time to human capital accumulation, h_{it} is individual human capital and $l_{it} = 1 - \delta$ is individual labor. In contrast to Equation (5.51), (5.57) has no externality in production. This externality now appears in the accumulation equation. Thus, instead of an accumulation equation like (5.52), where each worker builds on his or her own human capital, Eaton and Eckstein (1997) assume that all residents of city i learn from the same aggregate knowledge base H_{it}:

$$h_{it} - h_{it-1} = bH_{it}\delta. \tag{5.58}$$

It may seem natural, as before, to think of the city's knowledge base as the sum of the human capital of all residents $H_{it} = h_{it}N_{it}$. However, having dynamic scale effects in Equation (5.58) would imply that cities of different population size experience different growth rates. Ultimately, the output of the entire economy would be dominated by that

[36] Davis et al. (2011) conduct a similar exercise within a neoclassical model of growth with physical capital and no human capital. They find a much smaller contribution of agglomeration to aggregate growth of about 10%.

of the largest city, where output per worker would grow increasingly faster than in other cities.

An alternative way to think about the city's knowledge base H_{it} would be to equate it with the average human capital in the city: $H_{it} = \overline{h}_{it}$. This raises three problems. The first is that city size no longer matters since production now only depends on the individual's human capital and this accumulates at a rate that does not depend on city size. If urban costs increase with a city's size, efficiency then calls for the smallest possible cities. So instead of having one city of exploding size, we have all cities disappear. The second issue is that the process of economic growth can take place in each city separately and independently. This is arguably counterfactual. A third problem arises when we introduce some heterogeneity in individual human capital levels. Because Equation (5.58) now implies that an individual's human capital increases more rapidly in cities with higher average human capital, this heterogeneity provides a strong incentive for sorting and leads again to faster growth in some cities.

To avoid these three problems, Eaton and Eckstein (1997) propose a more complicated production function with static agglomeration economies as in Equation (5.51). Although assuming agglomeration economies in production "solves" the problem created by the lack of scale effects, it means this is no longer a model with dynamic agglomeration economies. Agglomeration effects essentially remain static. In response to the second issue of each city being a separate economy able to generate self-sustaining growth alone, Eaton and Eckstein (1997) equate the city knowledge base with the weighted sum of the average human capital of other cities: $H_{it} = \sum_j \phi_{ij}\overline{h}_{jt}$ where the weights ϕ_{ij} may depend on the distance between cities. While this still allows cities to be isolated growing economies, this process of diffusion is intuitively appealing. Finally, the third problem of sorting is "solved" by considering *ex ante* identical workers and a steady state with symmetric growth in all cities so that workers remain identical.

The literature has followed two alternative strategies to reintroduce dynamic agglomeration economies without having one city dominate the entire urban system. The first is to limit how much can be learned by, for instance, imposing a finite lifetime as in Glaeser (1999). The second strategy is to model the diffusion of innovations as Duranton and Puga (2001). Let us summarize these two approaches.

In a model of skill transmission inspired by Jovanovic and Rob (1989), Jovanovic and Nyarko (1995), Glaeser (1999) formalizes the notion that the proximity to individuals with greater skills facilitates the acquisition of skills.[37] Glaeser (1999) considers overlapping generations of risk-neutral individuals who live for two periods (young and then old). Workers can be skilled or unskilled, and this affects their productivity; the output of an unskilled worker is lower than that of a skilled worker.

[37] This model is generalized and exposed more formally in Duranton and Puga (2004).

Each worker is born unskilled and chooses whether to spend her youth in the hinterland or in the city. In the hinterland, the cost of living is low but a worker remains unskilled. In the city, the cost of living is higher but a worker may become skilled after successfully meeting with an (old) skilled worker. The probability of a successful meeting increases with the number of skilled workers in the city. The surplus created by this successful acquisition of skills is split between the young apprentice and his or her old master. When old, workers chose whether to relocate. Old unskilled workers can no longer become skilled so, given the higher cost of living in the city, they always live in the hinterland. Old skilled workers, however, may offset the higher cost of living in the city with their share of the surplus created by teaching young apprentices.

Provided the benefits from becoming skilled are sufficiently large and provided the probability of meeting a skilled worker in the city is sufficiently high, there is a steady state in which young workers move to the city. Those that become skilled then stay in the city while those that do not become skilled go to the hinterland in their second period.

In a different model of learning, Duranton and Puga (2001) propose a diffusion mechanism where the benefits from learning in one city can be exploited in another.[38] In this model, an entrepreneur can introduce a new product by paying a fixed cost of entry. At first, entrepreneurs need a period of experimentation to realize their full potential—they may have a project, but may not know all the details of the product to be made, what components to use, or what kind of workers to hire. There are many possible ways to implement this project, but one is better than all others.

More specifically, entrepreneurs can choose between many production processes, each associated with a different set of inputs. The ideal production process, which differs across entrepreneurs, is initially unknown. An entrepreneur can try to discover his or her ideal production process by sampling at most one production process each period and using it for prototype production. As soon as an entrepreneur samples his or her ideal production process, he or she knows this is it and can start mass-production. A proportion of firms randomly exit every period to ensure that new firms keep entering and learning is never exhausted.

The use of a particular production process, either for prototype production or mass-production, requires physical proximity with the corresponding input producers. As in the model described in Section 5.5.1, input producers benefit from static agglomeration economies. The cost of using a given production process diminishes as more local firms use the same type of process because they can share intermediate suppliers. At the same time, relocating production across cities is costly, so entrepreneurs who have not yet discovered their ideal production process benefit from locating in a very diversified local economy to facilitate their learning. They would also like to face many suppliers for each

[38] In Duranton and Puga (2001), the diffusion of innovations relies explicitly on factor mobility. This differs from the literature in international trade that models diffusion mechanisms occurring through the trade of goods or, directly, through diffusion spillovers (e.g. Grossman and Helpman, 1991a).

set of inputs to enjoy lower costs. However, urban crowding places a limit on city size and consequently on how many processes can be widely used in a city.

Provided learning is important and moving costs are neither too high nor too low, an interesting equilibrium where both diversified and specialized cities arise endogenously can be sustained. It reconciles the needs for diversity and specialization along the life-cycle of firms. Entrepreneurs develop new products in cities with a diversified production structure. It allows them to sample easily and discover their ideal set of inputs. After discovering this ideal set of inputs, entrepreneurs are no longer interested in urban diversity. Because input producers in different sectors do not benefit from each other directly, industrial diversity makes cities more costly. As a result, entrepreneurs who have discovered their ideal set of inputs move away from a diversified city to a specialized city so that they can benefit from agglomeration effects in the production of those inputs. Moving costs cannot be too high for relocation to occur after learning, nor so low that an entrepreneur can easily learn by constantly relocating. Further, the gains from learning need to be high enough to justify the foregone static agglomeration economies in the early phases. In this sense, we can think of diversified cities as nursery cities where learning takes place and specialized cities as the places where the production of mature goods occurs.

The nursery cities model of Duranton and Puga (2001) proposes a theory of how innovation takes place and diffuses in space, while also matching observed patterns of firm relocations and a number of other facts about cities such as the coexistence of specialized and diversified cities (Duranton and Puga, 2000). It can also be used to explain why, even if innovation and learning concentrate in a few large and diverse cities, this does not imply that smaller and more specialized cities will disappear. Instead, the diffusion of innovations to exploit them in small specialized cities frees up large and diverse cities to concentrate in continuously feeding the growth process with new ideas.

5.6.3 Human Capital

Empirically, the strong association between city human capital and city population growth has been noted for some time. Glaeser et al. (1995), Simon and Nardinelli (1996), and Simon (1998) estimate regressions of the following form:

$$\Delta_{t+1,t} \log N_i = \beta_0 + \beta_1 \log N_{it} + \beta_2 h_{it} + X_{it}\beta_3' + \epsilon_{it}, \tag{5.59}$$

where the dependent variable is the change in log population or log employment between t and $t+1$ in city i. The explanatory variable of interest h_{it} is a measure of human capital at time t. Finally, X_{it} is a set of controls for other engines of growth, which often includes region dummies, and initial population is also controlled for. To measure human capital, early work used a range of education variables (e.g. Glaeser et al. 1995) or rough proxies (such as the number of business professionals in Simon and Nardinelli, 1996, for

19th century England). More recent work (e.g. Simon and Nardinelli, 2002; Glaeser and Saiz, 2004) prefers the share of university graduates since this more discriminant measure of human capital is usually associated with stronger effects.

Note that our growth model from Section 5.6.1 can be used to motivate this specification. Dividing Equation (5.54) valued at time $t+1$ from the same equation valued at time t, and taking logs, we obtain:

$$\Delta_{t+1,t} \log N_i = \frac{\alpha+\sigma}{\gamma-\sigma} \Delta_{t+1,t} h_{it}. \tag{5.60}$$

The main difference, leaving aside the controls and the error term, is that the theoretical Equation (5.60) relates changes in population to changes in human capital whereas the empirical specification (5.59) relates changes in population to initial levels of human capital. However, if we assume, as in the regression relating city growth to roads in Section 5.2, that population adjusts slowly to any changes in human capital, we end up with a regression of changes on initial levels instead with initial population as an additional control (see Equations (5.14) and (5.15)).

In a thorough investigation of the relationship between human capital and city growth across US metropolitan areas between 1970 and 2000, Glaeser and Saiz (2004) conclude that one standard deviation in the share of university graduates in a city's workforce is associated with a quarter of a standard deviation of population growth during the following decade. Put differently, for an average city, a 1% point higher share of university graduates is associated with around 0.5% population growth over the subsequent decade. This finding is representative of the findings in the rest of the literature.[39]

The strong association between human capital and city growth might be spurious for a number of reasons. For instance, more educated workers may be more mobile (or equivalently have stronger incentives to move) and, as a result, end up being over represented in fast-growing cities. Alternatively, the effect may be stronger than estimated. This would occur, for example, if cities with more stringent zoning restrictions, which experience slower population growth, also retain a more educated workforce.

To investigate these concerns and to show that the effect of human capital on city growth is most likely causal, Glaeser and Saiz (2004) perform a number of robustness checks. First, they show that education levels affect city growth even after controlling for a wide array of city characteristics. Second, they show that the relationship between education levels and city growth holds when looking only at variations within cities over time. That is, a given city tends to grow faster during periods when its population is more educated. This indicates that the relationship between human capital and city growth is not driven by unobserved permanent characteristics that make cities grow faster

[39] The main exception is Glaeser et al. (2011). They fail to find a positive association between human capital and subsequent county population growth in the eastern and central United States for a few decades in the last 200 years.

and attract more educated workers. Finally, to account for the possibility of a common determinant of both city growth and human capital, they use instrumental variables. To obtain an exogenous determinant of human capital in cities, they follow Moretti (2004a) and use the foundation of land grant colleges as an instrumental variable. Starting in 1862, land grant colleges were created in each state to foster agricultural and engineering education. They were usually placed in cities that were conveniently located (typically a central location in a state). Shapiro (2006) shows that these cities were not more educated before 1900 but gradually became more educated as the grant colleges developed, often turning into major universities. Glaeser and Saiz (2004), like Shapiro (2006), find that instrumenting city human capital by the presence of land grant colleges strongly suggests that the effect of education on city growth is causal and, if anything, leads to higher coefficients than indicated by the simple association in the data.

The literature has also provided less direct evidence about the role of human capital in city growth by investigating the channels through which it percolates. The model in Section 5.6.1 proposes some direct benefits in production occurring through human capital externalities in cities (see Equation (5.51)). The notion that smart, educated people benefit from being surrounded by other smart, educated people has received support in the literature. Following Rauch (1993), Moretti (2004a,b) finds robust evidence of large external effects of university education on city wages and productivity.

The human capital externalities of the model in Section 5.6.1 are micro-founded in Duranton and Puga (2013) through a link between human capital and entrepreneurship. Entrepreneurs may be over represented among more educated workers. If this is the case, a more educated city is also a more entrepreneurial city, where more new firms are created and existing firms grow faster. Stronger population growth then naturally follows. We explore empirical evidence of this channel in greater depth below. For now, we note that when attempting to disentangle between different channels through which human capital affects city growth, Glaeser and Saiz (2004) and Shapiro (2006) provide evidence that most of the effects of human capital percolate through a productivity channel, either learning and human capital externalities or entrepreneurship and firm growth. De la Roca and Puga (2012) explicitly study learning effects, using rich administrative data for Spain that tracks workers' full employment histories. They find that, by working in bigger cities, workers not only obtain an immediate static earnings premium, as in the model of Section 5.5.1, but are also able to accumulate more valuable experience, which increases their earnings faster. The additional value of experience accumulated in bigger cities persists even after workers move away and is even stronger for those with higher initial ability. This is evidence of the importance of learning in cities, providing support for the idea that cities foster the accumulation of human capital.

Higher productivity is not the only possible channel through which human capital can affect city growth. It could also be the case that more educated cities develop better amenities. These amenities are attractive to workers from other cities, particularly educated workers. Although Glaeser and Saiz (2004) and Shapiro (2006) only find modest support regarding the importance of amenities created by the presence of a skilled workforce, Diamond (2013) stresses this channel to explain the divergence in the skill composition of US cities in the last 30 years. The tension between these divergent findings will hopefully be resolved by future research.

A difficulty with human capital externalities and most forms of knowledge spillovers is that they are hard to track directly since they do not leave a paper trail. There is however one outcome of interactions that leaves some paper trail behind: innovations, when they are patented, contain citations to other patents. In their pioneering work, Jaffe et al. (1993) show a local bias in citation patterns. A patent is more likely to be cited by a subsequent patent for which the inventor lives in the same US metropolitan areas than by a "similar" patent for which the inventor lives in a different area. While this initial finding has been shown to be sensitive to what one means by "similar" and how one defines the control group for citing patents (Thompson and Fox-Kean, 2005), more recent work has established it on firmer grounds (Murata et al. 2013) and evidenced a host of other phenomena associated with knowledge spillovers in innovative activity. For instance, Agrawal et al. (2006) show that citations for a given patent are also disproportionately often more likely to occur in locations where the cited inventor was living prior to obtaining this patent. In other research, Kerr (2010) shows that, for a given technology, patenting growth in a city is stronger after a breakthrough innovation and that this growth differential is higher for technologies that depend more heavily on immigrant innovators, who are arguably more mobile. It is beyond the scope of this chapter to review this broad literature. We refer the reader instead to the survey of Carlino and Kerr (2013).

5.6.4 Entrepreneurship

To investigate the effect of agglomeration on city growth Glaeser et al. (1992) propose the following regression:

$$\Delta_{t+1,t} \log N_i^j = \beta_0 + \beta_1 Spec_{it}^j + \beta_2 Div_{it}^j + \beta_3 EstSize_{it}^j + X_{it}^j \beta_4' + \epsilon_{it}^j, \tag{5.61}$$

where the dependent variable is the change in log employment between t and $t+1$ in city i and sector j. The use of log employment as the dependent variable is motivated by a positive link from productivity growth to employment growth. The explanatory variables are a measure of initial specialization, $Spec_{it}^j$, a measure of sectoral diversity faced by sector k in city i, Div_{it}^j, a measure of establishment size, $EstSize_{it}^j$, and a set of other controls X_{it}^j, such as wages, the national growth of sector j during the same period, and initial employment in the city and sector.

The main results of Glaeser et al. (1992) are a negative coefficient on initial specialization, a negative coefficient on establishment size, and a strongly positive coefficient on diversity. The effects are quantitatively large. A standard deviation in specialization or diversity is associated with about 10% of a standard deviation in employment growth.

A standard deviation in establishment size is associated with nearly a quarter of a standard deviation in employment growth. These results have been subsequently replicated in many countries and generally confirmed. See for instance Combes (2000) for France or Cingano and Schivardi (2004) for Italy. An important qualification of these findings by Henderson et al. (1995) is that diversity appears to be particularly important for high-tech industries whereas specialization seems to play a positive role for mature industries. These results are consistent with those of the model of Duranton and Puga (2001) described above. In another important paper, Feldman and Audretsch (1999) use a measure of innovation instead of employment growth as dependent variable. They find a positive association between innovation and sectoral diversity (provided this diversity is relevant to the sector) and a negative association between innovation and specialization.

The regression described by Equation (5.61) does not directly tie into the model described in Section 5.6.1 nor into any of the frameworks described in Section 5.6.2. In their work, Glaeser et al. (1992) interpret the coefficients on specialization, diversity, and establishment size as dynamic externalities affecting local employment growth in sectors. In particular, the coefficient on average establishment size is interpreted as a competition effect (or even a Porter effect after Porter, 1990). This interpretation is far-fetched since there is no obvious mapping of establishment size into the toughness of competition. In many reasonable models of industrial organization, tougher competition actually leads to larger firms (Sutton, 1991). It may be more reasonable to think of $EstSize^j_{it}$ as a broad measure of entrepreneurship, since higher entrepreneurship will lead to more start-ups, which will generally be smaller in size than more mature firms.[40] This, in turn, is consistent with a suggestion initially made by Chinitz (1961) in his classic comparison of New York and Pittsburgh about the importance of small firms and entrepreneurship as a key determinant of the prosperity of cities. This would also be consistent with interpreting entrepreneurship as a form of human capital that would be particularly important in explaining the evolution of cities.

The regression described by Equation (5.61) suffers from another interpretation issue. It is hard to separate mean-reversion in employment caused by measurement error from the true effect of initial specialization since initial employment in the city and sector must be used to compute initial specialization.

A third problem of interpretation, noted by Combes et al. (2004) and Cingano and Schivardi (2004), is that the link between employment growth and productivity growth need not be positive. In a sector with constant markups, if the price elasticity of demand is larger than 1, an increase in productivity implies a higher revenue and an increase in employment. However, in sectors where demand is less elastic, the opposite holds. At the

[40] This then begs the question of whether establishment size is a good measure of entrepreneurship and more generally raises the legitimate question of how best to measure entrepreneurship. In the case of a regression, like (5.61), Glaeser and Kerr (2009) show that the results are the same with alternative measures of entrepreneurship such as the number of start-ups.

level of entire industries, fast productivity growth will often lead to declining employment (as illustrated by many traditional manufacturing industries where the ability to produce goods has risen much faster than demand). This could also occur for sectors within cities when goods are differentiated across cities. This does not mean that regression (5.61) cannot uncover the agglomeration determinants of urban growth. It simply suggests some caution when interpreting any positive effect of diversity, specialization, or establishment size. It need not be the case that diversity fosters productivity which in turn fosters employment growth. To explore this issue in more depth, Cingano and Schivardi (2004) suggest running the following regression:

$$\Delta_{t+1,t} \log \text{TFP}_i^j = \beta_0 + \beta_1 Spec_{it}^j + \beta_2 Div_{it}^j + \beta_3 EstSize_{it}^j + X_{it}^j \beta_4' + \epsilon_{it}^j, \tag{5.62}$$

which mirrors Equation (5.61) but uses growth in average firm-level total factor productivity in a city and industry instead of employment growth as dependent variable.

Interestingly, while the estimation of Equation (5.61) by Cingano and Schivardi (2004) generally confirms the findings of Glaeser et al. (1992), their estimation of Equation (5.62) yields a positive coefficient on specialization, an insignificant coefficient on diversity, and weak results regarding establishment size.[41] The difference in the sign of the coefficient on specialization is consistent with the intriguing possibility raised above: specialization may have strong effects on productivity and, because of inelastic demand, negative effects on employment.

A fourth issue is whether any effect of specialization, diversity, or establishment size can be interpreted as evidence of dynamic externalities. Dynamic externalities imply that the level of, say, establishment size, has an effect on the growth of employment. Static externalities, on the other hand, imply that establishment size measured in level has an effect on the level of employment. Put differently, with static externalities it is the first difference in establishment size which affects the growth rate of employment. To distinguish between static and dynamic effects, it would then seem natural to run the following regression:

$$\Delta_{t+1,t} \log N_i^j = \beta_0 + \beta_1 \Delta_{t+1,t} EstSize_i^j + \beta_2 EstSize_{it}^j + X_{it}^j \beta_4' + \epsilon_{it}^j. \tag{5.63}$$

A positive coefficient on establishment size would be consistent with dynamic externalities whereas a positive coefficient on the change in establishment size would be consistent with static externalities. This interpretation is problematic because a gradual adjustment

[41] Glaeser et al. (1992) also run a regression akin to (5.62) but use the change in log wage in cities and sectors as dependent variable instead of total factor productivity growth. They find tiny effect associated with their specialization variable and strong positive coefficients on initial employment in the city and sector. They also find a small positive coefficient on diversity and a negative coefficient on the number of establishments. To the extent that wages growth reflects productivity growth, these results are roughly consistent with those of Cingano and Schivardi (2004).

of employment following a change in $EstSize_{it}^{j}$ implies that even with only static externalities we could estimate a positive value for β_2. This is the same argument as with the gradual adjustment of population which follows on transportation improvements discussed above in Equations (5.14) and (5.15).

To improve on regression (5.63), a possibility is to estimate models that examine the dynamics of both the number of establishment and their size with perhaps a rich lag structure to assess how much and how fast past values of both variables affect their contemporaneous values. Combes et al. (2004) estimate the following type of autoregressive system:

$$\begin{cases} \Delta_{t+1,t} \log m_i^j & = \beta_0 + \beta_1 m_{it}^j + \beta_2 EstSize_{it}^j + X_{it}^j \beta_3' + \epsilon_{it}^j, \\ \Delta_{t+1,t} \log EstSize_i^j & = \beta_4 + \beta_5 m_{it}^j + \beta_6 EstSize_{it}^j + X_{it}^j \beta_7' + \varepsilon_{it}^j, \end{cases} \tag{5.64}$$

where m_i^j is the number of establishments in sector j and city i and t measures years. Relative to Equation (5.63), the system estimated in (5.64) decomposes the growth of employment in a city and industry into the growth in the number of establishments and the growth in their employment size. Combes et al. (2004) also estimate systems with longer and richer lag structures. They find that a shorter lag structure like the one in Equation (5.64) performs well. In turn, this suggests that the explanatory variables affect employment and establishment size fast. This is consistent with local externalities being static and not dynamic. They also find that the number of establishments is more sensitive to the local structure of economic activity than establishment size. This last result is consistent with the more recent finding of Glaeser and Kerr (2009) that much of local entrepreneurship can be explained by the presence of many small suppliers. Rosenthal and Strange (2010) also highlight the importance of small establishments and suggest that their benefits arise from the greater diversity of specialized suppliers that they provide to local firms.

While interesting and insightful, the work discussed so far does not solve the endogeneity of the key explanatory variables in these regressions. This problem has been neglected by the literature. This is perhaps because regressions like (5.61) use growth over a period as dependent variable and establishment size at the beginning of the period as explanatory variable. However, using a predetermined variable as explanatory variable in a regression does not guarantee its exogeneity. Local entrepreneurs could enter in large numbers in a city and sector if they foresee strong future demand. That expectations of future growth should trigger entry today is only natural. This is the nature of business.

Glaeser et al. (2010) examine whether the presence of many small firms in a city and sector is driven by the demand for entrepreneurship or by its supply. To the extent that they can be captured by higher sales per worker, demand factors do not appear to matter. Their findings point instead at the importance of the supply of entrepreneurship. This indirect approach, however, does not entirely solve the causality issue. To tackle it head on, Glaeser et al. (2012) take an instrumental variable approach. Returning to Chinitz's (1961) initial comparison of Pittsburgh and New York, they use the idea that cities closer to mines have been influenced by large mining firms. In turn, large firms are expected to reduce entrepreneurship by providing attractive employment opportunities for highly skilled workers. Large firms may also breed a local culture of company men which also reduces entrepreneurship. Indeed, proximity to historical mines is associated with larger establishments today even in completely unrelated sectors. Using this instrument, they estimate an even larger effect of entrepreneurship on city growth than the one measured directly from the data. Because a mining past can be associated with a general decline in manufacturing, Glaeser et al. (2012) replicate their main findings for cities outside the rust belt. These findings also hold when, instead of focusing on overall employment, they only look at service sectors only remotely tied to mining. Overall, these results are supportive of the notion that entrepreneurship is an important engine of city growth.

5.7. RANDOM URBAN GROWTH

In our exposition of random urban growth models, we do not proceed as above with first a theoretical model followed by a discussion of the empirics. Instead, it is convenient to start with a discussion of a key fact about the size distribution of cities before presenting statistical processes that can account for this fact. We then discuss recent attempts at grounding these statistical processes into economic models before returning to a discussion of empirical issues.

5.7.1 The Empirics of Zipf's Law

Since Auerbach (1913), the distribution of city sizes has often been approximated with a Pareto distribution. To do this, a popular way is to rank cities in a country from the largest to the smallest and regress the rank on city population N_i in the following manner:

$$\log \mathrm{Rank}_i = \beta_0 - \xi \log N_i + \epsilon_i. \tag{5.65}$$

The estimated coefficient ξ is the exponent, or shape parameter, of the Pareto distribution.[42] Zipf's law (after Zipf, 1949) corresponds to the statement that $\xi = 1$. This implies that the expected size of the second largest city is half the size of that of the largest, that of the third largest is a third of that of the largest, etc.[43]

[42] Regression (5.65) is not a standard regression. First, because the dependent variable is computed directly from the explanatory variable, measurement error on the "true" size also affects the rank and thus leads to a downward bias for the standard errors with OLS. In addition, when $\xi = 1$, the ratio of the largest to the second largest city is equal to 2 in expectations but its 95% confidence interval is 1 to 20. Put differently, the largest city is on average more than twice as large as the second largest city. This biases the OLS estimate of ξ with small samples. See Gabaix and Ibragimov (2011) for a simple and elegant solution to this problem. See also the excellent survey of Gabaix and Ioannides (2004).

[43] The deterministic reformulation of Zipf's law is usually referred to as the rank-size rule.

The empirical validity of Zipf's law is hotly debated. The classic cross-country assessment of Rosen and Resnick (1980) is ambiguous because their average Pareto exponent of 1.14 for 44 countries has been interpreted as evidence both for and against Zipf's law. Follow-up work by Soo (2005) broadly confirms the results of Rosen and Resnick (1980).

A lot of the debate has centered around the validity of Zipf's law for US cities. Using less than 200 US cities, Krugman (1996) and Gabaix (1999a) conclude at a near perfect fit. On the other hand, Black and Henderson (2003) and Eeckhout (2004) dismiss Zipf's law. Black and Henderson (2003) use data for metropolitan areas for the entire 20th century. They argue that the Pareto exponent is "far" from 1 at around 0.8 and that the linearity of the relationship between log size and log rank is questionable. Eeckhout (2004) uses data for US places and argues that their size distribution is better described by a log normal than by a Pareto distribution. Rozenfeld et al. (2011) use high-resolution data for the United States and aggregate settlements that are close to each other into cities. When defined from the bottom up, they find that Zipf's law holds very well for cities with population above 10,000. Giesen et al. (2010) argue that, for a number of countries, a distribution that is Pareto for both tails and log normal for its body (double Pareto log normal) provides a better fit to the data. In the same spirit, Ioannides and Skouras (2013) estimate a variety of log-normal and Pareto shapes allowing for some switching between them or a mixture of both. They highlight the importance of the excellent fit of the Pareto in the upper tail where most of the population lives and some fragility in the lower tail where the results depend on the definition of cities being used.

Stepping back from these seemingly contradictory claims, the empirical debate is mainly about three issues. The first is what constitutes a proper definition for cities. Ideally, this definition should be given by the model at hand. As made clear above, many urban models have commuting patterns at their core. Practically, this argues in favor of defining cities from commuting patterns. However, the notion of spatial continuity used by Rozenfeld et al. (2011) is also legitimate since urban models also imply that cities should be constituted of contiguous commercial and residential areas with agriculture beyond the urban fringe.[44]

[44] In practice, both types of definitions run into a number of problems. With commuting-based definitions, (sub-metropolitan) jurisdictions are aggregated to a given core when they send a minimum fraction of their workers to this core. The procedure is repeated until no jurisdiction remains to be aggregated to the resulting metropolitan area. However, these jurisdictions are themselves arbitrary (and sometimes extremely large in the west of the United States). The threshold of commuters is also arbitrary and the set of resulting metropolitan areas might be sensitive to this. Definitions based on spatial continuity also need to rely on some arbitrary level of distance with no development (or close to none) to separate metropolitan areas. For cities with green belt, spatial contiguity may also restrict the metropolitan area to be the area within the green belt when, in many cases, workers may commute from outside this green belt in large numbers. See Duranton (2013) for further discussion.

The second issue in this debate is about whether we observe a Pareto distribution. When distributions have the same number of parameters to be estimated, like with Pareto and log normal, they can be compared directly in terms of goodness of fit. This is nonetheless problematic because the goodness of fit may be different in different parts of the distribution. The Pareto distribution may offer a better fit in the upper tail whereas the log normal may fit the body of the distribution better. In addition, distributions often have different numbers of parameters. Distributions with more parameters are expected to provide mechanically a better fit. For instance, a mixture of Pareto and log normal is bound to do better than a simple Pareto or a simple log normal. The standard approach is then to rely on specification tests that weight the fit of a distribution relative to its number of parameters. The usefulness of this approach is questionable because the penalty associated with more parameters in those tests is arbitrary.

Even if one is willing to accept that city sizes are drawn from a Pareto distribution, the third issue is, whether the Pareto shape parameter is equal to 1 or not. This can readily be tested using standard levels of statistical significance relative to unity for $\hat{\xi}$ as estimated in regression (5.65). This approach is nonetheless debatable. With enough data points, one can always reject any sharp hypothesis like $\xi = 1$. In practice, the standard errors around ξ are fairly large even for urban systems with many cities so that it is hard to reject Zipf's law. Of course, it is also hard to reject distributions which are quite far from Zipf's law.[45]

In the end, the more relevant question is not so much whether the distribution of city sizes satisfies Zipf's law or not, but whether looking at this distribution through the lens of Zipf's law is useful. We believe it is, for two reasons. First, Zipf's law provides a reasonable first approximation, at least for the upper tail of the distribution. Second, because both the regularities of Zipf's law and the observed empirical deviations from it can be used to guide the modeling of economic processes underlying city size distributions (Gabaix, 1999a).[46]

[45] Gabaix and Ibragimov (2011) show that the standard error on ξ is asymptotically $\sqrt{2/n}\,\xi$ where n is the number of observations. With 100 cities, it is not possible to reject that $\hat{\xi} = 1.38$ statistically differs from unity at 5%. Even with 1000 cities, $\hat{\xi} = 1.09$ cannot be rejected as being different from unity.

[46] An alternative way to proceed is proposed in Duranton (2007) where the (non-Zipf) predictions of the model are measured directly against the empirical reality. This is in contrast with much of the extant literature, which often proposes a model that may or may not yield Zipf's law, compares it to this benchmark, and then in turn compares the benchmark to the empirical reality. Comparing the predictions of a model directly to the data is more straightforward and avoids the pitfalls mentioned above. However, this is not without problems either. Some of the results of a model may depend on a choice of auxiliary parameters about which not much is known. Consequently, too many degrees of freedom might be available for a meaningful assessment of what really matters for the model. There is also a risk of overextending conclusions reached based on a particular dataset or country that may not be representative of a broader tendency.

5.7.2 The Statistics of Zipf's Law

Let us now explore the statistical processes that lead to Zipf's law. There are two (related) avenues: multiplicative and additive processes.

Following Gabaix (1999a) and Gabaix (1999b), multiplicative processes have attracted a lot of attention. These processes are referred to as Kesten processes (after Kesten, 1973). We borrow from Gabaix and Ioannides (2004) and consider an economy where total population and the number of cities are both fixed. Between time $t-1$ and t, city i grows according to $N_{it} = (1+g_{it})N_{it-1}$. We impose Gibrat's law (after Gibrat, 1931): g_{it} is identically and independently distributed for every city with density $f(g)$. After T periods the size of city i is:

$$\begin{aligned}\log N_{iT} &= \log N_{i0} + \sum_{t=1}^{t=T} \log(1+g_{it}) \\ &\approx \log N_{i0} + \sum_{t=1}^{t=T} g_{it}.\end{aligned} \tag{5.66}$$

We note that the approximation in this equation holds only when the shocks are small enough. By the central limit theorem, over time $\log N_{iT}$ approaches a normal distribution and the distribution of N_{iT} thus becomes log normal. This distribution of city sizes does not admit a steady state and its variance keeps increasing.

To obtain a steady state, one needs to impose a lower bound to city sizes. This prevents cities from becoming too small. Let $M_t(N)$ denote the share of cities with population size N or higher at time t. This can be calculated as the share of cities that experience a growth rate g between time $t-1$ and time t from a size of at least N/g at $t-1$, aggregated over the different possible values of g:

$$M_t(N) = \int_0^{+\infty} M_{t-1}\left(\frac{N}{g}\right) f(g)\mathrm{d}g. \tag{5.67}$$

At the steady state (and it can be shown that there is one when cities cannot fall below a small threshold), we have:

$$M(N) = \int_0^{+\infty} M\left(\frac{N}{g}\right) f(g)\mathrm{d}g. \tag{5.68}$$

We can then verify that Zipf's law, that is $M(N) = a/N$ (where a is a constant), is the steady state we are looking for. Inserting this into Equation (5.68) implies:

$$\int_0^{+\infty} gf(g)\mathrm{d}g = 1, \tag{5.69}$$

which must hold since aggregate population is constant.[47]

[47] See Gabaix (1999a) for a complete proof. Note also that the same proof applies with non-constant total population if one normalizes city sizes to represent population shares instead of population numbers.

More intuitively, without a lower bound on city sizes, their distribution is single-peaked with thin tails at both ends. This is because very few cities consistently get positive or negative shocks. With a lower bound on city sizes, things change dramatically because the thin lower tail disappears and there is instead a maximum of the density function at the lower bound. Preventing cities from becoming too small also allows the upper tail to be fed by more cities. As a result, it is fatter. This lower bound also allows for the existence of a steady state instead of an ever-widening distribution.

The seemingly innocuous assumption of a lower bound on city sizes is enough to generate a very different outcome. Without a lower bound, city sizes follow a log normal distribution. With a lower bound, city sizes follow a Pareto distribution. This suggests a relative theoretical fragility of these statistical processes since the final outcome depends heavily on an auxiliary assumption that will be extremely hard to test. In turn, this puts some of the empirical debates about whether the size distribution of cities is best described by a Pareto or by a log normal back into perspective.

The main alternative to the multiplicative process just described are additive processes. The first was proposed by Simon (1955). In essence, Simon's model assumes that aggregate population grows over time by discrete increments. With some probability, a new lump goes to form a new city. Otherwise, it is added to an existing city. The probability that any particular city gets it is proportional to its population. This mechanism generates a Pareto distribution for city sizes. The Pareto exponent falls to one at the limit as the probability of new cities being created goes to zero.[48]

Despite important differences between them, multiplicative and additive processes both have some version of Gibrat's law at their core, either directly through multiplicative shocks or through increases of fixed size that occur proportionately to population.

5.7.3 The Economics of Zipf's Law

Among existing models of random growth with an economic content, that proposed by Eeckhout (2004) is the simplest. There is a continuum of cities. Labor is the only factor of production. There are aggregate decreasing returns at the city level, which are modeled through congestion costs that make output decrease with elasticity $-\gamma$, with respect to city size; and agglomeration economies that simultaneously make output increase with elasticity σ, with respect to city size, with $\sigma < \gamma$. In addition, city i experiences a labor productivity shock B_{it} at time t. Hence, output per worker in city i is $B_{it}N_{it}^{\sigma-\gamma}$, where N_{it} denotes its population at time t. We note that this modeling of cities differs from what we have used so far. In the trade-off between agglomeration and dispersion, dispersion forces always dominate here and optimal city size is zero. The assumption of a fixed number of cities then becomes crucial: if workers could move to new sites, cities would disappear, as larger cities only offer net disadvantages.

[48] For technical details, see the expositions of Krugman (1996) and Duranton (2006).

Free mobility then implies the equalization of output per worker across all cities. Even though each city faces shocks, the law of large numbers applies in aggregate so that output per worker is deterministic. After normalizing output per worker to unity, the equilibrium size of city i is given by:

$$N_{it} = B_{it}^{\frac{1}{\gamma-\sigma}}. \tag{5.70}$$

With small i.i.d. shocks, productivity evolves according to $B_{it} = (1 + g_{it})B_{it-1}$. It is easy to see that after T periods, we have:

$$\log N_{iT} \approx \log N_{i0} + \frac{1}{\gamma - \sigma} \sum_{t=1}^{t=T} g_{it}. \tag{5.71}$$

Equation (5.71) is derived in the same way as Equation (5.66). The main difference is that instead of imposing arbitrary population shocks, the model assumes cumulative productivity shocks. In a setting where free mobility implies that population is a power function of productivity (Equation (5.70)), the log normal distribution of city productivity maps into a log normal distribution of city population. Consistent with the argument made above, adding a reflexive lower bound for city size to Eeckhout's (2004) model would imply a Pareto distribution instead of a log normal distribution for city sizes.

The model of Rossi-Hansberg and Wright (2007) also relies on multiplicative and cumulative productivity shocks.[49] A key difference with Eeckhout (2004) is that shocks occur for an entire industry and there are no idiosyncratic productivity differences between cities. The other main difference between this model and other random urban growth models is that it explicitly treats cities as an equilibrium between agglomeration and dispersion forces. This is important since it shows that random growth models can accommodate a standard modeling of cities. In fact, we can write a version of Rossi-Hansberg and Wright's (2007) model simply by adding random shocks to the productivity shifter in the systems of cities model we developed in Section 5.5. If these shocks are multiplicative and cumulative, the productivity shifter in sector j evolves according to $B_{t+1}^{j} = (1 + g_{t+1}^{j})B_{t}^{j}$, where the shocks g_{t}^{j} are identically and independently distributed. Adding a time subscript to Equation (5.49), we can write optimal city size (and equilibrium city size in the presence of competitive developers) as:

$$N_{it} = \left(B_{t}^{j}\frac{\sigma}{\tau}\right)^{\frac{1}{\gamma-\sigma}}. \tag{5.72}$$

[49] Zipf's law is obtained in two cases by Rossi-Hansberg and Wright (2007). The first is the case described here with permanent industry shocks. The second is a situation with temporary shocks which affect factor accumulation. For alternative ways to generate Zipf's law with cumulative shocks see also Córdoba (2008).

Following the approach used to derive Equation (5.66), after T periods, we have:

$$\log N_{iT} \approx \log N_{i0} + \frac{1}{\gamma - \sigma} \sum_{t=1}^{t=T} g_t^j. \tag{5.73}$$

This is exactly as Equation (5.71), except that now the cumulative productivity term is sector-specific instead of city-specific. Thus, again, the distribution of city sizes is log normal. Adding a lower bound for productivity by sector leads N_{iT} to instead be Pareto distributed.

Despite the similarity of Equations (5.71) and (5.73), the underlying dynamics of Eeckhout (2004), Rossi-Hansberg and Wright (2007) are quite different. In Rossi-Hansberg and Wright (2007), utility is a concave function of city size and productivity shocks are common to all cities specializing in the same sector. By Equation (5.42), utility equalization across cities requires the value of output per worker net of commuting cost expenditures to be the same everywhere. Then, when a sector j experiences a small positive shock B_t^j, optimal size for cities specializing in that sector increases as a result. If all existing cities specializing in sector j increased their population to this new, larger size, the resulting increase in aggregate output in that sector would lower its price so that developers in cities specializing in sector j could not compete for residents until some developers exited and output prices rose again. At the new equilibrium, there will be fewer but larger cities specializing in sector j. Sectors that have received a sequence of higher productivity shocks, have a larger optimal city size and thus fewer cities. More precisely, the Pareto distribution of sectoral productivity maps directly into a Pareto distribution for optimal city sizes through Equation (5.73). We note that this Pareto outcome crucially relies on cities being of optimal size.

Gabaix (1999a) considers a model where workers are mobile only at the beginning of their life, when they need to pick a city. At time t population in city i is made up of the N_{it}^y young workers who choose to locate there and the fraction $1 - \delta$ of the previous period population who survive:

$$N_{it} = N_{it}^y + (1 - \delta) N_{it-1}. \tag{5.74}$$

Workers derive utility in a multiplicatively separable fashion from the consumption of a homogenous freely tradable good and from a local amenity:

$$u_{it} = A_{it} w_{it}. \tag{5.75}$$

The level of amenity in each city i, A_{it}, is independently drawn every period from a common distribution. This reduces the location choice for young workers to a static utility maximization problem. The production function is homogenous of degree one in young workers N_{it}^y and incumbent residents. For simplicity, assume a Cobb-Douglas functional form: $Y_{it} = (N_{it}^{y})^{\alpha}[(1-\delta)N_{it-1}]^{1-\alpha}$. This implies the following wage for young workers:

$$w_{it} = \alpha \left(\frac{(1-\delta)N_{it-1}}{N_{it}^{y}} \right)^{1-\alpha}. \tag{5.76}$$

The number of young workers going to each city in each period adjusts to equalize contemporaneous utility to some common level: $u_{it} = \overline{u}$. Combining this with Equations (5.74)–(5.76) yields the growth rate for city i between periods $t-1$ and t as:

$$g_{it} \equiv \frac{N_{it} - N_{it-1}}{N_{it-1}} = (1-\delta) \left(\frac{\alpha A_{it}}{\overline{u}} \right)^{\frac{1}{1-\alpha}} - \delta. \tag{5.77}$$

This growth rate is identically and independently distributed for all cities, regardless of their size. Since we are back to the evolution of city sizes given by Equation (5.66), city sizes follow a log normal distribution. With a reflective lower bound for city sizes, Zipf's law applies instead. There are two differences with the previous two models. First, the shocks apply to amenities and not to technology. Second, the shocks are temporary, not permanent. An interesting part of Gabaix's model is how temporary shocks have permanent effects. This arises because the wage of young workers depends only on the ratio of young mobile workers to immobile incumbents because of constant returns in production and because young workers become immobile after their original location choice.

The models of Gabaix (1999a), Eeckhout (2004), and Rossi-Hansberg and Wright (2007) are the three main multiplicative random growth models. Duranton (2006, 2007) proposes two related economic mechanisms that lead to additive random growth.

Duranton (2006) builds on Romer's (1990) endogenous growth model. Discrete innovations occur with probability proportional to research activity. Local spillovers make research activity in each location proportional to the number of local products. With mobile workers and no cost nor benefits from cities, the number of local products is proportional to population. Hence, in equilibrium, small discrete innovations occur in cities with probability proportional to their population size. Note that innovations need to be discrete to avoid the law of large numbers from applying, which would eliminate the randomness from the urban growth process. Innovations lead either to local production of the new product or, with some probability, to production at a new location where some required natural resource is available. Cities grow in population as a result of the increase in labor demand for producing a new product that follows an innovation. In essence, this model puts a geographical structure on a discrete version of Romer (1990). As shown by Duranton (2006), this model maps directly into Simon (1955) and generates Zipf's law as a limit case when the probability of new city formation tends to zero.[50]

[50] This modeling also avoids some pitfalls of Simon (1955) which converges slowly. The cumulative and exponential nature of the growth process in Romer (1990) ensures that shocks, although additive, occur more frequently as time passes, which leads to much faster convergence.

Duranton (2007) uses a related model which builds instead on the Schumpeterian growth model of Grossman and Helpman (1991b). In this framework, profit-driven research tries to develop the next generation of a product up a quality ladder. A success gives it a monopoly which lapses when the next innovation on the same product occurs. Products are discrete to ensure the necessary granularity for shocks to affect cities. Again, local spillovers tie research on a given product to the location of its production. The core of the model is that research might succeed in improving the products it seeks to improve (same-product innovation) or, sometimes, because of serendipity in the research process, it might succeed in improving another product (cross-product innovation).

With same-product innovation, the location of activity is unchanged by innovation and successful new innovators only replace incumbent producers in the same city. With cross-product innovation, the old version of the improved product stops being produced where it used to be and the new version starts being produced in the city where the innovation took place. This typically leads to a relocation of production with a population gain for the innovating city and a loss for the city of the incumbent producer.

To prevent cities from disappearing forever, the model also assumes that there is a core product in each city that cannot move. Symmetry and the absence of other costs and benefits from cities also ensure that city population is proportional to the number of products manufactured locally.

In steady state, this model does not quite lead to Zipf's law because the arrival of new products is not exactly proportional to city size. Because they already have more products, large cities have fewer of them to capture from elsewhere. On the other hand, the smallest cities with only one fixed product can only grow. Hence, growth is less than proportional to city size and this leads to a distribution of city sizes that is less skewed than Zipf's law. This somewhat lower expected growth at the upper end of the distribution is an empirically relevant feature of the US city size distribution (Ioannides and Overman, 2003). More generally, a calibration of the model does well at replicating the US city size distribution. Unlike other models of random growth, this model does not focus exclusively on the size distribution of cities. It also replicates the fast churning of industries across cities, a well-documented fact (Simon, 2004; Duranton, 2007; Findeisen and Suedekum, 2008).

5.7.4 The Tension Between Random Urban Growth Models and Other Models of Urban Growth

The main difference between random urban growth models and the classical urban growth models we considered in Sections 5.2–5.6 regards the role of shocks. In the latter approaches, urban growth is driven by city characteristics and what is left unexplained is treated as a residual. In random growth models, the residual is everything. Far from being a nice complementarity between two classes of models, this is a source of important tensions.

From a theoretical standpoint, it is possible to combine ingredients from traditional and random growth models to have urban growth driven by a combination of substantive determinants and random shocks. Following Duranton and Puga (2013), let us take our urban growth model of Section 5.6.1, where human capital accumulation drives aggregate growth and urban growth, and add sector-specific random shocks. If these are multiplicative and cumulative, the productivity shifter in sector j evolves according to $B_t^j = (1 + g_t^j)B_{t-1}^j$. From Equation (5.54), city size at time t is given by:

$$N_{it} = \left(\frac{\sigma}{\tau} B_t^j (1-\delta)^{1-\alpha} h_{it}^{\alpha+\sigma}\right)^{\frac{1}{\gamma-\sigma}} . \tag{5.78}$$

Dividing this equation valued at time T from the same equation valued at time 0, and taking logs, we obtain:

$$\log N_{iT} \approx \log N_{i0} + \frac{\alpha+\sigma}{\gamma-\sigma}(\log h_{iT} - \log h_{i0}) + \frac{1}{\gamma-\sigma}\sum_{t=1}^{t=T} g_t^j . \tag{5.79}$$

Now urban growth has both a systematic component arising from human capital accumulation and a random component arising from sectoral shocks. If we assume, as in Equation (5.52), that human capital grows at the same rate in every city, we have $\log h_{iT} - \log h_{i0} \approx Tb\delta$. Then, imposing a lower bound to sectoral productivity results in a Pareto distribution for city sizes. At the same time, human capital accumulation makes cities experience parallel growth in expectation.[51] With this specification, there is no theoretical incompatibility between classical and random urban growth models.

However, when the rate at which human capital accumulates depends on the level of human capital in the city, the growth of a city becomes a function of its initial human capital. Assume for instance that workers can choose how much human capital to accumulate. Because of complementarities in learning, it can be that workers optimally invest more in human capital accumulation in more educated cities so that the fraction of time spent learning δ is now an increasing function of city average human capital: $\delta(\bar{h}_{it})$. Then, by the same argument as above, $\log h_{iT} - \log h_{i0} \approx Tbf(\bar{h}_{i0})$. We no longer obtain Zipf's law because in Equation (5.79) the effect of this systematic driver of urban growth eventually dwarfs the cumulative effect of the sectoral shocks $\frac{1}{\gamma-\sigma}\sum_{t=1}^{t=T} g_t^j$.

To understand better the tension between classical and random urban growth models, consider a simple urban growth regression:

$$\Delta_{t+1,t} \log N_i = \beta_0 + \beta_1 \log N_{it} + D_{it}\beta_2' + \epsilon_{it}, \tag{5.80}$$

where the growth of city i between t and $t+1$ depends on its population size at time t, some drivers of urban growth D_{it}, and a random term ϵ_{it}. As a starting point, it is

[51] This is not perfectly parallel growth, since random shocks mean that expected growth rates are equal across all cities whereas actual growth rates are not.

useful to think of the classical urban growth models we considered in Sections 5.2–5.6 as focusing on N_{it} and D_{it}, whereas random growth models focus on ϵ_{it}. Formally, the question is whether Gibrat's law (and hence Zipf's law), as generated by random urban growth models, is compatible with $\beta_1 \neq 0$ or $\beta_2' \neq 0$ and whether these situations are empirically relevant.

It is best to discuss the issues surrounding initial population size ($\beta_1 \neq 0$) and those regarding systematic drivers of city growth ($\beta_2' \neq 0$) separately. Starting with initial population size we note that, while there is some disagreement in the literature about the importance of mean-reversion in city population data (e.g. Black and Henderson, 2003, vs. Eeckhout, 2004), past city population is more often than not a significant determinant of city growth and its coefficient often appears with a negative sign in urban growth regressions.[52]

A first source of mean-reversion could be found in measurement error. Taking the simplified version of Rossi-Hansberg and Wright (2007) presented above, true population growth in city between $t-1$ and t is given by the unobserved sectoral shock g_t^j. The level of population is nonetheless observed with error so that we observe $N_{it}e^{\mu_{it}}$ instead of N_{it} as population at time t. If μ_{it} is i.i.d., this has two implications. First, over two consecutive periods, there is a negative correlation between growth and initial size since, for instance, a large positive measurement shock in $t-1$ makes for both a higher initial population at $t-1$ and a lower growth rate between $t-1$ and t. Second, the observed growth rate is $\epsilon_{it} = g_{it} + \mu_{it} - \mu_{it-1}$. In turn, this implies:

$$\log N_{iT} = \log N_{i0} + \beta_0 + \mu_{iT} - \mu_{i0} + \sum_{t=1}^{t=T} g_{it}. \tag{5.81}$$

This equation is compatible with Equation (5.79). As argued by Gabaix and Ioannides (2004) if the tail of the summation in g is fatter than that of μ, Zipf's law should still occur in steady state. Intuitively, mean-reversion does not matter provided it is dominated by the cumulated Gibrat's shocks. A similar argument would hold if the population was not mismeasured but instead subjected to real temporary shocks around optimal city size. Hence, Zipf's law need not rely on a strong version of Gibrat's law where $\beta_1 = 0$. Instead, it can hold with a weaker version of Gibrat's law, where $\beta_1 = 0$. This said, much remains to be done on this issue. We need to know what is the weakest version of Gibrat's law compatible with Zipf's law. For instance, an AR (1) error structure like $\epsilon_{it} = g_{it} + \rho\epsilon_{it-1}$ does not converge to log normal for N_T without further (Gaussian) assumptions about g.[53]

[52] Black and Henderson (2003) find a highly significant coefficient for $\beta_1 = -0.02$ in the case of US cities across decades of the 20th century. Covering an even longer time period, both Glaeser et al. (2011) and Desmet and Rappaport (2013) also find significant departures from Gibrat's law.

[53] Such autoregressive processes are important in this context given the strong persistence of population shocks (Rappaport, 2004).

Turning to the other determinants of urban growth, let us return to Equation (5.80), assume for simplicity $\beta_1 = 0$, allow for β_2 to be time varying, and consider that $\epsilon_{it} = g_{it}$, which is i.i.d. After simplification, we obtain:

$$\log N_{iT} = \log N_{i0} + \beta_0 + \sum_{t=1}^{t=T} D_{it}\beta'_{2t} + \sum_{t=1}^{t=T} g_{it}. \tag{5.82}$$

This equation corresponds to the predictions of the model of Section 5.6.2 where the growth rate of human capital is not constant across cities but is instead driven by some city characteristic (e.g. the local presence of strong research universities). It is now easy to understand that any term $D_{it}\beta'_{2t} = D_i\beta'_2$ that is constant in magnitude over time and differs across cities would lead to divergence in the long run and a distribution that differs from Zipf's law. This suggests a major incompatibility between classical and random urban growth models.

There are a number of ways around this incompatibility. First, the upper tail of the city size distribution may remain Pareto despite different growth trends. To understand this point, consider two groups of cities, fast- and slow-growing cities. Provided the lower bound city size for each group of cities grows with its trend, there is a Pareto distribution emerging for each group of cities and divergence between the two groups. At any point in time the overall distribution will be a mixture of two Pareto distributions, both with a slope coefficient minus one. Above the largest of the two lower bounds, this distribution will be Pareto.[54]

Second, classical and random urban growth models are also compatible when the effects of $D_{it}\beta'_{2t}$ are short lived, that is, when there is mean-reversion in β_{2t} or in D_{it}. Mean-reversion in β_{2t} corresponds to the situation where a permanent characteristic has a positive effect over a period of time and negative effect over another. In the United States for instance, it is possible that hot summers deterred population growth before the development of air-conditioning but promoted it after this. Proximity to coal and iron was arguably a factor of growth during the late 19th and early 20th century that became irrelevant after. Glaeser et al. (2011) provide formal support for this argument looking at the growth of counties in the Eastern and Central United States over a 200-year period. They show that many determinants of county population growth such as geography and climate are not stable over time.

Mean-reversion in D_{it} corresponds instead to the situation where the determinants of growth are temporary in cities. For instance, it could be that receiving roads is a factor of urban growth as suggested by Duranton and Turner (2012) but that new roads are allocated

[54] Skouras (2009) considers a different but related argument. Among groups of cities with the same constant average size, any group that follows Zipf's law will eventually dominate the upper tail.

proportionately to population.[55] In a slightly different vein, a number of papers highlight the importance of specific one-off technology shocks that affect urban growth. Duranton and Puga (2005) emphasize the availability of communication technologies allowing firms to separate their management from their production activities leading cities to specialize by function and no longer by sector (see also Ioannides et al. 2008, for another take on the effect of communication technologies on cities). Desmet and Rossi-Hansberg (2009) focus on the maturation of economic activities which are concentrated when new and gradually diffuse as they mature. Under some conditions, a series of one-off shocks like these may be able to bridge the gap between classical and random urban growth models.

More generally, what growth regressions and classical urban models treat as explanatory variables in some cases need to be thought of as the shocks in random growth models. This observation suggests that shocks in the context random growth models need not be equated with residuals in urban growth regressions.

Different time horizons between classical and random growth models may go a long way toward making them compatible with each other. Classical urban growth models, which constitute the theoretical underpinning of standard urban growth regressions, may be looking at the growth of cities around a particular period whereas random growth models may have a much longer time horizon. In that case, classical urban growth models help us uncover short run proximate factors of urban growth whereas random growth models help us understand the fundamental mechanics that drive urban growth in the long run.

Two further possibilities can be entertained to reconcile random and classical urban growth models. The first is that there might be a number of city characteristics distributed such that the effect of the entire vector of characteristics is about the same in all cities. In that case, the underlying trend for all cities would be the same and Zipf's law would occur in steady state. While an exact equalization across the effects of all characteristics across cities would be highly unlikely, some negative correlations across drivers of urban growth are not unthinkable.[56]

The second possibility is that Zipf's law may occur as the outcome of a static model while parallel city growth occurs for entirely unrelated reasons. Hsu (2012) proposes a microfounded model of central place theory which can generate Zipf's law. Lee and Li (2013) propose a model where city population depends multiplicatively on their many characteristics which are i.i.d. This leads to the static counterpart of Eeckhout's (2004)

[55] Duranton and Turner (2012) show that the 1947 plan which guided the early development of the US interstate highway system allocated highways to cities on average proportionately to their population. More recent highway developments are clearly less than proportional to population.

[56] For instance, cities with nice landscapes are also likely to suffer from greater construction costs and more generally a greater scarcity of usable land.

model. Behrens et al. (2012) also obtain Zipf's law in a model of sorting across cities.[57] These papers are interesting because they show that Zipf's law need not be the outcome of a random growth model but could arise for other reasons. Nevertheless, these static Zipf's law models imply strong restriction on urban growth since parallel growth is needed to retain Zipf's law. The recent empirical findings of Desmet and Rappaport (2013) are consistent with this type of argument. They find that the US city size distribution first settled to its current form and only then began to satisfy a mild form of parallel growth (through Gibrat's law).

Finally, it should be kept in mind that random growth models mainly offer theories of the growth of individual cities, not theories of the growth of all cities. For instance, random growth models have little to say about the increase in average city size over the last 200 years. Classical urban growth models propose both theories of the growth of all cities as well as theories of the growth of particular cities. Even if random growth models turned out to be a good explanation of urban evolutions, that would not prevent better and cheaper commuting technologies to be one important driver of the growth of all cities.

5.8. CONCLUSION

We have identified four key drivers of the population growth of cities in developed economies. First, transportation and housing supply. Second, amenities. Third, agglomeration effects, in particular those related to human capital and entrepreneurship. And fourth, technology and shocks to specific cities or industries.

The empirical case for these drivers rests first on cross-city growth regressions. Identifying causal factors in regressions of city population growth in cross-section is fraught with difficulties. Applications of this type of methodology to cross-country growth in income per capita have rightfully come under attack in the past (Durlauf et al. 2005). The exercise is arguably easier in the case of city population since there is less heterogeneity in the data. For instance, data on educational achievement is more directly comparable between Baltimore and Miami than between Belarus and Malawi. The number of explanatory factors for city population growth within a country is also much smaller than the number of possible causes of income growth across countries since many variables can be, as a first approximation, held constant within a unified country. In the last decade, the literature on city growth has also repeatedly tackled fundamental inference concerns heads on, relying in particular on instrumental variables.

[57] In addition, Henderson and Venables (2009) can generate Zipf's law from an underlying power law in site quality. In a very different model, Berliant and Watanabe (2009) generate empirically relevant size distributions. In their model, cities receive shocks by industries and only the best will produce. This leads to city sizes being determined by extreme value distributions which can be parameterized to fit existing distributions. A detailed analysis of static Zipf's law models is outside our scope here.

The literature on drivers of city growth nicely ties into the main modeling approaches used to study the economics of cities the monocentric model for housing and transportation; the model of cross-city compensating differentials for amenities; models of microfounded agglomeration economies for agglomeration effects; and human capital and random growth models for technology and sectoral shocks. Hence, the literature reviewed in this chapter goes beyond isolating specific drivers of city growth. It also provides empirical support for the core theoretical models of cities.

In the work reviewed above, close links between theory and empirics have turned out to be very useful. They allow going beyond the estimation of the elasticity of city growth with respect to a specific driver to examine other implications of these theories. For instance, in monocentric models of cities lower transportation costs imply not only population growth but also greater suburbanization, increased land consumption, etc. Many of these extra predictions have been examined in the literature and receive strong empirical support.

This said, the success of this literature is only partial and much remains to be done. We identify several areas of interest for future work. First, most of the theories we relied on are static and only offer predictions based on comparative statics. Related to this, many results depend crucially on workers being homogenous and perfectly mobile. Dynamic urban models with heterogeneous agents and explicit mobility costs should be a key priority for theory. This will provide new insights into the evolution of cities and help us consider adjustment processes explicitly. In turn, this will hopefully lead to new empirical approaches that push the study of urban dynamics beyond cross-city growth regressions and avoid the ambiguities that mar the interpretation of many results in the literature.

Furthermore, some potential drivers of the growth of cities are yet to be explored. The biggest gap is arguably studying the effects of municipal and city governments, local policies, and public finance. In addition, many empirical results should be strengthened and alternative empirical strategies developed to confirm them. Although a convincing empirical framework that examines all existing drivers of urban growth at the same time is too ambitious a goal, exploring drivers of urban growth in isolation is not satisfactory. Some explanations need to be confronted. For instance, the links between human capital and entrepreneurship need to be clarified. Also, both infrastructure and amenities drive city growth but infrastructure-rich places are often amenity poor and vice versa. Engines of city growth might substitute for one another or instead, perhaps, complement each other. Understanding the relationships between drivers of urban growth is of academic interest but it could also be highly relevant to design urban growth strategies.

As argued in the Introduction, the growth of cities potentially offers a unique window into the broader issue of the determinants of economic growth and technological progress. This is where the results have been least satisfactory. Little in the study of the growth of cities so far has really illuminated how growth and technological progress take place. For instance, as made clear in this chapter, there is good evidence that average education in cities has a causal effect on their subsequent population growth. While this is important and interesting to urban economists, providing useful insights for growth economists will require convincing evidence about a much more detailed causal chain looking into how innovation takes place in cities, how workers learn from each other, and how knowledge diffuses between workers.

ACKNOWLEDGMENTS

We thank Frédéric Robert-Nicoud, Esteban Rossi-Hansberg, and conference and seminar participants at the Paris School of Economics and the 2011 North American Meeting of the Regional Science Association for comments and discussions. This research was developed in part when Duranton was visiting the Paris School of Economics, whose financial support is gratefully acknowledged. Puga gratefully acknowledges funding from the European Community's Seventh Framework Programme, under ERC Advanced Grant agreement 269869.

REFERENCES

Abdel-Rahman, Hesham M., Fujita, Masahisa, 1990. Product variety, Marshallian externalities, and city sizes. Journal of Regional Science 30 (2), 165–183.

Aghion, Philippe, Howitt, Peter, 1992. A model of growth through creative destruction. Econometrica 60 (2), 323–351.

Agrawal, Ajay, Cockburn, Iain, McHale, John, 2006. Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships. Journal of Economic Geography 6 (5), 571–591.

Albouy, David, 2008. Are big cities really bad places to live? Improving quality-of-life estimates across cities. Working Paper 14472, National Bureau of Economic Research.

Alonso, William, 1964. Location and Land Use; Toward a General Theory of Land Rent. Harvard University Press, Cambridge, MA.

Altonji, Joseph G., David Card, 1991. The effects of immigration on the labor market outcomes of less-skilled natives. In: John M. Abowd, Richard B. Freeman (Eds.), Immigration, Trade and the Labor Market. Chicago University Press, Chicago, IL, pp. 201–234.

Anas, Alex, Arnott, Richard, Small, Kenneth A., 1998. Urban spatial structure. Journal of Economic Literature 36 (3), 1426–1464.

Angrist, Joshua D., Pischke, Jörn-Steffen, 2008. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton.

Arnott, Richard J., Stiglitz, Joseph E., 1979. Aggregate land rents, expenditure on public goods, and optimal city size. Quarterly Journal of Economics 93 (4), 471–500.

Arnott, Richard J., Stiglitz, Joseph E., 1981. Aggregate land rents and aggregate transport costs. Economic Journal 91 (362), 331–347.

Auerbach, Felix, 1913. Das Gesetz der Bevölkerungskonzentration. Petermanns Geographische Mitteilungen 59, 73–76.

Bairoch, Paul, 1988. Cities and Economic Development: From the Dawn of History to the Present. University of Chicago Press, Chicago.

Baldwin, Richard E., 2001. Core-periphery model with forward-looking expectations. Regional Science and Urban Economics 31 (1), 21–49.

Barro, Robert J., 1991. Economic-growth in a cross-section of countries. Quarterly Journal of Economics 106 (2), 407–443.

Bartik, Timothy, 1991. Who Benefits from State and Local Economic Development Policies? W. E. Upjohn Institute for Employment Research, Kalamazoo MI.

Baum-Snow, Nathaniel, 2007. Did highways cause suburbanization? Quarterly Journal of Economics 122 (2), 775–805.

Baum-Snow, Nathaniel, Lutz, Byron F., 2011. School desegregation, school choice, and changes in residential location patterns by race. American Economic Review 101 (7), 3019–3046.

Becker, Randy, Vernon Henderson, J., 2000. Intra-industry specialization and urban development. In: Huriot, Jean-Marie, Thisse, Jacques-François (Eds.), Economics of Cities: Theoretical Perspectives. Cambridge University Press, Cambridge, pp. 138–166.

Behrens, Kristian, Duranton, Gilles, Robert-Nicoud, Frédéric, 2012. Productive cities: Sorting, selection, and agglomeration. Processed. Wharton School, University of Pennsylvania.

Berliant, Marcus, Hiroki Watanabe, 2009. Explaining the size distribution of cities: X-treme economies. Processed. Washington University in St. Louis.

Black, Duncan, Vernon Henderson, J., 1999. A theory of urban growth. Journal of Political Economy 107 (2), 252–284.

Black, Duncan, Henderson, Vernon, 2003. Urban evolution in the USA. Journal of Economic Geography 3 (4), 343–372.

Blomquist, Glenn C., Berger, Mark C., Hoehn, John P., 1988. New estimates of quality of life in urban areas. American Economic Review 78 (1), 89–107.

Boustan, Leah Platt, 2010. Was postwar suburbanization white flight? Evidence from the black migration. Quarterly Journal of Economics 125 (1), 417–443.

Brainard, Lael S., 1997. An empirical assessment of the proximity-concentration trade-off between multinational sales and trade. American Economic Review 87 (4), 520–544.

Brueckner, Jan K., Rosenthal, Stuart S., 2009. Gentrification and neighborhood housing cycles: will America's future downtowns be rich? Review of Economics and Statistics 91 (4), 725–743.

Burchfield, Marcy, Overman, Henry G., Puga, Diego, Turner, Matthew A., 2006. Causes of sprawl: a portrait from space. Quarterly Journal of Economics 121 (2), 587–633.

Card, David, 2001. Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. Journal of Labor Economics 19 (1), 22–64.

Carlino, Gerald A., Kerr, William R., 2013. Agglomeration and innovation. Processed. Harvard University.

Carlino, Gerald A., Albert Saiz, 2008. City beautiful. Working Paper 08–22, Federal Reserve Bank of Philadelphia.

Cheshire, Paul C., Magrini, Stefano, 2006. Population growth in European cities: Weather matters - but only nationally. Regional Studies 40 (1), 23–37.

Chinitz, Benjamin, 1961. Contrasts in agglomeration: New York and Pittsburgh. American Economic Review Papers and Proceedings 51 (2), 279–289.

Ciccone, Antonio, Hall, Robert E., 1996. Productivity and the density of economic activity. American Economic Review 86 (1), 54–70.

Cingano, Federico, Schivardi, Fabiano, 2004. Identifying the sources of local productivity growth. Journal of the European Economic Association 2 (4), 720–742.

Combes, Pierre-Philippe, 2000. Economic structure and local growth: France, 1984–1993. Journal of Urban Economics 47 (3), 329–355.

Combes, Pierre-Philippe, Duranton, Gilles, Gobillon, Laurent, 2008. Spatial wage disparities: Sorting matters! Journal of Urban Economics 63 (2), 723–742.

Combes, Pierre-Philippe, Duranton, Gilles, Gobillon, Laurent, 2012a. The costs of agglomeration: Land prices in French cities. Processed. University of Pennsylvania.

Combes, Pierre-Philippe, Duranton, Gilles, Gobillon, Laurent, Puga, Diego, Roux, Sébastien, 2012b. The productivity advantages of large cities: Distinguishing agglomeration from firm selection. Econometrica 80 (6), 2543–2594.

Combes, Pierre-Philippe, Gilles Duranton, Laurent Gobillon, Sébastien Roux, 2010. Estimating agglomeration effects with history, geology, and worker fixed-effects. In: Edward L. Glaeser (Ed.), Agglomeration Economics. Chicago University Press, Chicago, IL, pp. 15–65.

Combes, Pierre-Philippe, Magnac, Thierry, Robin, Jean-Marc, 2004. The dynamics of local employment in France. Journal of Urban Economics 56 (2), 217–243.

Córdoba, Juan-Carlos, 2008. On the distribution of city sizes. Journal of Urban Economics 63 (1), 177–197.

Cuberes, David, 2011. Sequential city growth: empirical evidence. Journal of Urban Economics 69 (2), 229–239.

Cullen, Julie Berry, Levitt, Stephen D., 1999. Crime, urban flight, and the consequences for cities. Review of Economics and Statistics 81 (2), 159–169.

Davis, Morris, Fisher, Jonas D.M., Whited, Toni M., 2011. Macroeconomic implications of agglomeration. Processed. University of Wisconsin.

De la Roca, Jorge, Diego Puga, 2012. Learning by working in big cities. Processed, CEMFI.

de Vries, Jan, 1984. European Urbanization: 1500–1800. Methuen, London.

Desmet, Klaus, Jordan Rappaport, 2013. The settlement of the United States, 1800 to 2000: The long transition towards Gibrat's law. Discussion Paper 9353, Centre for Economic Policy Research.

Desmet, Klaus, Rossi-Hansberg, Esteban, 2009. Spatial growth and industry age. Journal of Economic Theory 144 (6), 2477–2502.

Diamond, Rebecca, 2013. The determinants and welfare implications of US workers' diverging location choices by skill: 1980–2000. Processed. Harvard University.

Dixit, Avinash K., Stiglitz, Joseph E., 1977. Monopolistic competition and optimum product diversity. American Economic Review 67 (3), 297–308.

Duby, George. 1981–1983. Histoire de la France Urbaine. Le Seuil, Paris.

Duranton, Gilles, 2006. Some foundations for Zipf's law: product proliferation and local spillovers. Regional Science and Urban Economics 36 (4), 542–563.

Duranton, Gilles, 2007. Urban evolutions: the fast, the slow, and the still. American Economic Review 97 (1), 197–221.

Duranton, Gilles, 2013. Delineating metropolitan areas: Measuring spatial labour market networks through commuting patterns. Processed. Wharton School, University of Pennsylvania.

Duranton, Gilles, Morrow, Peter M., Turner, Matthew A., 2013. Roads and trade: evidence from the US. Processed. University of Toronto.

Duranton, Gilles, Puga, Diego, 2000. Diversity and specialisation in cities: why, where and when does it matter? Urban Studies 37 (3), 533–555.

Duranton, Gilles, Puga, Diego, 2001. Nursery cities: Urban diversity, process innovation, and the life cycle of products. American Economic Review 91 (5), 1454–1477.

Duranton, Gilles, Puga, Diego, 2004. Micro-foundations of urban agglomeration economies. In: Henderson, Vernon, Thisse, Jacques-François (Eds.), Handbook of Regional and Urban Economics, vol 4. North-Holland, Amsterdam, 2063–2117.

Duranton, Gilles, Puga, Diego, 2005. From sectoral to functional urban specialisation. Journal of Urban Economics 57 (2), 343–370.

Duranton, Gilles, Puga, Diego, 2013. Urban growth: systematic, idiosyncratic and random determinants and their aggregate implications. Processed. Wharton School, University of Pennsylvania.

Duranton, Gilles, Turner, Matthew A., 2011. The fundamental law of road congestion: Evidence from US cities. American Economic Review 101 (6), 2616–2652.

Duranton, Gilles, Turner, Matthew A., 2012. Urban growth and transportation. Review of Economic Studies 79 (4), 1407–1440.

Durlauf, Steven N., Johnson, Paul A., Jonathan, R.W., Temple., 2005. Growth econometrics. In: Aghion, Philippe, Durlauf, Steven N. (Eds.), Handbook of Economic Growth, vol 1. North-Holland, Amsterdam, 555–677.

Eaton, Jonathan, Eckstein, Zvi, 1997. Cities and growth: theory and evidence from France and Japan. Regional Science and Urban Economics 27 (4–5), 443–474.

Eeckhout, Jan, 2004. Gibrat's law for (All) cities. American Economic Review 94 (5), 1429–1451.

Ethier, Wilfred J., 1982. National and international returns to scale in the modern theory of international trade. American Economic Review 72 (3), 389–405.

Feldman, Maryann P., Audretsch, David B., 1999. Innovation in cities: Science-based diversity, specialization and localized competition. European Economic Review 43 (2), 409–429.

Findeisen, Sebastian, Suedekum, Jens, 2008. Industry churning and the evolution of cities: evidence for Germany. Journal of Urban Economics 64 (2), 326–339.

Fischel, William A., 2000. Zoning and land use regulations. In: Boudewijn, Bouckaert, De Geest, Gerrit (Eds.), Encyclopedia of Law and Economics, vol 2. Edward Elgar, Cheltenham, 403–442.

Flatters, Frank, Vernon Henderson, J., Mieszkowski, Peter, 1974. Public goods, efficiency, and regional fiscal equalization. Journal of Public Economics 3 (2), 99–112.

Fujita, Masahisa, 1988. A monopolistic competition model of spatial agglomeration: a differentiated product approach. Regional Science and Urban Economics 18 (1), 87–124.

Fujita, Masahisa, 1989. Urban Economic Theory: Land Use and City Size. Cambridge University Press, Cambridge.

Fujita, Masahisa, Ogawa, Hideaki, 1982. Multiple equilibria and structural transition of non-monocentric urban configurations. Regional Science and Urban Economics 12 (2), 161–196.

Fujita, Masahisa, Thisse, Jacques-François, 2002. Economics of Agglomeration: Cities, Industrial Location, and Regional Growth. Cambridge University Press, Cambridge.

Gabaix, Xavier, 1999a. Zipf's law for cities: an explanation. Quarterly Journal of Economics 114 (3), 739–767.

Gabaix, Xavier, 1999b. Zipf's law and the growth of cities. American Economic Review Papers and Proceedings 89 (2), 129–132.

Gabaix, Xavier, Ibragimov, Rustam, 2011. Rank-1/2: a simple way to improve the OLS estimation of tail exponents. Journal of Business Economics and Statistics 29 (1), 24–39.

Gabaix, Xavier, Ioannides, Yannis M., 2004. The evolution of city size distributions. In: Henderson, Vernon, Thisse, Jacques-François (Eds.), Handbook of Regional and Urban Economics, vol 4. North-Holland, Amsterdam, 2341–2378.

Gibrat, Robert, 1931. Les inégalités économiques; applications: aux inégalités des richesses, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle, la loi de l'effet proportionnel. Paris: Librairie du Recueil Sirey.

Giesen, Kristian, Zimmermann, Arndt, Suedekum, Jens, 2010. The size distribution across all cities — Double Pareto lognormal strikes. Journal of Urban Economics 68 (2), 129–137.

Glaeser, Edward L., 1999. Learning in cities. Journal of Urban Economics 46 (2), 254–277.

Glaeser, Edward L., Gyourko, Joseph, 2005. Urban decline and durable housing. Journal of Political Economy 113 (2), 345–375.

Glaeser, Edward L., Gyourko, Joseph, Saks, Raven, 2005. Why is Manhattan so expensive? Regulation and the rise in housing prices. Journal of Law and Economics 48 (2), 331–369.

Glaeser, Edward L., Gyourko, Joseph, Saks, Raven E., 2006. Urban growth and housing supply. Journal of Economic Geography 6 (1), 71–89.

Glaeser, Edward L., Matthew Kahn., 2001. Decentralized employment and the transformation of the American city. Brookings-Wharton Papers on Urban Affairs:1–47.

Glaeser, Edward L., Kahn, Matthew E., 2004. Sprawl and urban growth. In: Henderson, Vernon, Thisse, Jacques-François (Eds.), Handbook of Regional and Urban Economics, vol 4. North-Holland, Amsterdam, 2481–2527.

Glaeser, Edward L., Kahn, Matthew E., Rappaport, Jordan, 2008. Why do the poor live in cities? The role of public transportation. Journal of Urban Economics 63 (1), 1–24.

Glaeser, Edward L., Kallal, Heidi, Scheinkman, José A., Schleifer, Andrei, 1992. Growth in cities. Journal of Political Economy 100 (6), 1126–1152.

Glaeser, Edward L., Kerr, William R., 2009. Local industrial conditions and entrepreneurship: How much of the spatial distribution can we explain? Journal of Economics and Management Strategy 18 (3), 623–663.

Glaeser, Edward L., Kerr, William R., Giacomo, A.M., Ponzetto., 2010. Clusters of entrepreneurship. Journal of Urban Economics 67 (1), 150–168.

Glaeser, Edward L., Kolko, Jed, Saiz, Albert, 2001. Consumer city. Journal of Economic Geography 1 (1), 27–50.

Glaeser, Edward L., Maré, David C., 2001. Cities and skills. Journal of Labor Economics 19 (2), 316–342.

Glaeser, Edward L., Kerr, Sari Pekkala, Kerr, William R., 2012. Entrepreneurship and urban growth: An empirical assessment with historical mines. Processed. Harvard University.

Glaeser, Edward L., Giacomo, A.M., Ponzetto, Kristina Tobio., 2011. Cities, skills, and regional change. Processed. Harvard University.

Glaeser, Edward L., Saiz, Albert, 2004. The rise of the skilled city. Brookings-Wharton Papers on Urban Affairs 5, 47–95.

Glaeser, Edward L., Scheinkman, José A., Shleifer, Andrei, 1995. Economic-growth in a cross-section of cities. Journal of Monetary Economics 36 (1), 117–143.

Glaeser, Edward L., Tobio, Kristina, 2008. The rise of the sunbelt. Southern Economic Journal 74 (3), 610–643.

Glaeser, Edward L., Ward, Bryce A., 2009. The causes and consequences of land use regulation: evidence from Greater Boston. Journal of Urban Economics 65 (3), 265–278.

Greenstone, Michael, Hornbeck, Richard, Moretti, Enrico, 2010. Identifying agglomeration spillovers: evidence from winners and losers of large plant openings. Journal of Political Economy 118 (3), 536–598.

Grossman, Gene M., Elhanan Helpman., 1991a. Innovation and Growth in the World Economy. MIT Press, Cambridge, MA.

Grossman, Gene M., Helpman, Elhanan, 1991b. Quality ladders in the theory of growth. Review of Economic Studies 58 (1), 43–61.

Gyourko, Joseph, Mayer, Christopher, Sinai, Todd, forthcoming. Superstar cities. American Economic Journal, Economic Policy.

Gyourko, Joseph, Saiz, Albert, Summers, Anita A., 2008. A new measure of the local regulatory environment for housing markets: The Wharton residential land use regulatory index. Urban Studies 45 (3), 693–729.

Helpman, Elhanan. 1998. The size of regions. In: David Pines, Efraim Sadka, Itzhak Zilcha (Eds.), Topics in Public Economics. Cambridge University Press, New York, NY, pp. 33–54.

Helsley, Robert W., Strange, William C., 1990. Matching and agglomeration economies in a system of cities. Regional Science and Urban Economics 20 (2), 189–212.

Henderson, J. Vernon, 1974. The sizes and types of cities. American Economic Review 64 (4), 640–656.

Henderson, J. Vernon, 2003. Marshall's scale economies. Journal of Urban Economics 53 (1), 1–28.

Henderson, J. Vernon, 2005. Urbanization and growth. In: Aghion, Philippe, Durlauf, Steven N. (Eds.), Handbook of Economic Growth, vol 1B. North-Holland, Amsterdam, pp. 1543–1591.

Henderson, J. Vernon, Kuncoro, Ari, Turner, Matt, 1995. Industrial development in cities. Journal of Political Economy 103 (5), 1067–1090.

Henderson, J. Vernon, Venables, Anthony J., 2009. The dynamics of city formation. Review of Economic Dynamics 39 (2), 233–254.

Henderson, J. Vernon, Wang, Hyoung Gun, 2007. Urbanization and city growth: the role of institutions. Regional Science and Urban Economics 37 (3), 283–313.

Hilber, Christian, Robert-Nicoud, Frédéric, 2013. On the origins of land use regulations: theory and evidence from us metro areas. Journal of Urban Economics 75 (1), 29–43.

Holmes, Thomas J., 1998. The effect of state policies on the location of manufacturing: Evidence from state borders. Journal of Political Economy 106 (4), 667–705.

Hsu, Wen-Tai, 2012. Central place theory and city size distribution. Economic Journal 122 (563), 903–932.

Ioannides, Yannis, Skouras, Spyros, 2013. US city size distribution: Robustly Pareto, but only in the tail. Journal of Urban Economics 73 (1), 18–29.

Ioannides, Yannis M., Overman, Henry G., 2003. Zipf's law for cities: an empirical examination. Regional Science and Urban Economics 33 (2), 127–137.

Ioannides, Yannis M., Overman, Henry G., Rossi-Hansberg, Esteban, Schmidheiny, Kurt, 2008. The effect of ICT on urban structure. Economic Policy 23 (54), 201–242.

Jaffe, Adam B., Trajtenberg, Manuel, Henderson, Rebecca, 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. Quarterly Journal of Economics 108 (3), 577–598.

Jovanovic, Boyan, Nyarko, Yaw, 1995. The transfer of human capital. Journal of Economic Dynamics and Control 19 (5–7), 1033–1064.

Jovanovic, Boyan, Rob, Rafael, 1989. The growth and diffusion of knowledge. Review of Economic Studies 56 (4), 569–582.

Kerr, William R., 2010. Breakthrough inventions and migrating clusters of innovation. Journal of Urban Economics 67 (1), 46–60.

Kesten, Harry, 1973. Random difference equations and renewal theory for products of random matrices. Acta Mathematica 131 (1), 207–248.

Krugman, Paul, 1996. Confronting the mystery of urban hierarchy. Journal of the Japanese and International Economies 10 (4), 1120–1171.

Krugman, Paul R., 1980. Scale economies, product differentiation, and the pattern of trade. American Economic Review 70 (5), 950–959.

Leamer, Edward E., Levinsohn, James, 1995. International trade theory: the evidence. In: Grossman, Gene M., Rogoff, Kenneth (Eds.), Handbook of International Economics, vol 3. North-Holland, Amsterdam, pp. 1339–1394.

Lee, Sanghoon, Qiang Li, forthcoming. Uneven landscapes and the city size distribution. Journal of Urban Economics.

LeRoy, Stephen F., Sonstelie, Jon, 1983. The effects of urban spatial structure on travel demand in the United States. Journal of Urban Economics 13, 67–89.

Lucas, Robert Jr., E., 1988. On the mechanics of economic development. Journal of Monetary Economics 22 (1), 3–42.

Lucas, Robert Jr., E., Esteban Rossi-Hansberg., 2002. On the internal structure of cities. Econometrica 70 (4), 1445–1476.

Mankiw, N.Gregory, David, David Romer, Weil, N., 1992. A contribution to the empirics of economic growth. Quarterly Journal of Economics 107 (2), 407–437.

Marshall, Alfred, 1890. Principles of Economics. Macmillan, London.

McMillen, Daniel P., 2001. Nonparametric employment subcenter indentification. Journal of Urban Economics 50 (3), 448–473.

McMillen, Daniel P., 2006. Testing for monocentricity. In: Arnott, Richard J., McMillen, Daniel P. (Eds.), A Companion to Urban Economics. Blackwell, Oxford, pp. 128–140.

McMillen, Daniel P., Smith, Stefani C., 2003. The number of subcenters in large urban areas. Journal of Urban Economics 53 (3), 332–342.

Melitz, Marc, Gianmarco, I.P., Ottaviano., 2008. Market size, trade and productivity. Review of Economic Studies 75 (1), 295–316.

Mills, Edwin S., 1967. An aggregative model of resource allocation in a metropolitan area. American Economic Review Papers and Proceedings 57 (2), 197–210.

Moretti, Enrico, 2004a. Estimating the social return to higher education: Evidence from longitudinal and repeated cross-sectional data. Journal of Econometrics 121 (1), 175–212.

Moretti, Enrico, 2004b. Workers' education, spillovers, and productivity: Evidence from plant-level production functions. American Economic Review 94 (3), 656–690.

Moretti, Enrico, 2011. Local labor markets. In: Ashenfelter, Orley, Card, David (Eds.), Handbook of Labor Economics, vol 4. Elsevier, Amsterdam, 1237–1313.

Murata, Yasusada, Nakajima, Ryo, Okamoto, Ryosuke, Tamura, Ryuichi, 2013. Localized knowledge spillovers and patent citations: A distance-based approach. Processed. Nihon University.

Muth, Richard F., 1969. Cities and Housing. University of Chicago Press, Chicago.

Ortalo-Magné, François, Prat, Andrea, 2010. The political economy of housing supply. University of Wisconsin, Processed.

Ottaviano, Gianmarco I.P., Peri, Giovanni, 2006. The economic value of cultural diversity: evidence from US cities. Journal of Economic Geography 6 (1), 9–44.

Porter, Michael, 1990. The Competitive Advantage of Nations. Free Press, New York.

Puga, Diego, 2010. The magnitude and causes of agglomeration economies. Journal of Regional Science 50 (1), 203–219.

Rappaport, Jordan, 2004. Why are population flows so persistent? Journal of Urban Economics 56 (3), 554–580.

Rappaport, Jordan, 2007. Moving to nice weather. Regional Science and Urban Economics 37 (3), 375–398.

Rauch, James E., 1993. Productivity gains from geographic concentration of human-capital - evidence from the cities. Journal of Urban Economics 34 (3), 380–400.

Redding, Stephen J., Sturm, Daniel M., 2008. The costs of remoteness: Evidence from German division and reunification. Journal of International Economics 98 (5), 1766–1797.

Roback, Jennifer, 1982. Wages, rents, and the quality of life. Journal of Political Economy 90 (6), 1257–1278.

Romer, Paul M., 1986. Increasing returns and long-run growth. Journal of Political Economy 94 (5), 1002–1037.

Romer, Paul M., 1990. Endogenous technological-change. Journal of Political Economy 98 (5), S71–S102.

Rosen, Kenneth T., Resnick, Mitchel, 1980. The size distribution of cities—an examination of the Pareto law and primacy. Journal of Urban Economics 8 (2), 165–186.

Rosen, Sherwin. 1979. Wage-based indexes of urban quality of life. In: Peter N. Miezkowski, Mahlon R. Straszheim (Eds.), Current Issues in Urban Economics. Johns Hopkins University Press, Baltimore, MD, pp. 74–104.

Rosenthal, Stuart S., Strange, William, 2004. Evidence on the nature and sources of agglomeration economies. In: Henderson, Vernon, Thisse, Jacques-François (Eds.), Handbook of Regional and Urban Economics, vol 4. North-Holland, Amsterdam, pp. 2119–2171.

Rosenthal, Stuart S., Strange, W. C., 2010. Small establishments/big effects: Agglomeration, industrial organization and entrepreneurship. In: Edward L. Glaeser (Eds.), Agglomeration Economics. Chicago University Press, Chicago, IL, pp. 277–302.

Rossi-Hansberg, Esteban, Mark, L.J., Wright., 2007. Urban structure and growth. Review of Economic Studies 74 (2), 597–624.

Rozenfeld, Hernán D., Rybski, Diego, Gabaix, Xavier, Maske, Hernán A., 2011. The area and population of cities: New insights from a different perspective on cities. American Economic Review 101 (5), 2205–2225.

Saiz, Albert, 2010. The geographic determinants of housing supply. Quarterly Journal of Economics 125 (3), 1253–1296.

Serck-Hanssen, Jan. 1969. The optimal number of factories in a spatial market. In: Bos, Hendricus C. (Ed.), Towards Balanced International Growth. North-Holland, Amsterdam, pp. 269–282.

Shapiro, Jesse M., 2006. Smart cities: Quality of life, productivity, and the growth effects of human capital. Review of Economics and Statistics 88 (2), 324–335.

Simon, Curtis J., 1998. Human capital and metropolitan employment growth. Journal of Urban Economics 43 (2), 223–243.

Simon, Curtis J., 2004. Industrial reallocation across US cities, 1977–1997. Journal of Urban Economics 56 (1), 119–143.

Simon, Curtis J., Nardinelli, Clark, 1996. The talk of the town: Human capital, information, and the growth of English cities, 1861 to 1961. Explorations in Economic History 33 (3), 384–413.

Simon, Curtis J., Nardinelli, Clark, 2002. Human capital and the rise of American cities: 1900–1990. Regional Science and Urban Economics 32 (1), 59–96.

Simon, Herbert, 1955. On a class of skew distribution functions. Biometrika 42 (2), 425–440.

Skouras, Spyros, 2009. Explaining Zipf's law for US cities. Processed. Athens University of Economics and Business.

Soo, Kwok Tong, 2005. Zipf's law for cities: A cross country investigation. Regional Science and Urban Economics 35 (3), 239–263.

Starrett, David A., 1974. Principles of optimal location in a large homogeneous area. Journal of Economic Theory 9 (4), 418–448.

Stiglitz, Joseph E., 1977. The theory of local public goods. In: Feldstein, Martin S., Inman, Robert P. (Eds.), The Economics of Public Services. MacMillan Press, London, pp. 274–333.

Storper, Michael, Scott, Allen J., 2009. Rethinking human capital, creativity and urban growth. Journal of Economic Geography 9 (2), 47–167.

Sutton, John. 1991. Sunk Costs and Market Structure. The MIT Press, Cambridge, MA.

Sveikauskas, Leo, 1975. Productivity of cities. Quarterly Journal of Economics 89 (3), 393–413.

Syverson, Chad, 2004. Market structure and productivity: A concrete example. Journal of Political Economy 112 (6), 1181–1222.

Thompson, Peter, Fox-Kean, Melanie, 2005. Patent citations and the geography of knowledge spillovers: A reassessment. American Economic Review 95 (1), 450–460.

Thünen, Johann H., von, 1826. Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie. Hamburg: Perthes. English Translation: The Isolated State, Pergammon Press, Oxford, 1966.

Vickrey, William S., 1977. The city as a firm. In: Feldstein, Martin S., Inman, Robert P. (Eds.), The Economics of Public Services. MacMillan Press, London, pp. 334–343.

Zipf, George Kingsley, 1949. Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology. Addison Wesley, Cambridge.

CHAPTER SIX

Growth and Structural Transformation

Berthold Herrendorf[*], **Richard Rogerson**[†]
and Ákos Valentinyi[‡]
[*]Department of Economics, Arizona State University, Tempe, AZ 85287, USA
[†]Princeton University & NBER, Princeton, USA
[‡]Cardiff Business School, IE-CERSHAS & CEPR, UK

Abstract

Structural transformation refers to the reallocation of economic activity across the broad sectors agriculture, manufacturing, and services. This review article synthesizes and evaluates recent advances in the research on structural transformation. We begin by presenting the stylized facts of structural transformation across time and space. We then develop a multi-sector extension of the one-sector growth model that encompasses the main existing theories of structural transformation. We argue that this multi-sector model serves as a natural benchmark to study structural transformation and that it is able to account for many salient features of structural transformation. We also argue that this multi-sector model delivers new and sharper insights for understanding economic development, regional income convergence, aggregate productivity trends, hours worked, business cycles, wage inequality, and greenhouse gas emissions. We conclude by suggesting several directions for future research on structural transformation.

Keywords

Approximate balanced growth, Multi-sector growth model, Structural transformation, Stylized facts

JEL Classification Codes

O11, O14, O4

6.1. INTRODUCTION

The one-sector growth model has become the workhorse of modern macroeconomics. The popularity of the one-sector growth model is at least partly due to the fact that it captures in a minimalist fashion the essence of modern economic growth, which Kuznets (1973), in his Nobel Prize lecture described as the sustained increase in productivity and living standards. By virtue of being a minimalist structure, the one-sector growth model necessarily abstracts from several features of the process of economic growth. One of these is the process of structural transformation, that is, the reallocation of economic activity across the broad sectors agriculture, manufacturing, and services. Kuznets listed structural transformation as one of the six main features of modern economic growth. Structural transformation has also received a lot of attention in the policy debate of

Handbook of Economic Growth, Volume 2B
ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53540-5.00006-9

developed countries where various observers have claimed that the sectoral reallocation of economic activity is inefficient, and calls for government intervention. Understanding whether structural transformation arises as an efficient equilibrium outcome requires enriching the one-sector growth model to incorporate multiple sectors. More generally, this raises the question whether incorporating multiple sectors will sharpen or expand the insights that can be obtained from the one-sector growth model. Several researchers have recently begun to tackle these questions, and the objective of this chapter is to synthesize and evaluate their efforts.[1]

The first step in the broad line of research on structural transformation is to develop extensions of the one-sector growth model that are consistent with the stylized facts of structural transformation. Accordingly, we begin this chapter by presenting the stylized facts of structural transformation and then develop a multi-sector extension of the growth model that serves as a natural benchmark model to address the issue of structural transformation. Given the prominent role attributed to theories of balanced growth in the literature using the one-sector growth model, we start by asking whether it is possible to simultaneously deliver structural transformation and balanced growth. Recent work has identified several versions of the growth model that achieve this. We present the results of this work in the context of our benchmark multi-sector model.

It turns out that the conditions under which one can simultaneously generate balanced growth and structural transformation are rather strict, and that under these conditions the multi-sector model is not able to account for the broad set of empirical regularities that characterize structural transformation. We therefore argue that the literature on structural transformation has possibly placed too much attention on requiring exact balanced growth, and that it would be better served by settling for approximate balanced growth instead. Put somewhat differently, we think that progress in building better models of structural transformation will come from focusing on the forces behind structural transformation without insisting on exact balanced growth. While the corresponding efforts to uncover the forces behind structural transformation are relatively recent, we describe some headway that has been made. We argue that the recent work suggests that the benchmark multi-sector model with approximate balanced growth is able to account for many salient features of structural transformation for the US, both qualitatively and quantitatively.

Armed with an extension of the one-sector growth model that incorporates structural transformation in an empirically reasonable fashion, we seek to answer the question of whether modeling structural transformation indeed delivers new or sharper insights into issues of interest. We argue that the answer to this question is yes, and we present several specific examples from the literature to illustrate this. These examples have in common that taking into account changes in the sectoral composition of the economy

[1] A different aspect of structural transformation that Kuznets also noted is the movement of the population from rural into urban areas, which is typically accompanied by the movement of employment out of agriculture.

is crucial for understanding a variety of changes in aggregate outcomes. As we will see, this applies to important issues concerning economic development, regional income convergence, aggregate productivity trends, hours worked, business cycles, wage inequality, and greenhouse gas emissions.[2]

6.2. THE STYLIZED FACTS OF STRUCTURAL TRANSFORMATION

As mentioned in the introduction, structural transformation is defined as the reallocation of economic activity across three broad sectors (agriculture, manufacturing, and services) that accompanies the process of modern economic growth.[3] In this section, we present the stylized facts of structural transformation. While a sizeable literature on the topic already exists, including the notable early contributions of Clark (1957), Chenery (1960), Kuznets (1966), and Syrquin (1988),[4] we think that improvements in the quality of previous data and the appearance of new data sets make it worthwhile for us to summarize the current state of evidence.

Because the process of structural transformation continues throughout development, it is desirable to document its properties using relatively long time series for individual countries. The early studies that we cited above attempted to do this. However, the authors of these studies typically had to piece together data from various sources, necessarily creating issues about the comparability of their results across time and countries. In addition, the time period for which data was available was still relatively short. Recent efforts by various researchers to reconstruct historical data have increased the availability of appropriate long time series data for the purposes of documenting structural transformation. Although one still has to piece together data from different sources to generate long time series for most countries, time coverage has improved and compatibility is much less of a problem than it was in the past. We are going to use the Historical National Accounts Database of the University of Groningen as our primary historical data source, which we complement with several other data sources to increase the length of the periods covered.[5]

[2] Matsuyama (2008) and Ray (2010) also review the literature on structural transformation (or structural change, as Ray calls it). In contrast to them, we devote a large part of our review to documenting the stylized facts of structural transformation and to assessing whether multi-sector extensions of the standard growth model can account for them. Greenwood and Seshadri (2005) review the literature on economic transformation, which refers to broader issues than structural transformation, for example changes in the patterns of fertility and the movement of women out of the household into the labor market.

[3] We follow much of the literature and use the term manufacturing in this context to refer to all activity that falls outside of agriculture and services. It might seem to be more appropriate to refer to this category as industry, because manufacturing is just the largest component of it, but we prefer to reserve the term industry to refer to a generic production category.

[4] The list of subsequent papers is too large for us to attempt to include it in its entirety.

[5] Appendix A contains a detailed description about the historical data sources that we use. Many of them are also underlying the recent historical studies by Dennis and Iscan (2009) about structural transformation

While it is conceptually desirable to examine changes for individual countries over long time series, and there is now more opportunity to do so, limiting attention to individual countries narrows the perspective unnecessarily. To begin with, it effectively restricts the set of countries that can be studied to those that are currently rich, and so it leaves open the question of whether currently poor countries show the same regularities that currently rich countries showed when they were poor a century or two ago. Limiting attention to long time series data has the additional disadvantage that despite major improvements in constructing historical time series, they typically do not reach the quality of the best data sets for recent years. Therefore, we document the stylized facts of structural transformation also for five data sets that cover a relatively large set of developing and developed countries during the last 30 or so years: the Benchmark Studies of the International Comparisons Program as reported by the Penn World Table (PWT), EU KLEMS, the National Accounts from the United Nations Statistics Division, the OECD Consumption Expenditure Data, and the World Development Indicators (WDI).[6]

6.2.1 Measures of Structural Transformation

Before presenting any data, it is useful to briefly note some aspects of measuring economic development and structural transformation.

The two most common measures of economic development at the aggregate level are GDP per capita and some measure of productivity (typically GDP per worker or GDP per hour, depending upon data availability), each expressed in international dollars. While these two measures often coincide, there are cases in which they differ. For example, several European economies have similar values of GDP per hour as the US, but GDP per capita can be as much as 25% lower than in the US because hours per adult are much lower. Without knowing the exact context of the issue being addressed, it is unclear whether one should categorize these European countries as equally or less developed than the United States.

Having raised this issue, in this chapter we choose to always measure the level of development by GDP per capita in 1990 international dollars. Three reasons motivate this choice. First, in order to be able to identify threshold effects and the like, we insist on the comparability of the GDP numbers across different data sets, and GDP per capita is the only measure that is available for most countries and most of the time. Second, the standard models of structural transformation take labor supply to be exogenous, implying that they abstract from differences in hours worked. Third, since some of the models that we will consider emphasize the role of income effects for structural transformation, it seems appropriate to characterize the patterns of sectoral reallocation conditional on

in the United States and by Alvarez-Cuadrado and Poschke (2011) about structural transformation in 12 industrialized countries including the United States.

[6] We again refer the reader to Appendix A for details regarding the data sets and how we use them to construct measures of economic activity at the sector level.

income. Irrespective of these three reasons for using GDP per capita, we emphasize that most of our figures would look similar if instead we used one of the productivity measures when they are available.

We now turn to measuring structural transformation. The three most common measures of economic activity at the sectoral level are employment shares, value added shares, and final consumption expenditure shares. Employment shares are calculated either by using workers or hours worked by sector, depending on data availability. Value added shares and final consumption expenditure shares are typically expressed in current prices (nominal shares), but they may also be expressed in constant prices (real shares). While there is a tendency in the literature to view the different measures as interchangeable when documenting how economic activity is reallocated across sectors over time, one of the issues that we want to emphasize in this chapter is that they are in fact distinct. In particular, as we will discuss in detail later on, it is critical to be aware of the distinctions among the different measures when doing quantitative work because even when they display the same qualitative behavior, the quantitative implications may be quite different. Moreover, there are some striking cases in which they display differences even in the qualitative behavior.

Probably the most important reason for the differences between the measures of structural transformation is that employment shares and value added shares are related to production whereas final consumption expenditure shares are related to consumption. Production and consumption measures may display different behaviors because value added is not the same as final output.

A simple example will help to illustrate the distinction between value added and final goods that is relevant here. Consider the purchase of a cotton shirt from a retail establishment. Because the cotton shirt is a good as opposed to a service, in terms of final consumption expenditure, the entire expenditure will be measured as final consumption expenditure of the manufacturing sector. However, in terms of value added in production, the same purchase will be broken down into three pieces: a component from the agricultural sector (i.e. the cotton that was used in making the shirt), a component from the manufacturing sector (i.e. the processing of the cotton and the production of the shirt), and a component from the service sector (i.e. the distribution and retail trade services where the shirt was purchased).

The end result of this is that although the same sectoral labels are used when disaggregating GDP into value added and final expenditure, the resulting measures of sectoral shares are conceptually distinct. It follows that both quantities and prices may differ between value added and final expenditure, implying that there is no reason to expect the implied shares to exhibit similar behavior. This will be of particular relevance when connecting models of structural transformation to the data, which we will discuss in detail below.

The previous discussion emphasized the difference between production and consumption measures. However, even the two measures that focus on production might contain different information. One example comes from Kuznets (1966), who showed for the early part of US development that the employment share of services increased considerably at the same time that the value added share of services remained almost constant.

Having emphasized that each of the three measures of economic activity at the sectoral level is distinct, we also want to note that each of them has its limitations as a singular measure. For the case of sectoral employment shares, a key issue is that employment may not reflect changes in true labor input, for example, because there are systematic differences in hours worked or in human capital per worker across sectors that vary with the level of development. For the case of value added and consumption expenditure shares, a key issue arises from the need to distinguish between changes in quantities and prices. This is often difficult empirically because reliable data on relative price comparisons across countries are hard to come by. In addition, consumption and production need not coincide because of the presence of investment and of imports and exports, so that neither measure alone is sufficient.

6.2.2 Production Measures of Structural Transformation

In this subsection we document the patterns of structural transformation based on examining production measures in several different data sets. We first review the available historical time series evidence for currently rich economies. We then turn to the evidence for currently rich and poor countries.

6.2.2.1 Evidence from Long Time Series for Currently Rich Countries

We construct individual time series of sectoral employment shares and value added shares over the 19th and 20th century for the following 10 countries: Belgium, Finland, France, Japan, Korea, Netherlands, Spain, Sweden, United Kingdom, and United States.[7] Since the early data is sketchy and we want to highlight trends over long periods of time, we report the latest available observation for each decade, if any. We note that for these historical time series we only have measures based on production.

Figure 6.1 plots the historical time series. The vertical axis is either the share of employment or the share of value added in current prices in the three broad sectors of interest. The horizontal axis is the log of GDP per capita in 1990 international dollars as reported by Maddison. The figures clearly reveal what the literature views as the stylized facts of structural transformation. Over the last two centuries, increases in GDP per capita have been associated with decreases in both the employment share and the nominal value added share in agriculture, and increases in both the employment share and the nominal

[7] For a detailed description of the data sources, see the Appendix A.

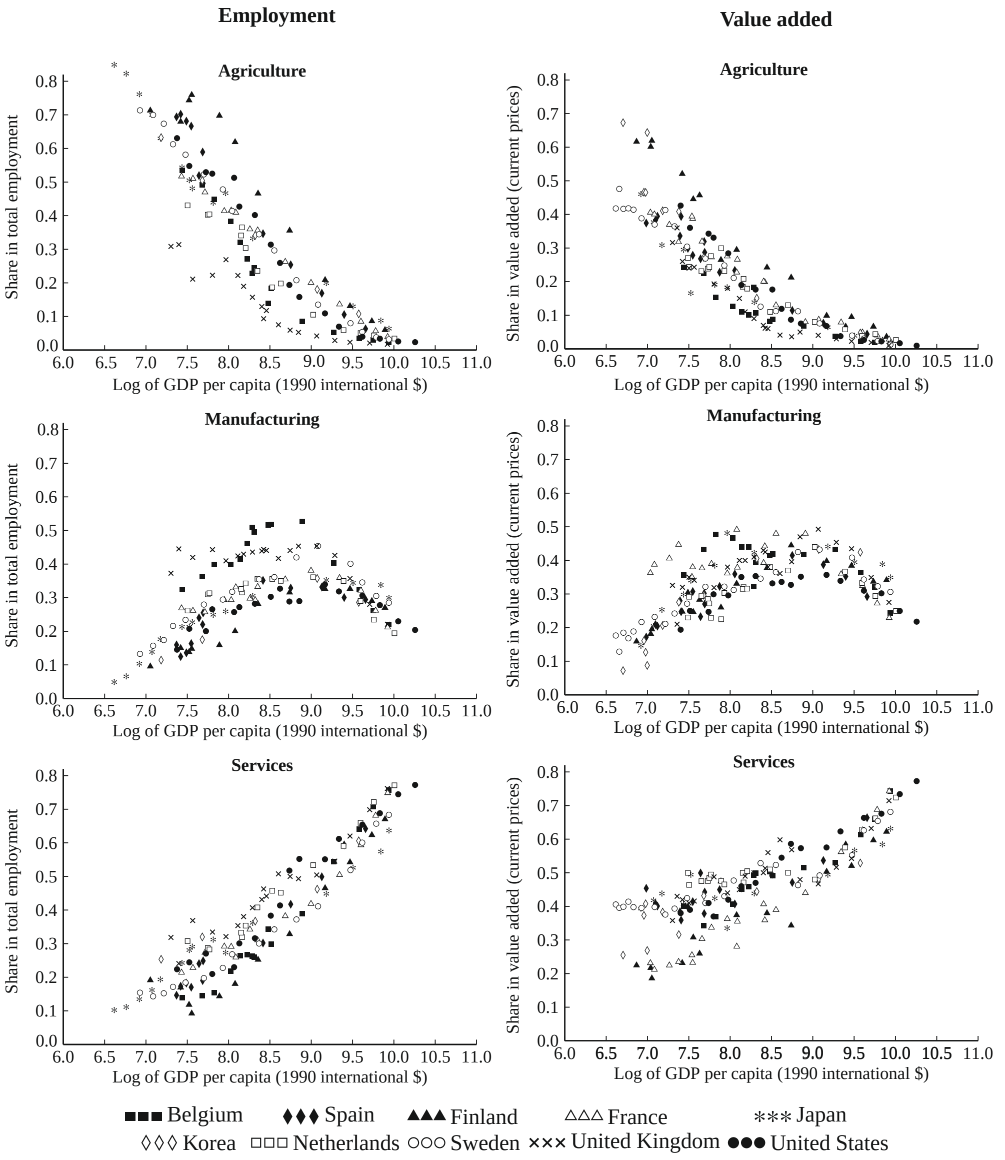


Figure 6.1 Sectoral shares of employment and value added—selected developed countries 1800–2000. *Source: Various historical statistics, see Appendix A.*

value added share in services. Manufacturing has behaved differently from the other two sectors: its employment and nominal value added shares follow a hump shape, that is, they are increasing for lower levels of development and decreasing for higher levels of development.

Figure 6.1 reveals several additional regularities that have been somewhat less appreciated in the context of structural transformation. First, focusing on the agricultural sector, we can see that for low levels of development, the value added share is considerably lower than the employment share. This finding is interesting in light of the fact that countries which are currently poor tend to have most of their workers in agriculture although agriculture is the least productive sector.[8] Second, focusing on the service sector, we see that both the employment share and the nominal value added share for the service sector are bounded away from zero even at very low levels of development; the lowest value added share of services is around 20% and the lowest employment share is around 10%.[9] Third, the figure for the nominal value added share in services suggests that there is an acceleration in the rate of increase when the log of GDP per capita reaches around 9.[10] Inspecting the graphs for the other two nominal value added shares more closely, we also note that the nominal value added share for manufacturing peaks around the same log GDP at which the nominal value added share for the service sector accelerates, so it appears that the accelerated increase in the value added share of services coincides with the onset of the decrease in the value added share for manufacturing sector.[11]

6.2.2.2 Evidence from Recent Panels for Currently Rich and Poor Countries

We now turn to an examination of production measures from several more recent data sets, which tend to be of higher quality than the historical data and which include also countries that are currently poor as well as additional variables (nominal versus real, hours versus employment). The goal of this subsection is to assess the stylized facts of structural transformation that we documented for the historical data, as well as to take advantage of the richer data available so as to examine additional dimensions of structural transformation.

Evidence from EU KLEMS

We start with EU KLEMS, which is compiled at the Groningen Growth and Development Center. The primary strength of EU KLEMS in documenting patterns in employment and value added shares is that it has the most complete information for all variables of interest, including sectoral hours worked, and that its value added data have been constructed from the national accounts of individual countries following a harmonized

[8] See Caselli (2005), and Restuccia et al. (2008) for evidence on this point.

[9] This finding is confirmed by the historical study of Broadberry et al. (2011), who present evidence for England during the 14th century that the employment share of services was around 20%.

[10] See Buera and Kaboski (2012a,b) for additional evidence on this point in a larger cross section of countries.

[11] While we do not develop this issue further here, Buera and Kaboski (2012b) also show that at low levels of GDP per capita the manufacturing sector expands more quickly than does the service sector.

procedure that aims to ensure cross-country comparability.[12] The primary weakness of EU KLEMS is that its coverage is limited to countries with relatively high income; South Korea during the early 1970s is the country with the lowest income in the sample.

We first document the evolution of the shares of sectoral hours worked and nominal value added as functions of the level of development for five non-European countries—i.e. Australia, Canada, Japan, Korea, and the United States—as well as for the aggregate of 15 EU countries.[13] The data are plotted in Figure 6.2. The vertical axis is either the share of total hours worked or the share of value added in current prices in the three broad sectors of interest. As before, the horizontal axis is the log of GDP per capita in 1990 international dollars from Maddison.

The plots in Figure 6.2 confirm several patterns from the historical times series. First, the shares of hours worked and nominal value added for agriculture tend to decrease with the level of development for all countries, whereas the shares for services tend to increase with the level of development for all countries. Second, taken as a whole, the data are consistent with a hump shape for the shares in the manufacturing sector, although all countries except for Korea have decreasing manufacturing shares. Third, the series for both shares as a function of GDP per capita are quite consistent across countries. That is, not only are the qualitative patterns very similar, but so too are the quantitative patterns. This is of particular interest given the considerable attention that has been placed on the role of openness in the growth miracle of Korea (Korea liberalized its manufacturing trade starting in the 1960s and became one of the most open countries in the world). Although, to a lesser extent, one could make similar statements for the case of Japan.

Although this last finding might tempt one to conclude that openness is not a quantitatively important determinant of sectoral allocations and structural transformation, we do want to caution the reader against jumping too quickly to this conclusion. Figure 6.3 shows the same series separately for the 15 EU countries. Although all countries display the same qualitative patterns, there is now substantial heterogeneity in the cross section at any given level of development. This is consistent with the view that these countries form a fairly integrated free-trade zone, thereby allowing for a high degree of specialization, and significant differences in how economic activity is allocated across broad sectors.[14]

Next, we turn our attention to possible differences between real and nominal shares of sectoral value added, where nominal refers to current prices and real refers to constant prices. Kuznets (1966) concluded that the early available data showed similar qualitative

[12] For example, a common industry classification was used and price indices were constructed in a similar way across countries. For more detail see O'Mahony and Timmer (2009), and Timmer et al. (2010).

[13] These are Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain, Sweden, and the United Kingdom.

[14] Some of the series that we consider later on in this section will reveal differences between Korea and the other countries.

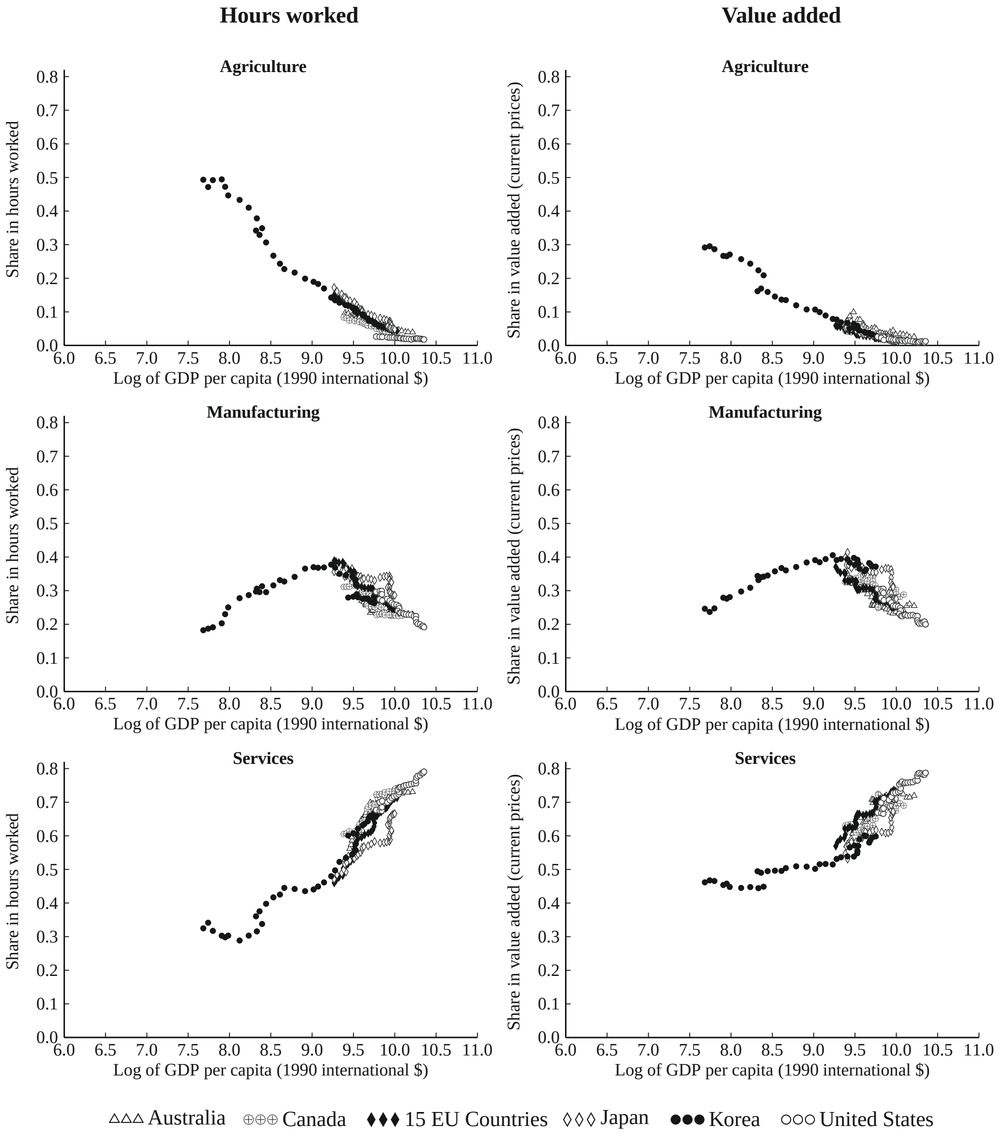


Figure 6.2 Sectoral shares of hours worked and nominal value added—5 non-EU countries and aggregate of 15 EU countries from EU KLEMS 1970–2007. *Source: EU KLEMS, PWT6.3.*

patterns for nominal and real shares. We revisit this comparison because EU KLEMS has more recent and higher quality data than were available to Kuznets. Figure 6.4 plots the real shares of sectoral value added in the left panel and, for comparison, the nominal shares from Figure 6.2 in the right panel. The plots show that the qualitative patterns of real

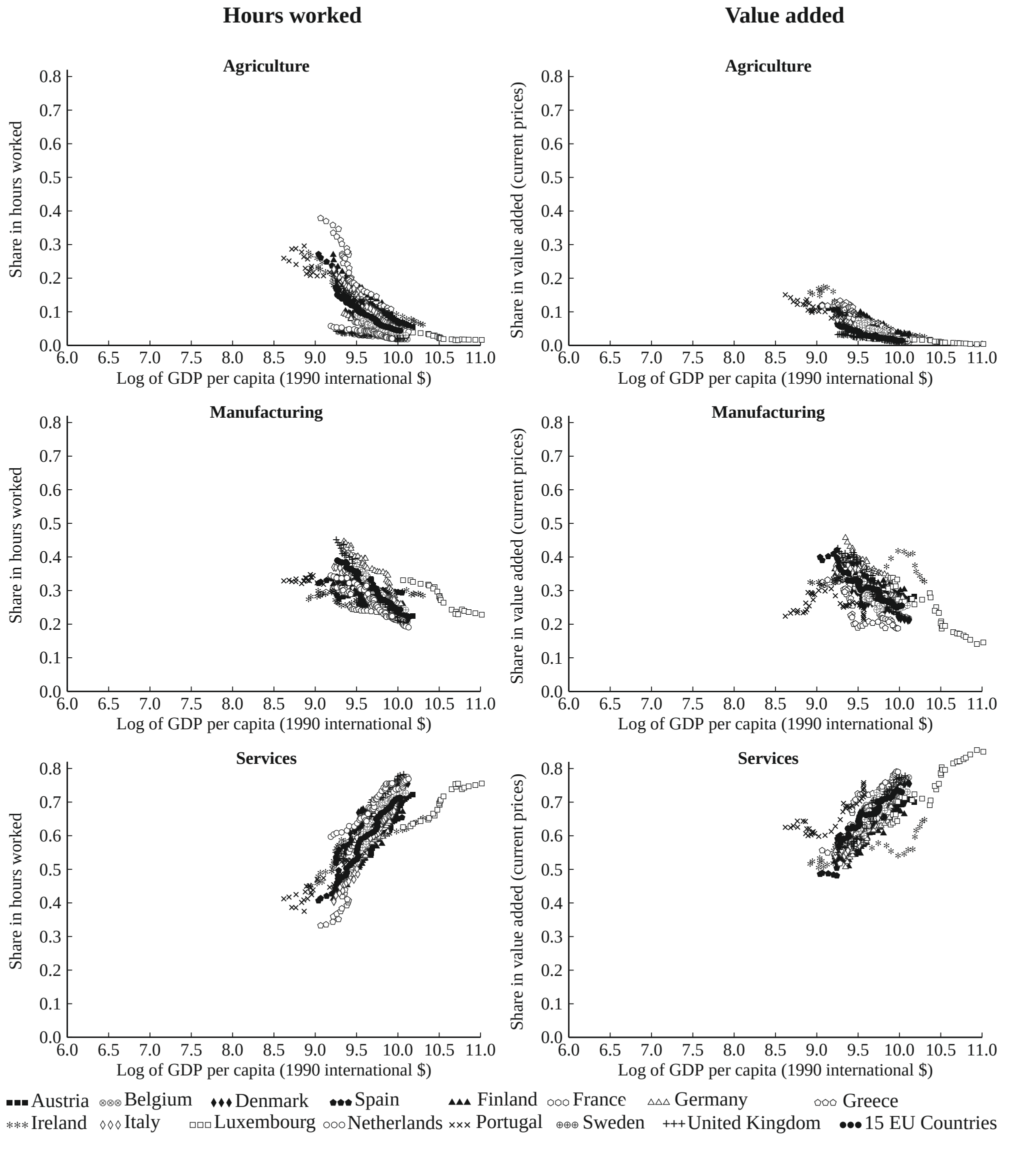


Figure 6.3 Sectoral shares of hours worked and nominal value added—15 EU countries from EU KLEMS 1970–2007. *Source: EU KLEMS, PWT6.3.*

and nominal value added shares are fairly similar to each other, confirming what Kuznets found for the earlier data.

One important exception is Korea where the manufacturing share rose to half of real value added, which is considerably higher than in the other countries on the graph.

At the same time, the manufacturing share of nominal value added flattened out around the maximum share for the other countries. Moreover, the real service share remained below the service share of the other countries, and actually fell somewhat. At the same time, the nominal service share stayed mostly flat. These observations imply that the price

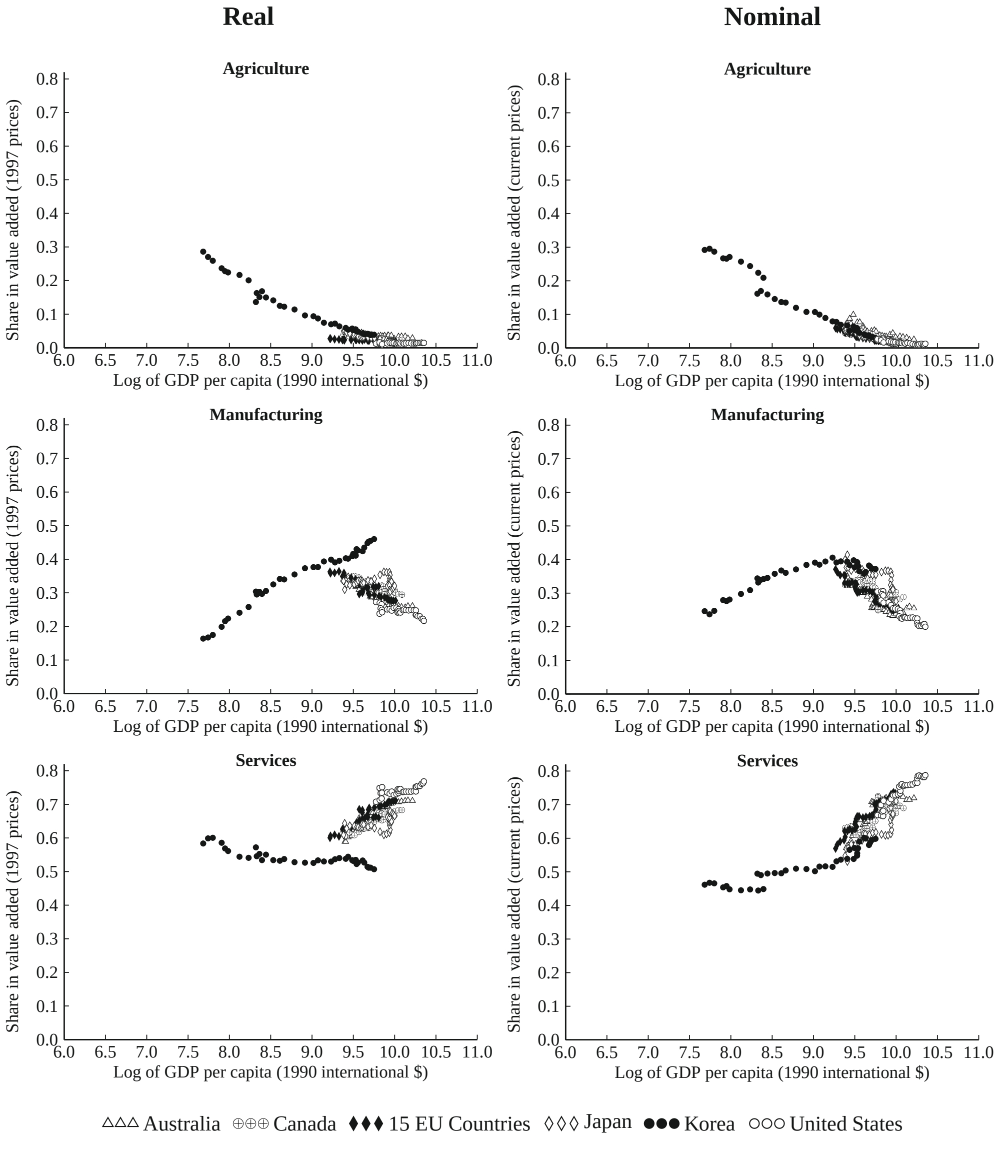


Figure 6.4 Sectoral shares of real and nominal value added—5 non-EU countries and aggregate of 15 EU countries from EU KLEMS 1970–2007. *Source: EU KLEMS, PWT6.3.*

of manufacturing relative to total value added fell by more in Korea than in the other countries. This is consistent with the view that during Korea's massive trade liberalization the relative price of manufactured goods fell considerably at the same time as the real growth rate of manufacturing increased considerably.[15]

Evidence from the WDI and the UN Statistics Division

As previously noted, the main shortcoming of both the historical data and of EU KLEMS is that the coverage is limited to countries that have fairly high income today. It is therefore of interest to verify whether the stylized facts of structural transformation extend to data sets that cover countries that are poor today. The two obvious data sets to use in this context are the World Development Indicators (WDI) and the National Accounts that the United Nations Statistics Division collects.

We use the WDI for employment by sector, which it reports since 1980 based on the data published by the International Labor Organization (ILO). We emphasize that these data are about employed workers instead of hours worked and are of considerably lower quality than those in EU KLEMS because there is much less harmonization underlying the construction of WDI data, which leads to comparability issues. Moreover, WDI employment data are not uniformly available over time for all countries.

We use the national accounts of the United Nations Statistics Division for value added by sector. Unlike the WDI, the UN Statistics Division provides continuous coverage for a large number of countries between 1970 and 2007 and makes an explicit effort to harmonize the national accounts data so as to ensure that they are comparable across different countries.

Figure 6.5 plots the sectoral employment shares from the WDI against GDP per capita from Maddison. The plots confirm that in terms of sectoral employment shares the basic qualitative regularities of structural transformation also hold outside the set of rich countries for which EU KLEMS has data. Specifically, it is the case again that the agricultural employment share decreases in the level of development and that the employment share of services increases in the level of development. Moreover, the employment share in manufacturing is strongly increasing at lower levels of development (log of GDP per worker smaller than 9) before flattening out and then decreasing somewhat for higher levels of development. While this pattern is consistent with a hump shape, we note that the downward sloping part is not very pronounced in the WDI data.

Not surprisingly, the plots also show that employment shares do take on much more extreme values than can be found in EU KLEMS. For example, now the employment share of agriculture can be as high as 70% and the employment shares of manufacturing and services can be as low as only 10%. Lastly, for a given level of development the plots

[15] Looking at sectoral employment shares, Bah (2008) documents that the process of structural transformation in many developing countries also looks different than the historical experiences of current rich countries.

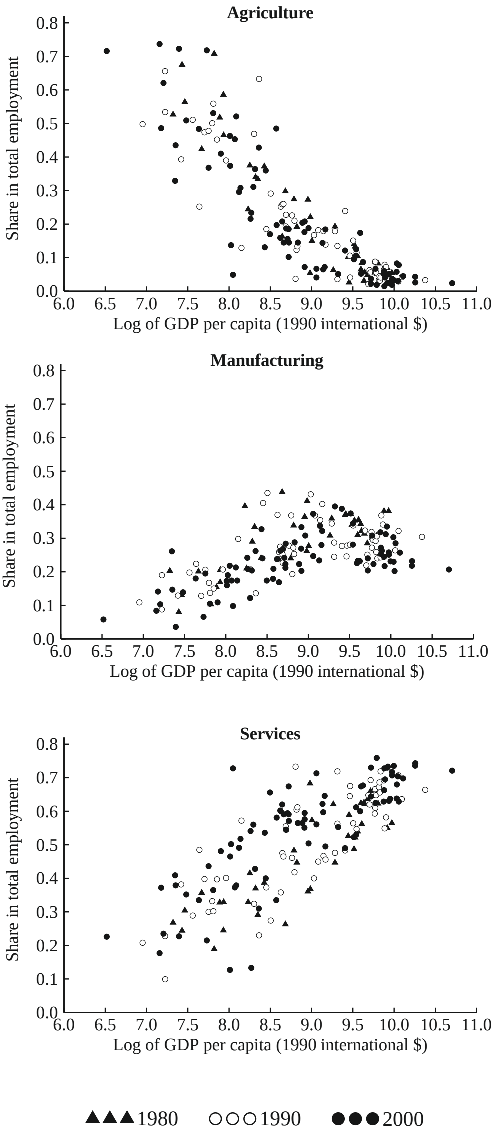


Figure 6.5 Sectoral shares of employment—cross sections from the WDI 1980–2000. *Source: World development indicators 2010.*

show much greater variability in the employment shares relative to what we found in the EU KLEMS data. The extent to which this simply reflects greater measurement error due to lack of comparability and other factors is an open question.

Figure 6.6 plots nominal value added shares by sector from the UN Statistics Division against GDP per capita from Maddison. Since these data have complete coverage for many rich and poor countries, they come close to a balanced panel. We therefore also plot the fitted nominal value added shares from panel regressions. This is intended as a way of summarizing some patterns in the data, instead of as a way of testing any theory. For each sector we regress nominal value added shares on country fixed effects and the level, square, and cube of GDP per worker.[16] We include countries for which no observations are missing, which were not communist, and which had more than a million inhabitants during 1970–2007. Appendix B contains the details regarding the construction of the panel of countries and Tables 6.1–6.3 in Appendix B contain the regression results.

The fitted curves reveal the same qualitative patterns that we have documented previously. It is of particular interest that the hump shape clearly emerges for manufacturing value added. Moreover, it is of interest that the fitted curve for services indicates an acceleration of the service share when the log of GDP per capita reaches a threshold value around 9 and the share of manufacturing value added peaks. Interestingly, this feature occurs at a similar threshold value also for the historical time series which we discussed above.

6.2.3 Consumption Measures of Structural Transformation

Lastly, we turn to the stylized facts of structural transformation when final consumption expenditure shares are used as a measure of economic activity at the sectoral level. We previously offered two main reasons why final consumption expenditure shares may

Table 6.1 Panel data analysis agriculture, 1970–2007

	Dependent variable: Agricultural share in value added					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
log GDP per capita	−0.121**	−0.489**	0.450*	−0.126**	−0.396**	0.169
	(0.001)	(0.021)	(0.184)	(0.015)	(0.067)	(0.274)
(log GDP per capita)^2		0.022**	−0.096**		0.017**	−0.056
		(0.001)	(0.022)		(0.004)	(0.035)
(log GDP per capita)^3			0.005**			0.003*
			(0.001)			(0.001)
Country fixed effects	No	No	No	Yes	Yes	Yes
R^2	0.751	0.783	0.786	0.751	0.781	0.784
N	3914	3914	3914	3914	3914	3914

Notes: Heteroscedasticity robust standard errors in parentheses.
* Significance level $p < 0.05$.
** Significance level $p < 0.01$.

[16] We report results for a cubic polynomial since adding higher-order terms did not have a significant effect on the fitted relationships.

exhibit different patterns than production value added shares: the presence of investment, imports, and exports and the fact that final consumption expenditure is a fundamentally distinct concept from value added produced. The goal of this subsection is to establish that these differences between consumption- and production-based measures do not matter much for agriculture and services, but can have important implications for manufacturing.

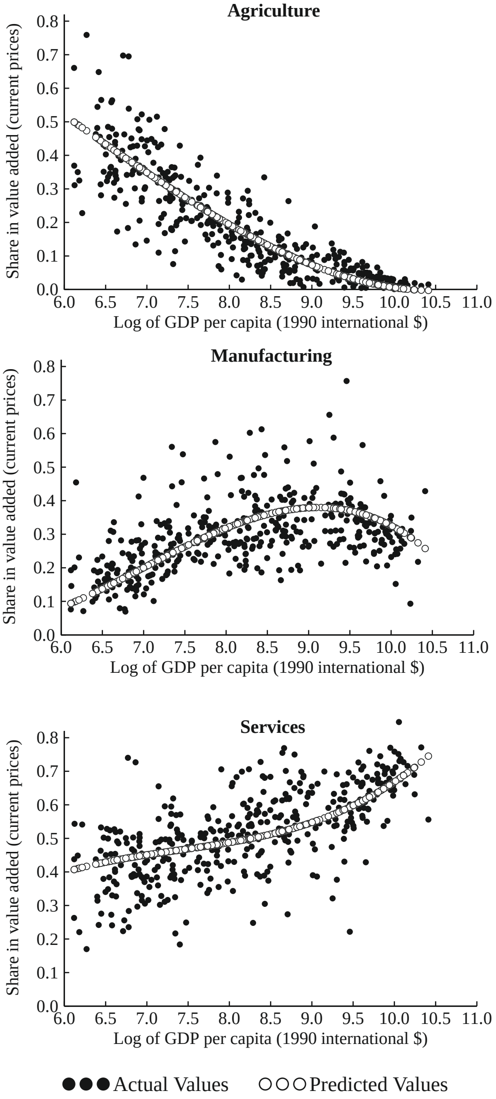


Figure 6.6 Sectoral shares of nominal value added—cross sections from UN national accounts 1975–2005. *Source: National accounts united nations, PWT6.3, own calculations.*

Comparable cross-country panel data on consumption expenditure by sector are much less available than such data on either employment or value added shares. We begin by presenting relatively long time series evidence for the US and the UK in Figure 6.7. The main message from the plots is that for these two countries, production and consumption measures display very similar behavior, both qualitatively and quantitatively. Specifically, nominal consumption shares for agriculture and services are decreasing and increasing over time, respectively, just as they were in the case for nominal value added shares, and the extent of the changes is quite similar too. Moreover, the consumption share for manufacturing displays a hump shape, just as it did in the case for the nominal value added share for manufacturing. Once again, the quantitative features are also similar, with the peak of the curves occurring at similar values of GDP per capita, and the extent of the decrease after the peak also being similar. One difference between consumption shares and value added shares is that the consumption share for manufacturing tends to be a few percentage points higher than the value added share for manufacturing. This occurs because of the fact that the consumption measure implicitly includes distribution services such as retail trade in its measure of manufacturing consumption.

We next consider two data sets on final consumption expenditure by sector: the OECD Consumption Expenditure Data Base and the Benchmark Studies of the International Comparisons Programme, as reported by the Penn World Table. The OECD data have reasonably long time series for several currently rich countries, namely, Australia, Canada, Japan, Korea, and the United States; as well as the seven EU countries, Austria, Denmark, Finland, France, Italy, the Netherlands, and the United Kingdom. The Benchmark Studies offer relatively large cross sections for the years 1980, 1985, and 1996. We define the sectors for consumption expenditure following the usual conventions; for example, we use food as the category closest to agriculture; for the details see Appendix A. For each data set, we

Table 6.2 Panel data analysis manufacturing, 1970–2007

	Dependent variable: Manufacturing share in value added					
	(1)	(2)	(3)	(4)	(5)	(6)
log GDP per capita	0.043**	0.447**	−1.196**	0.054**	0.497**	−1.252**
	(0.001)	(0.021)	(0.144)	(0.017)	(0.078)	(0.446)
(log GDP per capita)^2		−0.025**	0.182**		−0.028**	0.198**
		(0.001)	(0.018)		(0.005)	(0.058)
(log GDP per capita)^3			−0.009**			−0.009**
			(0.001)			(0.002)
R^2	0.234	0.331	0.352	0.234	0.331	0.348
N	3914	3914	3914	3914	3914	3914

Notes: Heteroscedasticity robust standard errors in parentheses.
** Significance level $p < 0.01$.

Table 6.3 Panel data analysis services, 1970–2007

	Dependent variable: Service share in value added					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
log GDP per capita	0.078**	0.041*	0.745**	0.072**	−0.101	1.084*
	(0.001)	(0.019)	(0.170)	(0.012)	(0.089)	(0.417)
(log GDP per capita)^2		0.002*	−0.086**		0.011†	−0.142*
		(0.001)	(0.021)		(0.006)	(0.055)
(log GDP per capita)^3			0.004**			0.006**
			(0.001)			(0.002)
R^2	0.493	0.493	0.496	0.493	0.485	0.476
N	3914	3914	3914	3914	3914	3914

Notes: Heteroscedasticity robust standard errors in parentheses.
† Significance level $p < 0.10$.
* Significance level $p < 0.05$.
** Significance level $p < 0.01$.

pool the data and plot the nominal consumption expenditure shares of the three sectors against GDP per capita measured in 1990 international dollars.

Figure 6.8 contains the plots for the OECD data and Figure 6.9 contains the plots for the Penn World Table data. Two patterns are immediate: the final expenditure share for food tends to decrease with the level of development while the final expenditure share for services tends to increase with development. These two patterns are qualitatively similar to the patterns that we have documented by using the production-based measures of economic activity at the sectoral level. However, when we examine the plot for manufacturing consumption we now see some differences. Of particular interest is Korea, whereas it exhibits the same hump shape as the other OECD countries for the nominal production value added share of manufacturing, we see that its consumption share of manufacturing is virtually flat during a period of rapid growth.

The data from the PWT for the manufacturing consumption share effectively show a cloud. While this plot is not necessarily inconsistent with a hump shape for each country coupled with level differences across countries, it suggests that differences between production and consumption measures may be a more common feature of the data in the larger sample of countries. We think this is an important issue that merits further work. If the link between consumption and production measures is different for current developing countries than it was for countries that developed earlier, then this may well have implications for the nature of the development path that these countries follow.[17]

[17] We are going to revisit this issue below when we discuss in detail our paper Herrendorf et al. (2009).

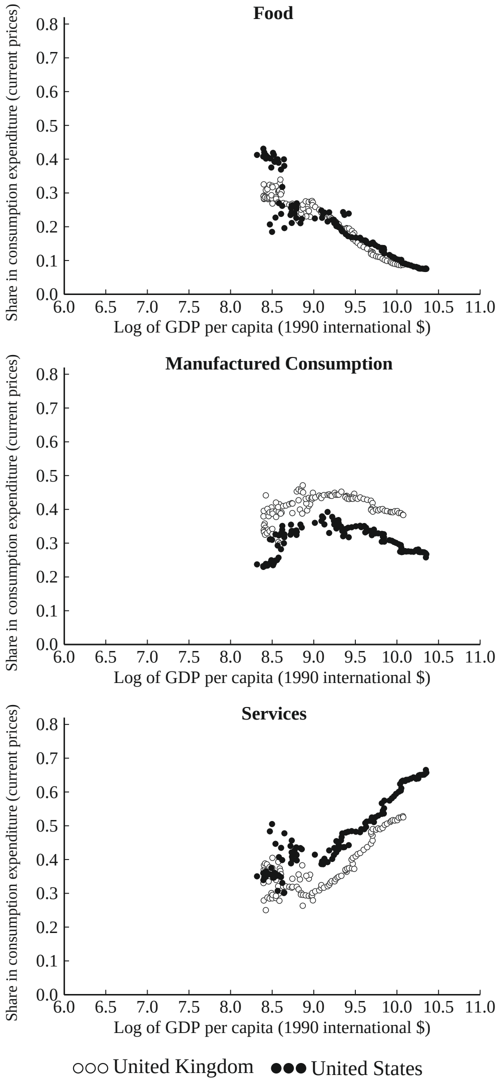


Figure 6.7 Sectoral shares of nominal consumption expenditure—US and UK 1900–2008. *Source: Various historical statistics, see Appendix A.*

6.3. MODELING STRUCTURAL TRANSFORMATION AND GROWTH

In this section we present a natural extension of the one-sector growth model that incorporates structural transformation. We develop our extension in two steps. In the first one, we consider the well-known, two-sector version of the growth model that has separate consumption and investment sectors. In the second step, we disaggregate consumption into the three components: agriculture, manufacturing, and services.

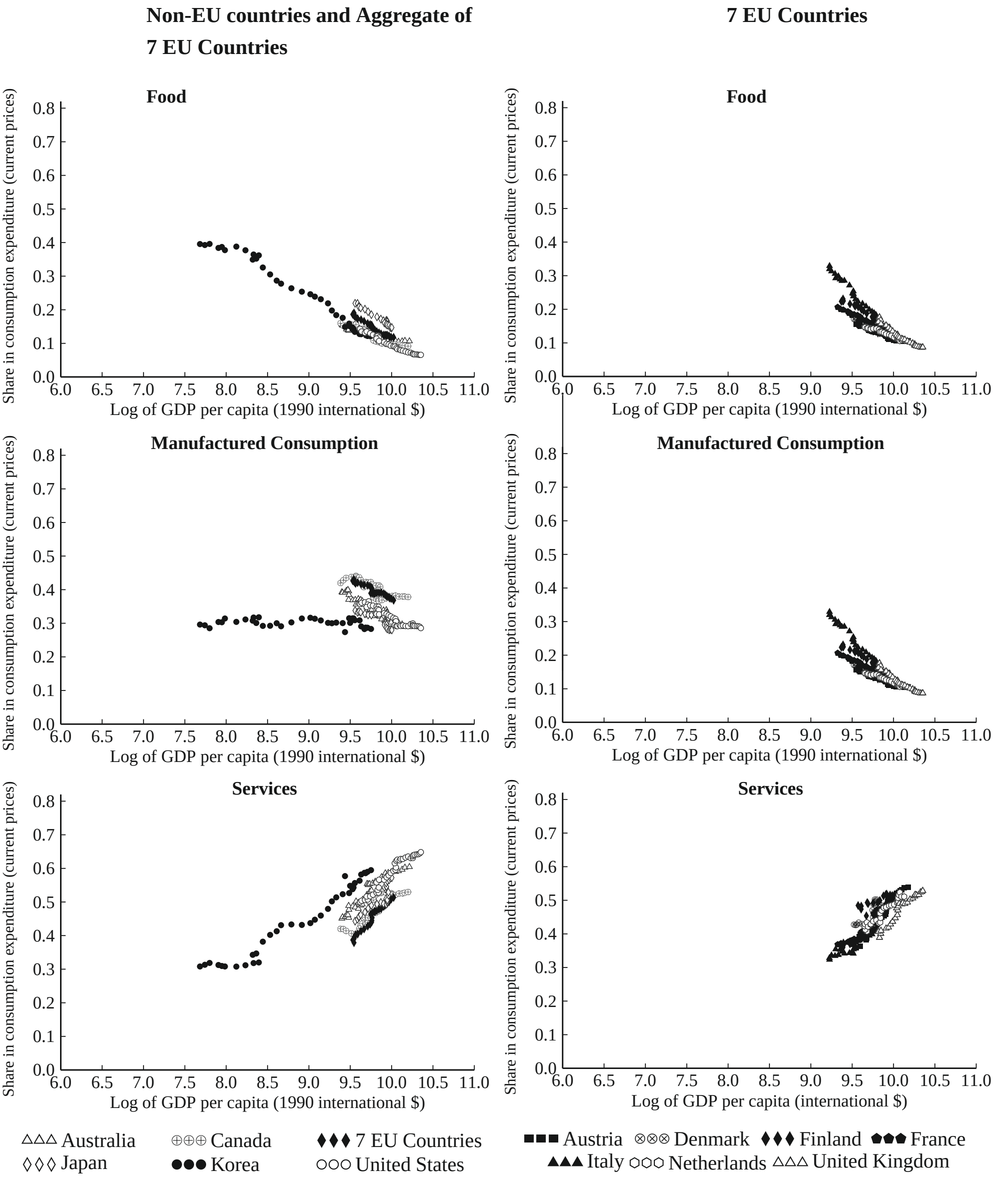


Figure 6.8 Sectoral shares of nominal consumption expenditure—various countries, OECD 1970–2007. *Source: OECD, EU KLEMS, PWT6.3.*

6.3.1 Background: A Two-Sector Version of the Growth Model

Our presentation of the two-sector growth model closely resembles that in Greenwood et al. (1997), which is a version of Uzawa (1963). We assume an infinitely lived stand-in

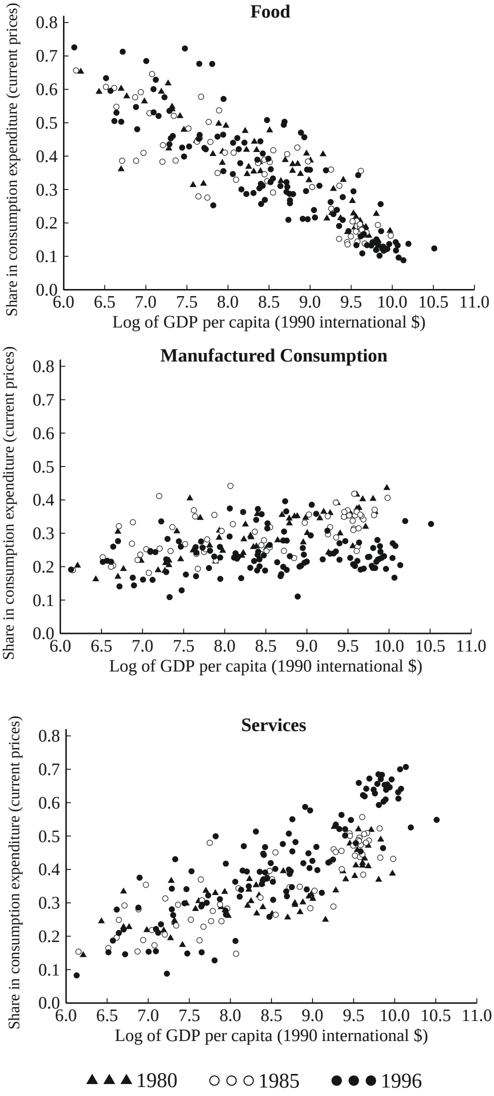


Figure 6.9 Sectoral shares of nominal consumption expenditure—cross sections from the ICP benchmark studies 1980, 1985, 1996. *Source: International comparisons programme (as reported in PWT).*

household with preferences over consumption sequences $\{C_t\}$ given by:

$$\sum_{t=0}^{\infty} \beta^t \log C_t, \tag{6.1}$$

where $0 < \beta < 1$ is the discount factor. Note that, for simplicity, preferences are such that the household does not value leisure. The household is endowed with one unit of productive time and a positive initial stock of capital, K_0.

There are two constant-returns-to-scale production functions which describe how consumption (C) and investment (X) are produced from capital (k) and labor (n). It is convenient to follow the literature and impose that the production functions are Cobb-Douglas and have the same capital share:

$$C_t = k_{ct}^{\theta}(A_{ct}n_{ct})^{1-\theta},$$
$$X_t = k_{xt}^{\theta}(A_{xt}n_{xt})^{1-\theta},$$

where A_{it} represents exogenous labor-augmenting technological progress in sector i. We adopt the notational convention of using upper-case letters to refer to aggregate variables.

Capital accumulates as usual:

$$K_{t+1} = (1-\delta)K_t + X_t,$$

where $0 < \delta < 1$ denotes the depreciation rate.

We assume that capital and labor are freely mobile between the two sectors so that feasibility requires that in each period:

$$K_t = k_{ct} + k_{xt},$$
$$1 = n_{ct} + n_{xt}.$$

As is standard, we study the competitive equilibrium for this economy. Although one can obtain the competitive-equilibrium allocations by solving a social planner's problem, we want to emphasize the role of relative prices and therefore consider a sequence-of-markets competitive equilibrium in which the price of the investment good is normalized to be equal to one in each period. The price of the consumption good relative to the investment good is denoted by P_t, the rental rate for capital is denoted by R_t, and the wage rate is denoted by W_t. We assume that the household accumulates capital and rents it to firms.

We begin our characterization of the equilibrium by establishing that the capital-to-labor ratios are equalized across sectors at each point in time. To see this, note that the first-order conditions for the stand-in firm in sector $i \in \{c, x\}$ are given by:

$$R_t = P_t\theta \left(\frac{k_{ct}}{n_{ct}}\right)^{\theta-1} A_{ct}^{1-\theta} = \theta \left(\frac{k_{xt}}{n_{xt}}\right)^{\theta-1} A_{xt}^{1-\theta},$$
$$W_t = P_t(1-\theta) \left(\frac{k_{ct}}{n_{ct}}\right)^{\theta} A_{ct}^{1-\theta} = (1-\theta) \left(\frac{k_{xt}}{n_{xt}}\right)^{\theta} A_{xt}^{1-\theta}.$$

Combining these two equations and rearranging gives an expression for the capital-to-labor ratio in sector $i \in \{c, x\}$:

$$\frac{k_{it}}{n_{it}} = \frac{\theta}{1-\theta}\frac{W_t}{R_t}.$$

It follows that the capital-to-labor ratio in each sector is the same and equals the aggregate capital-to-labor ratio[18]:

$$\frac{k_{ct}}{n_{ct}} = \frac{k_{xt}}{n_{xt}} = K_t. \tag{6.2}$$

Next, we establish that the equilibrium value of the relative price P_t is pinned down by technology. To see this, divide the first-order conditions for labor from the two sectors by each other and use the fact that sectoral capital-to-labor ratios are equalized. This gives:

$$P_t = \left(\frac{A_{xt}}{A_{ct}}\right)^{1-\theta}. \tag{6.3}$$

Equations (6.2) and (6.3) imply that:

$$P_t C_t = \left(\frac{k_{ct}}{n_{ct}}\right)^{\theta} P_t A_{ct}^{1-\theta} n_{ct} = K_t^{\theta} A_{xt}^{1-\theta} n_{ct}.$$

It follows that the model aggregates on the production side, that is, we can consider an aggregate production function that produces a single good that can be turned into either consumption or investment via a linear technology with marginal rate of transformation equal to P_t:

$$Y_t = X_t + P_t C_t = K_t^{\theta}(A_{xt})^{1-\theta}(n_{xt} + n_{ct}) = K_t^{\theta} A_{xt}^{1-\theta}. \tag{6.4}$$

Additionally, Equation (6.2) and the first-order conditions for the firm in the investment sector imply that the marginal products of the aggregate production function determine the rental rate of capital and the wage rate:

$$R_t = \theta K_t^{\theta-1} A_{xt}^{1-\theta}, \tag{6.5}$$
$$W_t = (1-\theta) K_t^{\theta} A_{xt}^{1-\theta}. \tag{6.6}$$

[18] To see this note that:

$$\frac{k_{ct}}{n_{ct}} n_{ct} + \frac{k_{xt}}{n_{xt}} n_{xt} = K_t(n_{ct} + n_{xt}) = K_t.$$

To characterize the competitive equilibrium further, we turn to the household side. The household's maximization problem is[19]:

$$\max_{\{C_t,K_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \log C_t \quad \text{st} \quad P_t C_t + K_{t+1} = (1-\delta+R_t)K_t + W_t.$$

Letting μ_t denote the current-value Lagrange multiplier on the period t budget constraint, the first-order conditions for C_t and K_t are:

$$\frac{\beta^t}{C_t} = \mu_t P_t,$$
$$1-\delta+R_t = \frac{\mu_{t-1}}{\mu_t}.$$

Combining these two equations gives the Euler equation:

$$\frac{1}{\beta}\frac{P_t C_t}{P_{t-1}C_{t-1}} = 1-\delta+R_t. \tag{6.7}$$

Using Equations (6.4) and (6.5), Equation (6.7) can be written as a second-order difference equation in the aggregate capital stock K_t. Given a value for the initial capital stock, this second-order difference equation together with a transversality condition determines the equilibrium sequence of capital stocks.

We are now ready to consider the possibility of a balanced growth path in this model. We start by assuming that both technologies improve at constant, though not necessarily equal, rates $\gamma_i > 0$:

$$\frac{A_{it+1}}{A_{it}} = 1+\gamma_i, \quad i = c, x.$$

The standard definition of balanced growth is that endogenous variables are constant or grow at constant rates. It turns out that this definition is too strict for models with structural transformation because the very nature of structural transformation is that the sectoral composition changes. We therefore follow the literature and use the weaker concept of generalized balanced growth path (GBGP), which only requires that the real interest rate is constant.

The motivation for requiring that the real interest rate be constant is that although it may exhibit short-term fluctuations, it does not show a long-term trend. This, of course,

[19] Note that if total consumption grows at a constant rate γ_c, which will be the case below when we consider generalized balanced growth, then the household's objective function remains finite, and so is well-defined.The reason for this is that:

$$\sum_{t=0}^{\infty} \beta^t \log C_t = \log C_0 \sum_{t=0}^{\infty} \beta^t + \log(\mathrm{Hrc}) \sum_{t=0}^{\infty} \beta^t t < \infty.$$

is one of the Kaldor facts. The next result shows that along a GBGP of our two-sector model the other four facts of Kaldor will also hold; that is, K_t and Y_t grow at constant rates and K_t/Y_t and R_tK_t/Y_t are constant.

Proposition 1. *If a GBGP exists, then the Kaldor facts hold along the GBGP.*

Proof. Since R_t is constant along a GBGP, it suffices to show that K_t, Y_t, and X_t all grow at rate γ_x.

The fact that R is constant and Equation (6.5) holds in period t and $t+1$ implies:

$$\frac{A_{xt+1}}{A_{xt}} = \frac{K_{t+1}}{K_t}. \tag{6.8}$$

It follows that K_t also grows at the constant rate of γ_x. Using $Y_t = A_{xt}^{1-\theta}K_t^{\theta}$, we have:

$$\frac{Y_{t+1}}{Y_t} = \left(\frac{A_{xt+1}}{A_{xt}}\right)^{1-\theta}\left(\frac{K_{t+1}}{K_t}\right)^{\theta}. \tag{6.9}$$

Using Equation (6.8) this gives:

$$\frac{Y_{t+1}}{Y_t} = (1+\gamma_x)^{\theta}(1+\gamma_x)^{1-\theta} = 1+\gamma_x. \tag{6.10}$$

In other words, Y grows at a constant rate. Moreover, constant growth of K necessarily implies constant growth of X. The fact that the aggregate technology is Cobb-Douglas implies that factor shares are constant even off a GBGP. □

If K_t grows at the constant rate γ_x, then the law of motion for capital implies that X_t must grow at the same constant rate. Equation (6.4) then implies that P_tC_t must also grow at this same rate. Substituting this growth rate into Equation (6.7) pins down the constant value of the rental rate of capital along a GBGP:

$$\frac{1}{\beta}(1+\gamma_x) = 1-\delta+R.$$

Given a value for A_{x0}, using this version of the Euler equation and the condition on the equilibrium rental rate (6.5), we obtain the unique value of K_0 along a GBGP:

$$K_0 = \left[\frac{\beta\theta}{(1+\gamma_x)-\beta(1-\delta)}\right]^{\frac{1}{1-\theta}} A_{x0}. \tag{6.11}$$

We note several features of this generalized balanced growth path. First, K_t and C_t grow at different rates along the GBGP. In particular, since (6.3) implies that P_t grows at gross rate $[(1+\gamma_x)/(1+\gamma_c)]^{1-\theta}$, and P_tC_t grows at gross rate $(1+\gamma_x)$, it follows that C_t grows at gross rate $(1+\gamma_x)^{\theta}\ (1+\gamma_c)^{1-\theta}$, i.e. a weighted average of the two sectoral growth rates in technology. Given that X_t grows at the same rate as both A_{xt} and K_t, it follows that sectoral employment and capital shares are constant along the balanced growth path. In other words, although in this model differential rates of technological progress lead to changes in relative prices of sectoral outputs, these price changes are not associated with any changes in factor allocations over time.

For future reference, it is of interest to note that although we assumed that technological progress in both sectors is constant over time, this is not required for the existence of a GBGP. In fact, because along the GBGP, the difference in technological progress only shows up in prices and not in allocations, it follows that the same results would apply even if the growth rate of technological progress in the consumption sector varied over time. This would have no effect on how capital and labor are allocated and would only show up in the behavior of the relative price P_t. Although in this case not all variables would grow at constant rates, it would still be true that the rental rate of capital would be constant and that Y_t and K_t would grow at the same constant rate. Thus, there would still be a GBGP.

6.3.2 A Benchmark Model of Growth and Structural Transformation

We use the model of the previous section as the starting point for our analysis of structural transformation in the context of the growth model.

6.3.2.1 Set up of the Benchmark Model

As in the previous section, we assume an infinitely lived stand-in household that has preferences characterized by (6.1) and is endowed with one unit of time and a positive initial capital stock. Different than in the previous section, we now assume that C_t is a composite of agricultural consumption (c_{at}), manufacturing consumption (c_{mt}), and service consumption (c_{st}):

$$C_t = \left[\omega_a^{\frac{1}{\varepsilon}} \left(c_{at} - \bar{c}_a\right)^{\frac{\varepsilon-1}{\varepsilon}} + \omega_m^{\frac{1}{\varepsilon}} \left(c_{mt}\right)^{\frac{\varepsilon-1}{\varepsilon}} + \omega_s^{\frac{1}{\varepsilon}} \left(c_{st} + \bar{c}_s\right)^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\varepsilon}{\varepsilon-1}}, \tag{6.12}$$

where $\bar{c}_i, \omega_i \geq 0$ and $\varepsilon > 0$. The functional form (6.12) is a parsimonious choice that allows us to capture two features on the demand side that are potentially important for understanding the reallocation of activity across these three sectors: how the demand of the household reacts to changes in income and in relative prices. In particular, the presence of the two terms $\bar{c}_a$ and $\bar{c}_s$ allows for the period utility function to be non-homothetic and therefore the possibility that changes in income will lead to changes in expenditure shares even if relative prices are constant. The parameter ε influences the elasticity of substitution between the three goods, and hence the response of nominal expenditure shares to changes in relative prices. Note, however, that in the above specification the elasticity of substitution is not equal to ε because it also depends on the non-homotheticity terms.

Note also that we raise the weights w_i by the exponent $1/\varepsilon$ to ensure that the generalized Leontief utility function is the limit as ε approaches 0:

$$\lim_{\varepsilon \to 0} C_t = \min\{w_a(c_a t - \overline{c_a}), w_m c_m t, w_s(c_s t + c_s)\}$$

We generalize the previous model to allow for four Cobb-Douglas production functions, one for each of the three consumption goods and one for the investment good. Formally, the production functions are given by[20]:

$$c_{it} = k_{it}^{\theta}(A_{it} n_{it})^{1-\theta}, \quad i \in \{a, m, s\}, \tag{6.13}$$

$$X_t = k_{xt}^{\theta}(A_{xt} n_{xt})^{1-\theta}. \tag{6.14}$$

There is a tradition in the literature of working with only three production functions, with the assumption that all investment is produced by the manufacturing sector. Under this assumption, the output of the manufacturing sector can be used as either consumption or investment whereas the output of the other two sectors can only be used as consumption. We have not adopted this specification for two reasons. First, despite the apparent reasonableness of the claim that investment is to first approximation produced exclusively by the manufacturing sector, it turns out that this is not supported by the data. Moreover, such an assumption is becoming increasingly at odds with the data over time, due at least in part to the fact that software is both a sizeable and increasing component of investment, and that most software innovation takes place in the service sector. In fact, for this reason total investment has exceeded the size of the entire manufacturing sector in the US since 2000. The second reason for considering a separate investment sector derives from evidence that technological progress in the investment sector has been more rapid than in the rest of the economy; see, for example Greenwood et al. (1997). Because the possibility of differential rates of technological progress across sectors will play a key role in the subsequent analysis, we want to allow for the possibility that this rate is different in the investment sector.

Capital is accumulated as usual:

$$K_{t+1} = (1 - \delta)K_t + X_t.$$

[20] We follow much of the literature in abstracting from the differences between physical capital and land and treating land as part of physical capital. We then restrict our attention to Cobb-Douglas production functions in capital and labor that have the same capital share in all sectors, which is analytically very convenient, because it implies that we can aggregate the sectoral production functions to an economy-wide Cobb-Douglas production function. In Section 6.5.1.2 we will explore to which extent the assumption of equal sectoral capital shares is borne out by the data. For now, we just mention that even if one thinks that sectoral capital shares (where capital includes land) are similar, then there are still important applications for which it is crucial that land is a fixed factor. For such applications, one needs to model land and physical capital separately.

As before, we assume that capital and labor are freely mobile.[21] With four sectors, the feasibility conditions now take the form:

$$\begin{aligned} K_t &= k_{at} + k_{mt} + k_{st} + k_{xt}, \\ 1 &= n_{at} + n_{mt} + n_{st} + n_{xt}. \end{aligned}$$

6.3.2.2 Equilibrium Properties of the Benchmark Model

We again consider a sequence-of-markets competitive equilibrium in which the price of the investment good is normalized to equal one in each period. The prices of the consumption goods relative to the investment good are denoted by p_{it}, $i \in \{a, m, s\}$. We again assume that the household accumulates capital and rents it to firms.

Several key properties of the two-sector model that we established above continue to hold in the four-sector model. Specifically, using the same logic as in the previous section, one can show that the capital-to-labor ratios are equalized across the four sectors at each point in time, and are equal to the aggregate capital-to-labor ratio:

$$\frac{k_{it}}{n_{it}} = K_t, \quad i = a, m, s, x. \tag{6.15}$$

Moreover, as before, relative prices are determined by technology:

$$p_{it} = \left(\frac{A_{xt}}{A_{it}}\right)^{1-\theta}, \quad i = a, m, s. \tag{6.16}$$

Using the above results, one can also show that our multi-sector model aggregates on the production side:

$$Y_t = p_{at}c_{at} + p_{mt}c_{mt} + p_{st}c_{st} + X_t = K_t^{\theta} A_{xt}^{1-\theta}. \tag{6.17}$$

Lastly, the first-order conditions from the firm problems, (6.5) and (6.6), are still valid.

On the household side, the model is more involved now. In particular, the household problem now takes the form:

$$\max_{\{c_{at},c_{mt},c_{st},K_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \log \left[\omega_a^{\frac{1}{\varepsilon}} (c_{at} - \bar{c}_a)^{\frac{\varepsilon-1}{\varepsilon}} + \omega_m^{\frac{1}{\varepsilon}} (c_{mt})^{\frac{\varepsilon-1}{\varepsilon}} + \omega_s^{\frac{1}{\varepsilon}} (c_{st} + \bar{c}_s)^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\varepsilon}{\varepsilon-1}}$$
$$\text{st} \quad p_{at}c_{at} + p_{mt}c_{mt} + p_{st}c_{st} + K_{t+1} = (1 - \delta + R_t)K_t + W_t.$$

In what follows, we show that this problem can be split into two subproblems: (i) how to allocate total income between total consumption and savings; and (ii) how to allocate total consumption expenditure between the three consumption goods. We develop a

[21] We discuss the case of restricted labor mobility in Section 6.6.2.

useful representation in which the first subproblem closely resembles the problem of the household in the two-sector model considered previously.

In order to have a well-defined household problem, we need to make sure that the consumption of agricultural goods will exceed the subsistence term $\bar{c}_a$ in each period. Even if this is the case, a corner solution may still arise in which the household chooses zero consumption of services. For now, we assume that the household problem is well defined and that its solution is interior in all periods. In Proposition 2 below, we offer a formal condition to ensure that this is the case along the GBGP. Essentially, this will boil down to requiring that in each period total consumption is large enough relative to the two terms $\bar{c}_a$ and $\bar{c}_s$.

The first-order conditions for an interior solution for the three consumption categories are:

$$\frac{1}{C_t}\omega_a^{\frac{1}{\varepsilon}}\left(c_{at}-\bar{c}_a\right)^{-\frac{1}{\varepsilon}}C_t^{\frac{1}{\varepsilon}} = \lambda_t p_{at}, \tag{6.18}$$

$$\frac{1}{C_t}\omega_m^{\frac{1}{\varepsilon}}\left(c_{mt}\right)^{-\frac{1}{\varepsilon}}C_t^{\frac{1}{\varepsilon}} = \lambda_t p_{mt}, \tag{6.19}$$

$$\frac{1}{C_t}\omega_s^{\frac{1}{\varepsilon}}\left(c_{st}+\bar{c}_s\right)^{-\frac{1}{\varepsilon}}C_t^{\frac{1}{\varepsilon}} = \lambda_t p_{st}, \tag{6.20}$$

where λ_t denotes the current-value Lagrange multiplier on the budget constraint in period t. If one raises each of the Equations (6.18)–(6.20) to the power $1-\varepsilon$, adds them, and uses the definition (6.12) of C_t, then one obtains:

$$\frac{1}{C_t} = \lambda_t\left[\omega_a(p_{at})^{1-\varepsilon}+\omega_m(p_{mt})^{1-\varepsilon}+\omega_s(p_{st})^{1-\varepsilon}\right]^{\frac{1}{1-\varepsilon}}. \tag{6.21}$$

Given that λ_t is the marginal value of an additional unit of expenditure in period t, it follows that the other term on the right-hand side is naturally interpreted as the price of a unit of composite consumption. In view of this, we will define the price index P_t by:

$$P_t \equiv \left[\omega_a\left(p_{at}\right)^{1-\varepsilon}+\omega_m\left(p_{mt}\right)^{1-\varepsilon}+\omega_s\left(p_{st}\right)^{1-\varepsilon}\right]^{\frac{1}{1-\varepsilon}}. \tag{6.22}$$

If one adds the three first-order conditions (6.18)–(6.20) and uses this definition of P_t, one also obtains:

$$p_{at}c_{at}+p_{mt}c_{mt}+p_{st}c_{st} = P_tC_t+p_{at}\bar{c}_a-p_{st}\bar{c}_s. \tag{6.23}$$

It follows that the household's maximization problem can be broken down into two subproblems:

(i) Intertemporal Problem. Allocate total income among the composite consumption good and savings:

$$\max_{\{C_t,K_{t+1}\}_{t=0}^{\infty}}\sum_{t=0}^{\infty}\beta^t\log C_t \quad \text{st} \quad P_tC_t+K_{t+1}=(1-\delta+r_t)K_t+w_t-p_{at}\bar{c}_a+p_{st}\bar{c}_s.$$

(ii) Static Problem. Allocate the period t consumption expenditure P_tC_t among the three consumption goods:

$$\max_{c_{at},c_{mt},c_{st}} \left[\omega_a^{\frac{1}{\varepsilon}}\left(c_{at}-\bar{c}_a\right)^{\frac{\varepsilon-1}{\varepsilon}}+\omega_m^{\frac{1}{\varepsilon}}\left(c_{mt}\right)^{\frac{\varepsilon-1}{\varepsilon}}+\omega_s^{\frac{1}{\varepsilon}}\left(c_{st}+\bar{c}_s\right)^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\varepsilon}{\varepsilon-1}}$$
$$\text{st} \quad p_{at}c_{at}+p_{mt}c_{mt}+p_{st}c_{st}=P_tC_t+p_{at}\bar{c}_a-p_{st}\bar{c}_s.$$

This representation nicely separates out the growth component of the model from the structural transformation component of the model. From the perspective of balanced growth in the aggregates K_t and C_t, the representation looks like the two-sector growth model with the exception of one detail: this economy behaves as if there is a time varying endowment, reflected by the term $-p_{at}\bar{c}_a+p_{st}\bar{c}_s$. If this endowment happens to be zero at all dates, then the equivalence to a standard two-sector model is exact. Be that as it may, the Euler equation is still of the form (6.7). Moreover, although the expression for the relative price P_t is somewhat more complicated in the current setting compared to the two-sector model, the equilibrium value of this relative price can still be determined directly from primitives without solving for the full equilibrium.

From the perspective of structural transformation, the above representation implies that we can focus on the solution to the static problem of allocating each period's consumption expenditure between the three consumption goods. The first-order conditions (6.18)–(6.20) characterize the solution to this static problem. For future reference, we note two useful implications of the first-order conditions. First, they impose conditions on the ratios of any two consumption goods:

$$\left(\frac{p_{at}}{p_{mt}}\right)^{\varepsilon}\frac{c_{at}-\bar{c}_a}{c_{mt}}=\frac{\omega_a}{\omega_m}, \tag{6.24}$$

$$\left(\frac{p_{st}}{p_{mt}}\right)^{\varepsilon}\frac{c_{st}+\bar{c}_s}{c_{mt}}=\frac{\omega_s}{\omega_m}. \tag{6.25}$$

Second, they impose a condition on the ratio of the expenditure on composite consumption and the expenditure on manufactured consumption:

$$\frac{P_tC_t}{p_{mt}c_{mt}}=\left[\frac{\omega_a}{\omega_m}\left(\frac{A_{mt}}{A_{at}}\right)^{(1-\theta)(1-\varepsilon)}+1+\frac{\omega_s}{\omega_m}\left(\frac{A_{mt}}{A_{st}}\right)^{(1-\theta)(1-\varepsilon)}\right]. \tag{6.26}$$

Equations (6.24)–(6.26) will play a key role below when we study the details of structural transformation within the framework of our four-sector model.

6.3.3 Connecting the Benchmark Model to Measures of Structural Transformation

Since we will eventually ask whether versions of this model can help us understand the stylized facts of structural transformation that we documented in Section 6.2, it is relevant to briefly discuss some issues related to how one connects the model just described to the various measures from the data that we have previously examined. While this might appear obvious, there are a couple of issues that require notice.

In Section 6.2, we disaggregated total value added into the value added of agriculture, manufacturing, and services; and measured the shares of these three sectors in total value added. To connect our model with these measures of sectoral activity, it is natural to assume that the sectoral production functions that we have specified in the benchmark model represent value added production functions. However, because we have modeled the investment sector as a separate sector, one also needs to allocate the value added from the investment sector among the other three sectors. The literature often assumes that the entire value added of the investment sector belongs to manufacturing. This assumption is inconsistent with the data, for the simple reason that in recent years, in the United States, the value added of the investment sector has exceeded the value added of the (total) manufacturing sector. An alternative is to allocate investment value added to the other sectors using constant shares. This is also at odds with the data, since as shown in Herrendorf et al. (2009), the increasing importance of software as a component of investment has led to an increase in the share of investment value added occurring in the service sector. Nonetheless, since it serves to facilitate transparency, we will adopt this alternative as a benchmark in the next section when we discuss the qualitative features of balanced growth paths in different special cases of the model. However, it should be kept in mind that movements in the sectoral distribution of investment value added shares could affect the predictions that we highlight. As a practical matter, while this effect can matter, it is probably not so relevant at the quantitative level because total investment is a relatively small share of GDP.

The second issue concerns how to connect the model with production value added data versus consumption expenditure data. Specifically, assuming that the sector production functions are interpreted as value added production functions leads to a difficulty when trying to connect the model with data on consumption expenditure shares. Because equilibrium requires that $c_{it} = k_{it}^{\theta}(A_{it}n_{it})^{1-\theta}$, it would seem natural to identify $p_{it}c_{it}/\sum_j p_{jt}c_{jt}$ as the model's measure of the nominal consumption share of sector i in period t. However, this share is not the appropriate measure for the nominal consumption expenditure share of sector i as measured in the data. To see why, let us return to the example discussed earlier of the purchase of a cotton shirt. To measure the contribution of this shirt to manufactured final consumption expenditure, we need to aggregate all value added that goes into the production of the shirt through the use of intermediate inputs from each of the three sectors. This requires us to take into account the input-output relationships about how value added is aggregated into final consumption expenditure. In contrast, the above definition of consumption shares includes only the value added that came from the manufacturing sector itself, and so it does not reflect how final consumption expenditure is measured in a world in which each sector uses intermediate inputs from the other sectors.

To avoid this problem, one could alternatively assume that $p_{it}c_{it}/\sum p_{jt}c_{jt}$ in the model does correspond to the nominal consumption expenditure share of sector i in period t as measured in the data. But since in equilibrium $c_{it} = k_{it}^{\theta}(A_{it}n_{it})^{1-\theta}$, it would then follow that $p_{it}k_{it}^{\theta}(A_{it}n_{it})^{1-\theta}$ is not an appropriate measure of value added from sector i in period t as measured in the data. Returning to the shirt example, this piece of c_{mt} now reflects the value added components from each of the three sectors that went into producing the final product, and so it cannot be the value added from one particular sector. In order to maintain consistency, it must be that the production functions summarize the labor and capital from the various stages of production that are used to produce final consumption expenditure. In order to obtain value added shares one would have to use (inverse) input-output relationships to unbundle the final consumption expenditure into its value added components. Moreover, since n_{it} now reflects all of the labor that went into producing the shirt at each of the various stages of production, it is also no longer the case either that n_{it} is an appropriate measure of the employment share of sector i in period t.

The bottom line from this discussion is that if one wants to have a model that can simultaneously address the shares of sectoral employment, value added, and consumption expenditure, then one will need to explicitly include the details of the input-output structure involved in transforming sectoral value added into sectoral consumption expenditure. We have chosen not to do this in order to preserve a greater degree of transparency in the presentation. In view of this, we need to keep in mind that when we discuss the model implications for the measures of structural transformation, we can either connect the production measures (employment shares and value added shares) to the data, implying that the consumption measure (consumption expenditure shares) does not have a close empirical counterpart, or we can connect the consumption measure to the data implying that the two production measures do not have close empirical counterparts. Whichever way we choose, our model will not be able to make statements about all three measures of structural transformation at the same time. Moreover, as we discuss later on in more detail, one should not assume that preference and technology parameters are invariant to the interpretation that one imposes on the model objects.

6.4. THE ECONOMIC FORCES BEHIND STRUCTURAL TRANSFORMATION: THEORETICAL ANALYSIS

The Kaldor facts regarding balanced growth over long periods of time have led the profession to focus on specifications of the one-sector neoclassical growth model that generate balanced growth. The evidence that we presented in Section 6.2 suggests that the continuing process of reallocation of activity across sectors coexists with the stable behavior of aggregate variables that characterizes balanced growth. It is therefore perhaps not surprising that the theoretical literature on structural transformation has looked for specifications of the previous model that give rise to a generalized balanced growth path along which structural transformation occurs. We begin this section by summarizing the results of this theoretical literature and its predictions for the nature of structural transformation. We close this section with a discussion of whether the focus on specifications that deliver exact balanced growth might be too stringent. Irrespective of whether this is the case, we believe that the search for specifications that deliver balanced growth and structural transformation has proven useful in helping researchers isolate various forces that are potentially important in shaping structural transformation.

6.4.1 Two Special Cases with Analytical Solutions

Our previous derivations put us in a position to easily summarize recent findings in the literature about the joint possibility of generalized balanced growth and structural transformation. In this subsection, we focus on two recent papers that emphasize different economic forces behind structural transformation, notably Kongsamut et al. (2001) and Ngai and Pissarides (2007).

6.4.1.1 Preliminaries

If we are to look for a balanced growth path it is natural to limit ourselves to situations in which technological progress is constant. We therefore assume:

$$\frac{A_{it+1}}{A_{it}} = 1 + \gamma_i, \quad i = a, m, c, x. \tag{6.27}$$

As previously noted, even if all aggregates grow at constant rates, it will typically not be the case that all sector-level variables grow at constant rates. We therefore follow the literature and focus on generalized balanced growth paths (GBGP), which are defined to be equilibrium paths along which the rental rate of capital is constant, i.e. $R_t = R$.

We next turn to the issue of whether there are specifications of the model for which a GBGP exists along which structural transformation occurs. At this stage we will simply pose this question from a qualitative perspective. Specifically, we will say that a GBGP exhibits structural transformation if either sectoral employment shares (n_{it}) or sectoral value added (or consumption expenditure) shares ($p_{it}c_{it}/Y_t$) are not constant for all three consumption sectors. The issue of generating the right patterns of structural transformation, both qualitatively and quantitatively, will be taken up later.

As a starting point it is useful to examine two special cases. The first special case makes the extreme assumption that the three consumption goods are perfect substitutes: $\bar{c}_a = \bar{c}_s = 0, \omega_a = \omega_m = \omega_s, A_{at} = A_{mt} = A_{st}$, and $\varepsilon \to \infty$. In this case, the model is identical at the aggregate level to the two-sector model in the previous section, and so it has a unique balanced growth path in terms of C_t and K_t. However, since the three consumption goods are perfect substitutes and have identical production functions, the allocation of labor and capital between the three sectors is indeterminate, beyond the restriction that capital-to-labor ratios must be the same in all sectors with positive output. Because of this indeterminacy it is obviously the case that one can accommodate whatever patterns one desires in terms of changes in either labor allocations or value added shares across sectors. However, since, as we have seen in Section 6.2 above, the features of structural transformation appear to be stable over time and across countries, this does not seem a very appealing way to account for structural transformation.

The second special case of interest assumes that $\bar{c}_a = \bar{c}_s = 0$ and $\varepsilon = 1$, so that the preference aggregator is Cobb Douglas. We do not present the details here, but one can show that the unique balanced growth path has constant sectoral labor and value added shares. This happens despite the fact that we have not restricted the relative rates of productivity growth among the three consumption sectors. Intuitively, with Cobb-Douglas preferences, employment and value added shares are independent of relative productivities. With sectoral employment and capital shares fixed, differences in relative productivities generate differences in relative outputs, but these differences in output are perfectly offset in terms of value added shares by changes in relative prices. While this special case gives rise to balanced growth and avoids the indeterminacy of the previous case, it does not give rise to structural transformation along the balanced growth path.

In what follows, we describe two scenarios that can generate structural transformation along a GBGP. Each of them can be understood as a departure from this second special case.

6.4.1.2 Case 1: Income Effects and Structural Transformation

Case 1 corresponds to the analysis found in Kongsamut et al. (2001) and represents the extreme scenario in which all structural change is driven by income effects that are generated by the non-homotheticity terms $\bar{c}_a$ and $\bar{c}_s$ when income changes but relative prices remain the same. For this case, we assume that technological progress is uniform across all consumption sectors ($\gamma_i = \gamma_j$ for all $i, j = a, m, s$) and that the parameter governing the elasticity of substitution among consumption goods is unity ($\varepsilon = 1$).[22] The consumption aggregator (6.12) then takes the well-known Stone-Geary form:

$$C_t = \omega_a \log (c_{at} - \bar{c}_a) + \omega_m \log (c_{mt}) + \omega_s \log (c_{st} + \bar{c}_s) . \tag{6.28}$$

With $\bar{c}_a$ and $\bar{c}_s$ positive it is easy to see intuitively how one may get structural transformation along a GBGP; as income grows, the non-homotheticity of the demands for the different consumption goods will lead to changes in the value added shares. However, there is a potential issue in obtaining generalized balanced growth when $\bar{c}_a$ and $\bar{c}_s$ are positive. To see this, recall the Euler equation (6.7) for the household problem. From this equation, if R_t

[22] Note that ε equals the elasticity of substitution only if $\bar{c}_a = \bar{c}_s = 0$.

is constant over time, then it must be that $P_t C_t$ grows at a constant rate. From the period-budget equation, (6.23), and noting that factor payments are equal to output, we have:

$$P_t C_t + p_{at}\bar{c}_a - p_{st}\bar{c}_s = K_t^{\theta} A_{xt}^{1-\theta} + (1-\delta)K_t - K_{t+1}. \tag{6.29}$$

Since the right-hand side grows at rate γ_x, $P_t C_t + p_{at}\bar{c}_a - p_{st}\bar{c}_s$ must also grow at rate γ_x. If $p_{a0}\bar{c}_a - p_{s0}\bar{c}_s$ is not zero, then $p_{at}\bar{c}_a - p_{st}\bar{c}_s$ will grow at rate γ_x only if relative prices also grow at rate γ_x. However, this contradicts the fact that p_{at} and p_{st} both grow at gross rate $[(1+\gamma_{xt})/(1+\gamma_{ct})]^{1-\theta}$, which is implied by expression (6.16). Hence, balanced growth requires that $p_{a0}\bar{c}_a - p_{s0}\bar{c}_s = 0$, which is equivalent to:

$$\frac{\bar{c}_a}{\bar{c}_s} = \left(\frac{A_{a0}}{A_{s0}}\right)^{1-\theta}. \tag{6.30}$$

Note that since both relative prices grow at the same rate, this condition implies that $p_{at}\bar{c}_a - p_{st}\bar{c}_s = 0$ at all dates t.[23]

Given condition (6.30), Equation (6.29) simply requires that $P_t C_t$ grows at rate γ_x. From the perspective of balanced growth this economy then looks very much like the two-sector model that we considered in the previous section. In particular, similar to that two-sector model, the share of labor and capital devoted to consumption versus investment is constant along a GBGP.

We make two remarks regarding condition (6.30). First, note that if either of $\bar{c}_a$ or $\bar{c}_s$ is positive, then they must both be positive. As we discuss in a later section, many papers have implicitly assumed that $\bar{c}_a > 0$ and $\bar{c}_s = 0$, which is inconsistent with condition (6.30). Second, this condition relates the parameters of preferences and technology to each other, and is therefore somewhat of a fragile condition. We shall return to this point later in this section.

Next we consider whether structural transformation occurs along the GBGP. To examine this note that if $\varepsilon = 1$, then (6.24) and (6.25) imply the Stone-Geary demand system:

$$c_{at} = \omega_a \frac{P_t C_t}{p_{at}} + \bar{c}_a, \tag{6.31}$$

$$c_{mt} = \omega_m \frac{P_t C_t}{p_{mt}}, \tag{6.32}$$

$$c_{st} = \omega_s \frac{P_t C_t}{p_{st}} - \bar{c}_s. \tag{6.33}$$

[23] This point illustrates that the assumption of the same rate of technological progress in the agriculture and service sectors is a necessary condition and not merely a simplification.

Moreover, the assumption that technology in all consumption sectors grow at the same rate implies that the relative prices of the three consumption goods are constant:

$$\frac{p_{it}}{P_t} = \frac{p_{i0}}{P_0}, \qquad i \in \{a, m, s\}.$$

Hence, c_{at}, c_{mt}, and c_{st} grow at a slower rate, at the same rate, and at a faster rate than C_t, respectively. Given that the relative prices of the three consumption goods are constant, it follows that $p_{it}c_{it}/P_tC_t$ is decreasing for agriculture, constant for manufacturing, and increasing for services. Since total consumption expenditures are a constant share of total output, it follows that these properties also carry over to both n_{it} and $p_{it}c_{it}/Y_t$.

In summary, and more formally, we have the following result:

Proposition 2. *Assume that condition (6.30) holds and that:*

$$\bar{c}_s \leq \omega_s \left(\frac{A_{s0}}{A_{x0}}\right)^{1-\theta} \left[K_0^\theta A_{x0}^{1-\theta} - (\gamma_x + \delta)K_0\right]. \tag{6.34}$$

where K_0 is given by (6.11).

Then there is a unique GBGP. Along the GBGP, the employment and nominal value added shares of the investment sector are constant. The employment and nominal value added shares are decreasing for agriculture, constant for manufacturing, and increasing for services.

Proof. We start by noting that it is straightforward to show that (6.11) implies that $K_0^\theta A_{x0}^{1-\theta} > (\gamma_x + \delta)K_0$. Hence, $P_0C_0 = K_0^\theta A_{x0}^{1-\theta} - (\gamma_x + \delta)K_0 > 0$ and condition (6.34) is well-defined. Condition (6.34) ensures that the right-hand side of (6.33) is positive at $t = 0$. Since the economy grows while relative prices remain constant, this implies that the right-hand side is positive for all t. In this case, Equations (6.31)–(6.33) are well defined and they have a unique interior solution for c_{at}, c_{mt}, c_{st}. The existence of a unique GBGP and the statements about the shares then follow directly from the previous discussion. □

6.4.1.3 Case 2: Relative Price Effects and Structural Transformation

The second case that we consider corresponds to the analysis found in Ngai and Pissarides (2007).[24] Whereas the previous case generated structural transformation purely via income changes and asked whether this could be consistent with balanced growth, Ngai and Pissarides consider the polar extreme case in which structural transformation is generated purely from changes in relative prices and ask whether this can be consistent with balanced growth. Accordingly, they assume that $\bar{c}_a = \bar{c}_s = 0$. In order to have relative price changes operating it is clearly necessary to have differential rates of technological progress among the three consumption goods sectors, so no restrictions will be placed on the relative

[24] This work builds on the important earlier contribution of Baumol (1967).

values of γ_i. Given our earlier discussion, however, we know that ε will have to take on a value other than unity.

The analysis of this case follows directly from our analysis of the two-sector model. Specifically, if the values of γ_a, γ_m, and γ_s are different, then the price index P_t will not grow at a constant rate. However, as noted at the end of the section on the two-sector model, this has no bearing on the existence of a unique GBGP; there still is a unique GBGP that features a constant share of labor and capital allocated to total consumption. Along the GBGP, the value of $P_t C_t$ will grow at the constant rate γ_x even though neither component grows at a constant rate.

To assess the implications for structural transformation we again turn to Equations (6.24) and (6.25). Using Equation (6.16) for relative prices, these two equations can now be written as:

$$\frac{c_{at}}{c_{mt}} = \frac{\omega_a}{\omega_m}\left(\frac{A_{at}}{A_{mt}}\right)^{\varepsilon(1-\theta)}, \tag{6.35}$$

$$\frac{c_{st}}{c_{mt}} = \frac{\omega_s}{\omega_m}\left(\frac{A_{st}}{A_{mt}}\right)^{\varepsilon(1-\theta)}. \tag{6.36}$$

Noting that $c_{it} = K_t^{\theta} A_{it}^{1-\theta} n_{it}$, we also have:

$$\frac{n_{at}}{n_{mt}} = \frac{\omega_a}{\omega_m}\left(\frac{A_{mt}}{A_{at}}\right)^{(1-\varepsilon)(1-\theta)}, \tag{6.37}$$

$$\frac{n_{st}}{n_{mt}} = \frac{\omega_s}{\omega_m}\left(\frac{A_{mt}}{A_{st}}\right)^{(1-\varepsilon)(1-\theta)}. \tag{6.38}$$

Recalling that labor allocated to the overall consumption sector is constant, it follows that if $\varepsilon = 1$, we have the earlier result that the n_{it} are constant in each of the three consumption sectors. So too are the values of $p_{it}c_{it}/P_t C_t$ and $p_{it}c_{it}/Y_t$. If ε differs from one, then the model can generate structural transformation along a GBGP as long as the rates of technological progress differ among the three consumption sectors. In contrast to Case 1, it is not true in this case that c_{mt} is a constant proportion of C_t, nor is true that C_t grows at a constant rate. Without imposing some additional structure, one cannot say more about the nature of structural transformation that occurs.

To simplify exposition, we focus on the special case in which technological progress is strongest in agriculture and weakest in services, that is, $\gamma_a > \gamma_m > \gamma_s$. If, in addition, we assume that $\varepsilon < 1$, then the above expressions imply that along a GBGP the values of $n_{it}, p_{it}c_{it}/P_t C_t$ and $p_{it}c_{it}/Y_t$ are decreasing for agriculture and increasing for services. The behavior of these values for manufacturing is ambiguous in terms of the direction of change, but the size of the change is bounded by the sizes of the change in the other two sectors. Proposition 5 of Ngai and Pissarides (2007) shows that the evolution of n_m in this case, will be either monotonically decreasing or hump-shaped.

More formally, we summarize the above discussion with the following proposition.

Proposition 3. *Let $\bar{c}_a = \bar{c}_s = 0, \varepsilon < 1, \gamma_a > \gamma_m > \gamma_s > 0$, and $\gamma_x > 0$.*

There is a unique GBGP. Along the GBGP, the shares of employment and nominal value added (in current prices) of the investment sector are constant; the shares of employment and nominal value added (in current prices) of the consumption sectors behave as follows: the agricultural shares decline; the services shares rise; the manufacturing shares decrease less than the agricultural shares and increase less than the service shares.

6.4.1.4 Qualitative Assessment

The previous subsections outlined two different theories of structural transformation in the context of generalized balanced growth. Although we postpone a more rigorous assessment of the economic mechanisms implicit in these two theories until a later section, it is still of interest at this point to assess the extent to which each of the theories taken individually can account for some of the broad patterns that we documented in Section 6.2. We will see that while each theory can qualitatively account for some of the patterns found earlier, each also has some limitations.

Given the qualifications that we have noted previously in connecting the model with data, we keep in mind that we can either connect the production measures (employment shares and value added shares) to the data, implying that the consumption measure (consumption expenditure shares) does not have a close empirical counterpart, or we can connect the consumption measure to the data, implying that the two production measures do not have close empirical counterparts. Whichever way we choose to proceed, our benchmark model will not be able to make statements about all three measures of structural transformation at the same time.

We begin with the model of Kongsamut et al. (2001). Since the investment sector uses a constant share of labor and accounts for a constant share of (nominal) output, it will not influence the trend behavior of any quantities if it is allocated across the three sectors in constant proportions. Assuming this and starting with the nominal production measures, we conclude that the model can account for the increase in the service sector shares and the decrease in the agricultural sector measures along its GBGP, but it does not generate a hump shape for the manufacturing sector measures. If one allows for the investment share of manufacturing to decrease over time, as is true in the US data, then the model could generate a decline in both production measures for manufacturing. The increasing share of services in investment would only accentuate the rising employment and nominal value added shares for services. Turning to the nominal consumption expenditure measures, the model can account for the increase in the service share, the near constancy of the manufacturing share, and the decrease in the agricultural share.

The model of Kongsamut et al. (2001) has two additional implications that are counterfactual. First, along its generalized balanced growth relative prices need to be constant. It follows that along a GBGP the real measures of structural transformation must display exactly the same properties as the nominal measures, which means that the model cannot account for the quantitative differences between the nominal and the real measures. Second, the model of Kongsamut et al. (2001) implies that in sufficiently poor economies, the household will consume a zero quantity of services and employment in services will also be zero. In contrast, we saw in Section 6.2 that even in the poorest countries service employment and value added are bounded away from zero.

Next we turn to the model of Ngai and Pissarides (2007). Once again we note that since along the GBGP the share of labor devoted to investment is constant and the nominal share of investment in output is constant, any constant allocation of investment across the three sectors will not influence of the trend properties. In this case, given the previously assumed ranking for the rates of technological progress, we conclude that structural transformation along the model's GBGP is qualitatively consistent with the evidence for employment and nominal value added shares in both agriculture and services. While the model does not necessarily deliver a hump shape for the manufacturing shares of employment and nominal valued added, it can deliver this for certain parameter values. Turning to the nominal consumption expenditure measures, the model can account for the increase in the service share and the decrease in the agricultural share, and can qualitatively produce hump-shaped dynamics for manufacturing, though this is not guaranteed.

However, the model of Ngai and Pissarides (2007) cannot account for the behavior of all real shares, irrespective of whether we use production or consumption related measures. In particular, given the assumptions about relative TFPs and the CES utility function being inelastic—i.e. $\varepsilon \in [0, 1)$, the model cannot generate the decreases in the real quantities of agriculture and manufacturing relative to services that we documented in Section 6.2 above. The reason for this is that with a CES utility function, nominal and real shares necessarily move in opposing directions. Given that the model accounts for the relative decline of the nominal shares of agriculture and manufacturing, this implies that it cannot account for the relative decline of the real shares. To see why nominal and real shares move in opposite directions, consider the implications of a decrease in the price of manufacturing relative to services. If $\varepsilon \in [0, 1)$, then the nominal quantity of manufacturing decreases relative to that of services whereas the real quantity of manufactured goods relative to services remains the same if $\varepsilon = 0$ and increases if $\varepsilon \in (0, 1)$.

In summary, although each of these two specifications can account for some of the qualitative patterns that we documented previously, neither of them is able to match all of the patterns. However, the previous discussion suggests that a model featuring both income and relative price effects might successfully match all of the patterns. For example, adding non-homotheticities to the Ngai-Pissarides model could, in principle, allow the model to generate a decrease in the quantity of manufacturing relative to services. While such a specification would not permit a balanced growth path, this is a more general issue to which we will return to later in this section.

6.4.2 Alternative Specifications

In the preceding analysis, we have summarized the results from two papers regarding the possibility of simultaneously having structural transformation and generalized balanced growth. We chose these two papers because they illustrate two different channels through which expenditure shares may change over time: income changes and relative price changes. In this subsection we describe some alternative formulations of these two channels that have appeared in the literature.

6.4.2.1 *Other Specifications Emphasizing the Effects of Income Changes*

Above we chose a specification of preferences where the effects of income changes on expenditure shares were captured by the non-homotheticity terms $\bar{c}_a$ and $\bar{c}_s$. While we think that this is a tractable and transparent way of introducing income effects, there are several alternative specifications of non-homothetic preferences in the literature. Here, we discuss some examples.

In the first quantitative analysis of structural transformation within the framework of the growth model, Echevarria (1997) generated effects from changes in income by using the following alternative specification of the intertemporal utility function:

$$\sum_{t=0}^{\infty} \beta^t \left[\alpha_a \log c_a + \alpha_m \log c_m + \alpha_s \log c_s - \eta \left(\frac{1}{c_a^{\rho_a}} + \frac{1}{c_m^{\rho_m}} + \frac{1}{c_s^{\rho_s}} \right) \right],$$

where $\alpha_i > 0, \eta, \rho_i \geq 0$. If $\eta = 0$ then the preferences reduce to a Cobb-Douglas specification, but if $\eta > 0$ and at least one of the $\rho_i > 0$, then the preferences are not homothetic. To see some of the features of this specification it is useful to examine the properties of the marginal utility of good i, which is given by:

$$MU_i(c_i) = \alpha_i c_i^{-1} + \eta \rho_i c_i^{-1-\rho_i}. \tag{6.39}$$

Note first that the marginal utility of each good will be infinite for zero consumption quantities, implying that the household chooses interior consumption quantities. The second term is positive if $\eta\rho_i > 0$. In this case, it goes to infinity as c_i becomes arbitrarily small and it goes to zero as c_i becomes arbitrarily large.

If, as in Echevarria's calibration, $\eta > 0$ and $\rho_a > \rho_m > \rho_s = 0$, then at low levels of income (and hence of consumption), there is a force in favor of higher c_a and c_m and of lower c_s, and the force is stronger for c_a than for c_m. In contrast, at high levels of income this force disappears. Intuitively, one can use the parameters η and ρ_i to achieve the same qualitative effects that are generated by the parameters $\bar{c}_a$ and $\bar{c}_s$ in our benchmark model.

The main advantage of Echevarria's specification of period utility is that an interior solution to the static period problem exists for any positive level of income. This is in contrast to what happens in our benchmark model, since if $\bar{c}_a > 0$ and the present value of income is lower than the present value of $\{p_{at}\bar{c}_a\}$, then the household cannot afford to purchase at least $\bar{c}_a$ units of the agricultural good in all periods and our period utility will not be defined in at least one period. From an analytical perspective, the disadvantage of Echevarria's specification is that it is not consistent with generalized balanced growth. The reason for this is the presence of the term $\eta c_j^{-\rho_j}$ in the period utility function. If $\eta = 0$, then period utility is of the homothetic log form and a GBGP exists. In contrast, if $\eta > 0$, then it is impossible for the value of total consumption, $\sum_{j\in\{a,m,s\}} p_{jt}c_{jt}$, to grow at the same constant rate at which technological progress grows. As we saw in Section 6.3.2 above, this would be required for a GBGP with constant real interest rate to exist.

A recent paper by Boppart (2011) explores more general preferences that are consistent with balanced growth. In particular, Boppart specifies indirect period utility functions that fall into the class of "price-independent-generalized-linearity" preferences defined by Muellbauer (1975, 1976). These preferences are more general than Gorman preferences in that they generate nonlinear Engel curves. Nonetheless, they aggregate and allow for a stand-in household. There are two advantages of using price-independent-generalized-linearity preferences in the context of structural transformation. First, they avoid the awkward feature of our benchmark specification that can lead to utility not being defined for sufficiently small income. Second, as Boppart establishes, they are consistent with balanced growth if the technology side is as we specified it above.

A different approach to generating effects from changes in income is Foellmi and Zweimüller (2008). Whereas our benchmark model implicitly aggregated individual consumption goods into three broad sectors and defined preferences over the amounts of the three resulting aggregates, these authors specify preferences over an unbounded mass of potential consumption goods. Preferences are such that for each good, marginal utility is finite at zero consumption and decreases to zero at some finite satiation level of consumption. Over time, as income increases, the mass of goods that are consumed increases, so that there is adjustment along both the intensive and the extensive margin. The order in which the goods will be introduced is uniquely determined by the model's primitives: all of the goods are symmetric from the perspective of production but are given different weights in preferences.[25]

The fact that new goods are consumed over time implies that labor will necessarily be reallocated across activities over time. In terms of basic economic forces, the key mechanism at work comes from the fact that different goods have different income elasticities. Different than in the specification of our benchmark model, however, any particular good in this model will have an income elasticity of zero asymptotically since at some date satiation will be reached.

[25] This type of preferences is sometimes called hierarchical preferences. It was first used by Murphy et al. (1989).

In order to connect their model to the standard facts of structural transformation, Foellmi and Zweimüller (2008) need to map individual goods into the three broad sectors. If they assume that agricultural goods are disproportionately the goods with high weights, that services are disproportionately the goods with low weights, and that manufacturing goods lie "in between" these two, then they can match the qualitative patterns presented earlier. As income grows and more of the less weighted goods are consumed, one obtains a declining share for agricultural goods, an increasing share for services, and a hump-shaped pattern for manufacturing. Foellmi and Zweimüller can also generate balanced growth with relatively standard assumptions. Specifically, if they assume that the weighting function on different goods has a power form and there is constant labor-augmenting technological progress that is common to the production of all goods, then their model gives rise to a GBGP. As they discuss in their paper, the assumption of a power function for the weighting function is analogous to the assumption of a constant elasticity utility function in the context of the standard one-sector growth model.

Relative to the results that we derived previously about income changes and structural transformation, the specification of Foellmi and Zweimüller (2008) delivers balanced growth and structural transformation in a more robust manner, in the sense that it does not need a condition similar to (6.30) that imposes a restriction on the parameters of preferences and technology. Moreover, it can also deliver a hump-shaped relationship between GDP per capita and the manufacturing shares. But a limitation of the specification of Foellmi and Zweimüller (2008) is that modeling structural transformation at the level of individual goods does not provide much guidance for how to connect the model with data at the level of broad sectors.[26]

Hall and Jones (2007) also develop a framework that can give rise to non-homothetic demand functions, though their focus is specifically on the rise of spending on health care, as opposed to the more general process of structural transformation. Nonetheless, this is of interest in the current context since increases in health care account for a significant part of the overall increase in the size of the service sector. In the basic model of Hall and Jones, utility in the current period is derived from a single good that represents all non-health consumption. The period utility function is homothetic and health consumption in period t provides no direct utility flow in period t but does influence the probability of survival to the next period. Intuitively, this model has features akin to the model with intensive-extensive margins that we discussed above. Specifically, a household can adjust

[26] Buera and Kaboski (2012a,b) adopt a similar preference structure as Foellmi and Zweimüller (2008), except that they stress the introduction of new goods and adjustment along the extensive margin. Other aspects of their analysis are quite different, however. We discuss their model in more detail later in this section and again in Section 6.7.6. For now we simply note that Buera and Kaboski (2012a) derive an explicit mapping from their preferences to a reduced-from representation of preferences over goods and services. The interesting feature of this mapping is that it includes a term that is analogous to our term $\bar{c}_s$, but rather than being a constant, its value changes over time as technological progress occurs.

along the intensive margin by spending more on consumption, or along the extensive margin by spending more on health care and therefore increasing the expected number of periods in which consumption occurs. As the level of consumption increases, the marginal utility from additional consumption at the intensive margin decreases relative to the marginal utility of living an additional period. This can generate an increasing expenditure share for health consumption as incomes rise, and therefore look like a model that features a non-homothetic period utility function over health and non-health consumption.[27]

6.4.2.2 Other Specifications Emphasizing Relative Price Effects

In the Ngai-Pissarides model analyzed as Case 2 above, sectoral reallocation of factors of production and nominal value added shares occurred as a result of relative output price changes along the balanced growth path. Relative price changes were in turn generated by having differential rates of technological progress across sectors. The literature has also noted that relative output price changes can result from changes in the relative prices of inputs if sectors vary in the intensity with which they use inputs and there are changes in the relative supply of factors. In this case, one can generate structural transformation via relative price changes even if technological change is neutral.

Two papers in the literature stress this mechanism. Caselli and Coleman (2001) focus on skilled and unskilled workers as the two inputs of interest, noting that non-agriculture is more skill-intensive than agriculture. They argue that the effective cost of education decreased in the first half of the 20th century, thereby increasing the relative supply of skilled workers, decreasing the relative price of non-agricultural goods, and moving resources out of agriculture.[28] Acemoglu and Guerrieri (2008) consider capital and labor as the two inputs of interest, and assume that sectors differ in their capital intensity. Since growth driven by technological change is associated with an increase in the capital-to-labor ratio, changes in relative supplies of capital and labor arise quite naturally.[29]

Here we sketch the basic idea within our benchmark model. Since the economics of the model of Acemoglu and Guerrieri (2008) is closest to that of Ngai and Pissarides (2007), except that the underlying cause of the relative price movements is different, we illustrate the basic idea by focusing on the implications for structural transformation of differences in the sectoral capital intensities. We assume that TFP growth is uniform across

[27] In a recent paper, Lawver (2011) uses a version of the model of Hall and Jones (2007) to measure the increase in the quality of health consumption.

[28] In Section 6.7.2, we will revisit this paper and discuss its implications for income convergence between regions.

[29] In a different context, Bar and Leukhina (2010) argue that non-agriculture is more labor intensive than agriculture, and that the increase in population associated with the demographic transition could help explain the initial expansion of the non-agricultural sector in the context of England during the time of the Industrial Revolution.

the three consumption sectors and define A_t by $A_t \equiv A_{it}^{1-\theta_i}$ for $i \in \{a, m, s\}$. The capital intensities differ across sectors so that the sectoral production functions (6.13) become:

$$c_{it} = A_t k_{it}^{\theta_i} n_{it}^{1-\theta_i}, \quad i \in \{a, m, s\}. \tag{6.40}$$

All other features of the environment are the same as in the benchmark model described earlier.

The first-order conditions for the stand-in firm in sector $i \in \{a, m, s\}$ are now given by:

$$R_t = p_{it}\theta_i A_t \left(\frac{k_{it}}{n_{it}}\right)^{\theta_i - 1}, \tag{6.41}$$

$$W_t = p_{it}(1-\theta_i)A_t \left(\frac{k_{it}}{n_{it}}\right)^{\theta_i}. \tag{6.42}$$

Dividing these equations by each other gives:

$$\frac{1-\theta_i}{\theta_i}\frac{k_{it}}{n_{it}} = \frac{1-\theta_j}{\theta_j}\frac{k_{jt}}{n_{jt}}. \tag{6.43}$$

Two implications follow from this equation. First, sectors with larger capital shares have larger capital-to-labor ratios; second, the capital-to-labor ratio grows at the same rate in all sectors.

To derive the implications for relative prices, substitute (6.43) into (6.42) and rearrange to obtain:

$$\frac{p_{it}}{p_{jt}} = \Omega_{ij}\left(\frac{k_{it}}{n_{it}}\right)^{\theta_j - \theta_i} \qquad i, j \in \{a, m, s\}, \tag{6.44}$$

where Ω_{ij} is a constant that depends on the capital shares. Since the capital-to-labor ratios of all sectors grow at the same rate, Equation (6.44) implies that for any pair of sectors, the relative price of the sector with the higher capital share decreases as the aggregate capital stock grows. If one assumes:

$$\theta_a > \theta_m > \theta_s, \tag{6.45}$$

it follows that the price of services relative to manufacturing and of manufacturing relative to agriculture will both increase over time. This implication is of course analogous to what we derived in the context of the Ngai-Pissarides model when we assumed that $\gamma_a > \gamma_m > \gamma_s$.

It is important to note that the mechanism of Acemoglu and Guerrieri (2008) relies not only on differences in the sectoral capital intensities, but also on the fact that with Cobb-Douglas production functions the elasticity of substitution between capital and labor is equal to one. Indeed, Alvarez-Cuadrado et al. (2012) have recently pointed out that the relative price of sectoral output depends not only on sectoral TFP and capital intensity, but also on the elasticity of substitution between capital and labor. To see how the elasticity of substitution matters in this context, consider first the extreme case in which capital and labor are perfect substitutes. The capital intensity then does not matter at all for relative prices because firms can perfectly substitute labor for capital when capital is relatively expensive. In the other extreme case, capital and labor are perfect complements and the production function is of the Leontief form. The capital intensity then matters crucially for relative prices because one cannot substitute labor for capital when capital is relatively expensive. More generally, the effects of Acemoglu and Guerrieri (2008) are more important if the sectoral elasticity of substitution is smaller.

Although the specification of Acemoglu and Guerrieri (2008) can account for the changes in nominal value added shares, it cannot account for the changes in real value added shares. Moreover, it cannot generate the patterns in sectoral employment shares either.[30] To see why, note that using (6.43), it is straightforward to show that:

$$K = \left(\sum_{j=x,a,m,s} \frac{\theta_j}{1-\theta_j} n_j \right) \frac{1-\theta_i}{\theta_i} \frac{k_i}{n_i}. \tag{6.46}$$

Solving this expression for k_i/n_i and substituting the result into Equation (6.40) gives:

$$c_{it} = A_t K_t^{\theta_i} \left(\frac{\frac{\theta_i}{1-\theta_i}}{\sum \frac{\theta_j}{1-\theta_j} n_j} \right) n_{it}, \quad i \in \{a, m, s\}.$$

In the polar case of Leontief utility, c_{it}/c_{jt} is constant, so the previous equation implies that n_{it}/n_{jt} is constant too. For positive elasticities of substitution, changes in relative quantities are in the opposite direction of changes in relative prices. In other words, in the model of Acemoglu and Guerrieri, there cannot be structural transformation in terms of employment that is consistent with the fact that service employment increased at the same time as which its relative price increased too.

One important additional difference relative to the specification of Ngai and Pissarides (2007) is that the model of Acemoglu and Guerrieri (2008) has exact GBGP only asymptotically, and so the best we can hope for in this model is approximate generalized balanced growth. Below we discuss the difference between approximate and exact generalized balanced growth in more detail.

6.4.2.3 An Alternative View of Structural Transformation

In two recent papers, Buera and Kaboski (2012a,b) have offered a novel representation of structural transformation that implicitly involves elements of both of the special cases discussed previously. Here we offer a simple version of their framework to illustrate the forces at work. In Section 6.7.6, we discuss their specific implications in more detail.

[30] A similar issue is also present in Ngai and Pissarides (2007). We will discuss this in more detail later.

They consider an economy in which there are a continuum of services and a continuum of goods. For simplicity, in their economy goods are only useful as an input into the production of services, and each good is uniquely associated with the production of a specific service. Specifically, each good is produced using labor, and each service is produced using labor and its corresponding specialized good. They adopt a similar preference structure as Foellmi and Zweimüller (2008), but they assume that each service can only be consumed in the amounts of zero or one, so that increasing consumption will necessarily manifest itself along the extensive margin. From the consumer's perspective all services are symmetric. Consider the following special case of this structure as a benchmark. Assume a single household with one unit of time. Index the continuum of goods and services by z. The technology for producing each good z at time t is $g(z) = A(t)h_g(z)$, where $A(t)$ captures labor-augmenting technological change. The technology for producing each service z at time t is Leontief: $s(z) = \min\{A(t)\frac{1}{a}h_s(z), g(z)\}$, where $A(t)$ is the same in both production functions. Because each service is consumed in amount 1, it takes $(1 + a)/A(t)$ units of labor to produce one unit of service, so that total consumption (i.e. the total number of services that are consumed) will be given by $A(t)/(1 + a)$, and a fraction $1/(1 + a)$ of labor will be devoted to the goods sector. So, in this benchmark economy there is no structural transformation in terms of labor allocations between goods and services.

Buera and Kaboski generate interesting implications in this setting by extending it along two dimensions. First, they introduce the possibility of home produced services which also require labor and the specialized good. To create an interesting tradeoff between the choice of whether to produce a given service in the home or in the market, they assume that market production of services is more efficient. This could be modeled in different ways and differs in their two papers. To illustrate some basic workings of the model we assume that market production takes less of the good per unit of output, but that home produced services supply a proportionately higher utility flow. An illustrative example would be the choice between home produced transportation services (buying a car and driving yourself) versus market provided transportation services (buses, or taxis). While having a car increases convenience, the car will also be idle for considerable periods. Second, they introduce heterogeneity into the production side of the economy by assuming that higher z goods require more labor to be produced. This heterogeneity interacts with the choice of whether to produce a given service in the market or the home, since the more expensive it is to produce the durable, the greater is the penalty for home production which requires more of the durable per unit of output. Whether a good is produced in the home or the market in turn has implications for observed allocations of labor and market value added across market sectors, since having home produced services requires labor from the goods sector, but will not use any labor in the market service sector. In their model, as an economy develops the marginal services that are added represent services with higher benefits to market versus home production. The combination of technological change plus the changing nature of the marginal services being brought into the economy can introduce interesting dynamics for how activity shifts between the market and home sectors. If production shifts toward the market and away from the home, this will be recorded as an increase in the size of the market service sector relative to the goods sector.

As noted earlier, models with these types of preferences necessarily embody a nonhomotheticity. But the production heterogeneity in this model implicitly acts like differential technological growth across sectors since the marginal services that are added as an economy grows have differing relative productivity for home versus market production. A general message from this framework is that when thinking about growth and structural transformation it is important to think about the new goods and services that are associated with growth, and the movement of delivery of certain services between the home and market sectors, since the changing nature of activities in the market sector can have important implications for the measured sectoral allocation of market activity.

6.4.3 Approximate versus Exact Generalized Balanced Growth

Up to this point, our discussion has focused on analytic results concerning the possibility of jointly having generalized balanced growth and structural transformation. This is a natural starting point given the emphasis that the literature using the one-sector growth model places on balanced growth and that conditions under which balanced growth results in the one-sector model, are relatively weak—constant returns to scale production with labor-augmenting technical change and a period utility function with a constant intertemporal elasticity of substitution. The results that we have presented above for multi-sector models, however, have made it apparent that the conditions for jointly having generalized balanced growth and structural transformation become considerably more stringent—we now need that all production functions are Cobb-Douglas with the same capital share, that the period utility function exhibits a unitary elasticity of substitution, and in some cases that there is a particular relationship between preference and technology parameters. To the extent that there is good reason to believe that many of these conditions are not satisfied, models that impose them may be missing some key features of reality. In fact, some authors have dismissed income changes as an important source of structural transformation on the grounds that they are consistent with generalized balanced growth only under very fragile cross-restrictions on technology and preferences such as the one imposed in (6.30).

The previous discussion suggests that it may be ill advised to insist on generalized balanced growth in the context of structural transformation. To the extent that (generalized) balanced growth is merely a good approximation to what we see in the data in various countries over long periods of time, the more relevant question is whether there are specifications that can deliver structural transformation and approximate generalized balanced growth, which may occur under much less stringent conditions than exact generalized balanced growth.

To date there has not been much systematic analysis of the extent to which approximate generalized balanced growth is a robust feature of multi-sector versions of the growth model along the lines of those that we have considered. But several cases in the literature suggest that approximate generalized balanced growth may in fact be quite robust. To begin with, Kongsamut et al. (2001) consider numerical examples that depart from the exact conditions needed for generalized balanced growth in their setting and find that the equilibrium path does not deviate much from generalized balanced growth. In a similar context, Gollin et al. (2002) study a two-sector model with subsistence consumption in the agricultural sector but not in the other sector—a clear violation of the conditions needed to generate GBGP, but find relatively small variations of the interest rate when their model is calibrated to match the US data over the post 1950 period. Moreover, although the model in Acemoglu and Guerrieri (2008) only has an asymptotic GBGP, the results that they report for numerical simulations suggest that the model's behavior along a transition path is not that different from balanced growth.

The models just discussed have the feature that asymptotically structural transformation ceases to occur. For example, if structural transformation occurs as the result of the non-homothetic terms $\bar{c}_a$ and $\bar{c}_s$, then productivity increases will imply that in the limit the size of the two non-homothetic terms becomes arbitrarily small relative to consumption. Since we observe (approximate) balanced growth and structural transformation over very long periods in the data, it follows that any model that generates structural transformation purely while it is converging to an exact balanced growth path must have very long-lived dynamics in order to capture reality.[31]

6.5. THE ECONOMIC FORCES BEHIND STRUCTURAL TRANSFORMATION: EMPIRICAL ANALYSIS

The previous section has focused on models that could generate (approximate) generalized balanced growth and structural transformation as simultaneous outcomes. The various models that we reviewed emphasize different theories for the reallocation of activity across sectors that accompanies growth. In one class of theories, the key driving force is uniform technological progress, and the key propagation mechanism comes from income effects. In another class of theories, the key driving force is technological progress that differs across sectors and the key propagation mechanism comes from relative price effects in consumption. In a third class of models, the driving force is again uniform technological progress, but the propagation mechanism is a combination of different capital intensities or elasticities of substitution in production and relative price effects in consumption.

[31] Note that this statement does not apply to the model of Ngai and Pissarides (2007) which exhibits structural transformation both along the exact balanced growth path and in the limit.

Rather than focusing narrowly on the conditions required to generate exact balanced growth, we believe that the key to developing quantitative theories of structural transformation is to develop quantitative assessments of the various driving forces and propagation mechanisms that the literature has identified as potentially important. In this section we summarize the recent progress in this effort. We break this section into two subsections. The first considers the direct evidence regarding differences in rates of technological progress, capital intensities, and elasticities of substitution. The second considers the more general issue of the relative importance of the effects coming from changes in income and changes in relative prices.

6.5.1 Technological Differences Across Sectors

In this subsection we consider the evidence regarding technological differences across sectors along the two dimensions highlighted by the previous theories: differences in technological progress and differences in capital shares and in elasticities of substitution. We also assess the extent to which these differences are appropriate to generate the qualitative features found in the data regarding structural transformation.

6.5.1.1 Sectoral TFP Growth

Assumptions about TFP growth at the sectoral level played an important role in both of the theories of structural transformation that we highlighted. It is therefore of interest to ask what the empirical evidence is regarding relative growth rates in sectoral TFP. Although this would seem to be a relatively straightforward exercise, it is actually challenging to verify the properties of TFP growth in sectoral value added production functions in a cross-country setting. The main reason is that calculating sectoral TFPs requires data on real value added, capital and labor inputs, and the factor shares at the sector level. Unfortunately, these data are unavailable for most countries. One of the many issues is that in order to compute real value added one must have data on the real quantity of intermediate inputs, not just the value of intermediate inputs.

One data set that has the necessary information for a set of countries is EU KLEMS.[32] We begin, therefore, by using the EU KLEMS data starting in 1970 to compute TFP in the production of value added in agriculture, manufacturing, and services for the same set of countries as in Section 6.2: Australia, Canada, Japan, Korea, and the US; as well as the aggregate of 10 EU countries.[33] Figure 6.10 plots the sectoral TFPs for these countries. Given that we are interested in growth rates of TFP, we normalize TFP in 1990 for all

[32] See Timmer et al. (2010), particularly the chapter on structural change, for further discussion of the details of the EU KLEMS data on multifactor productivity. See also Duarte and Restuccia (2010) who document similar facts about TFP as we do here.

[33] The 10 EU countries are the EU member states for which EU KLEMS performs growth accounting: Austria, Belgium, Denmark, Spain, Finland, France, Germany, Italy, the Netherlands, and the United Kingdom.

sectors in all countries to be one. One message that emerges from Figure 6.10 is that there are indeed substantial differences in the growth rates of TFP across sectors. Moreover, we can see that the conditions of Ngai and Pissarides (2007) broadly hold for Australia, Canada, the EU 10, and the United States; averaging over the time period 1970–2007, TFP in agriculture shows the strongest growth while TFP in services shows the weakest growth. This is exactly what is needed for the observed reallocation of employment out of agriculture and manufacturing into the service sector in the model of Ngai and Pissarides (2007).

While data limitations make it difficult to obtain long time series evidence on sectoral TFP for a large sample of countries, our theory suggests an alternative method which requires fewer data. Specifically, in the analysis of our benchmark model we highlighted the fact that if sectoral production functions are Cobb-Douglas with equal capital shares then there is a direct inverse relationship in equilibrium between changes in relative prices and changes in relative productivities. Given appropriate data on prices, one could use this relationship to infer changes in relative productivity. Since long time series of price data is much more readily available than the data needed to measure TFP directly, this is an appealing alternative. However, in addition to requiring the assumption of Cobb-Douglas production functions with equal capital shares, there are two limitations to be noted. First, in our model we assumed that technological change was the only factor that varied over time. One can easily imagine policies or regulations that may also affect relative prices across sectors. If these factors are important for some countries during some periods, it may be misleading to assume that all relative price changes are driven by changes in relative productivities. Second, although price data do exist going quite far back in time, the price data that is required to infer relative productivity growth in value added production functions is the price per unit of value added. In contrast, in practice most available price indices correspond to final goods or to gross output.

Having noted these qualifications, we turn to the evidence documented by Alvarez-Cuadrado and Poschke (2011) about time series changes in the relative price of agriculture to non-agriculture for 11 advanced countries over the last two centuries. A key feature of these data is that the price of agriculture relative to non-agriculture changed its behavior during the last two centuries: while before World War II, it showed an increasing trend, after World War II it started to follow a decreasing trend. Interpreting these changes in relative prices as indicative of changes in relative TFPs, the implication is that prior to World War II, TFP growth in agriculture was actually lower than in non-agriculture.[34] The period before World War II also corresponds to the period that saw the largest movement out of agriculture. In contrast to the findings for data since 1970, the longer time series does not seem to be consistent with relative TFPs driving the labor reallocation from agriculture to non-agriculture.

[34] It should be noted that the evolution of agricultural TFP in Korea between 1970 and 2007 shows a similar U-shaped pattern (see Figure 6.10).

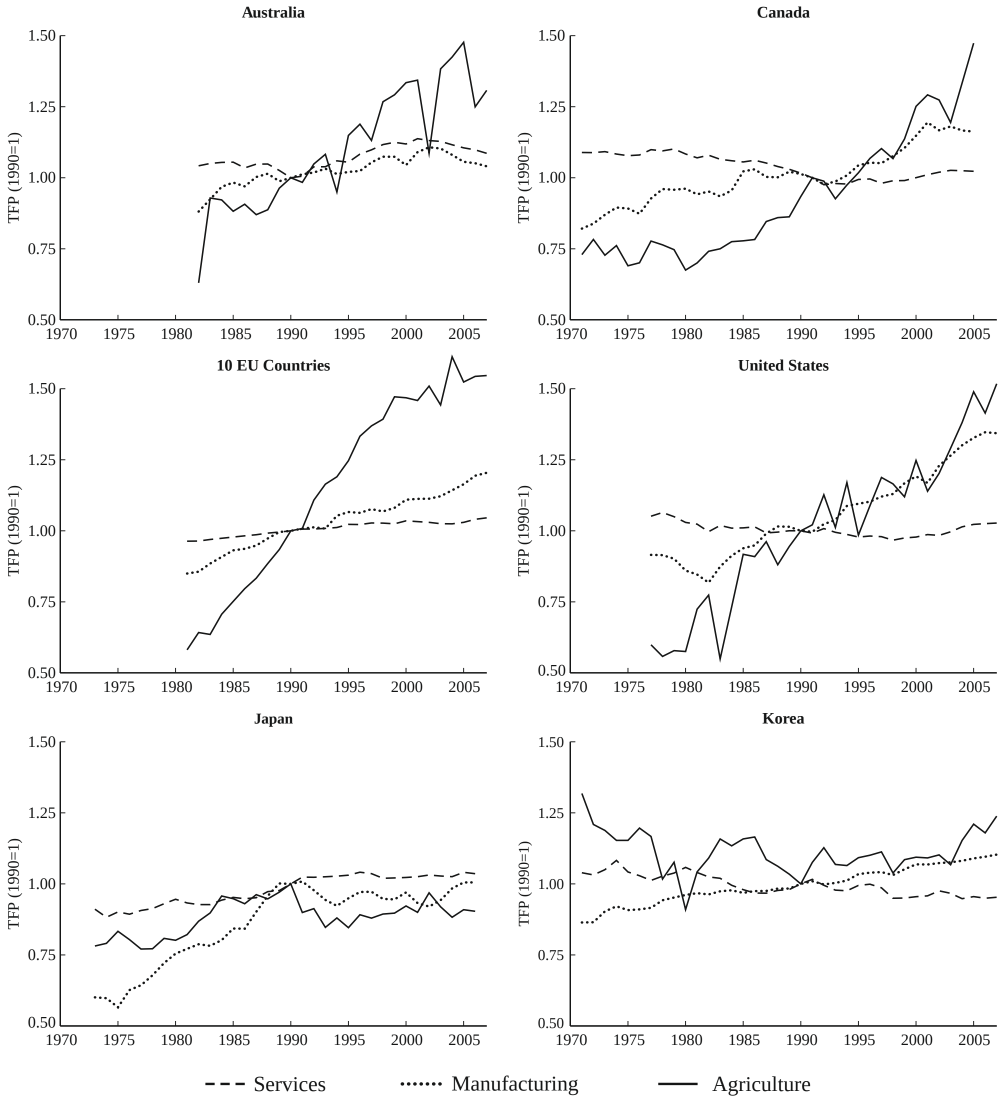


Figure 6.10 Sectoral TFP for selected countries—time series from EU KLEMS 1970–2007. *Source: EU KLEMS, WORLD KLEMS for Korea.*

By way of summary, we think there are two main conclusions that can be drawn from this evidence. First, there are systematic differences in TFP growth rates across sectors. After World War II, these differences appear to be consistent with what is needed to obtain the observed reallocation of employment out of agriculture and manufacturing into the service sector in the model of Ngai and Pissarides (2007). Second, the differences in TFP growth rates across sectors do not appear to be stable over very long periods of time, at least in the case of agriculture versus non-agriculture, which does not bode too well for the models of structural transformation and exact balanced growth that we highlighted previously.

6.5.1.2 Sectoral Differences in Capital Shares and Elasticities of Substitution

Next we consider the existing evidence regarding the potential role of differences in sectoral capital shares, as emphasized by Acemoglu and Guerrieri (2008), and of differences in sectoral elasticity of substitution, as emphasized by Alvarez-Cuadrado et al. (2012). Herrendorf et al. (2013) speak to these questions by assessing how structural transformation is affected by sectoral differences in labor-augmenting technological progress; substitutability between capital and labor; and capital intensity. Using post-war US data on sectoral value added, capital, and labor, they estimate CES production functions and compare them with Cobb-Douglas production functions with different and with equal capital shares. They find that labor-augmenting technological progress is faster in agriculture than in manufacturing and faster in manufacturing than in services; capital and labor are more easily substitutable in agriculture than in manufacturing and more easily substitutable in manufacturing than in services; agriculture is more capital intensive than services and services are more capital intensive than manufacturing.[35]

The findings of Herrendorf et al. (2013) have two implications for the importance of sectoral differences in capital shares and elasticity of substitution as driving forces behind structural transformation. First, in the face of an increasing capital-to-labor ratio, differences in capital shares cause reallocation from agriculture to manufacturing and from services to manufacturing. Second, differences in the elasticity of substitution partly neutralize the differences in the capital shares. In particular, while agriculture has by far the largest capital share it also has the highest substitutability between capital and labor, and in fact agriculture is the only sector for which capital and labor are more substitutable than the Cobb-Douglas case. Herrendorf et al. (2013) show that, as a result, sectoral differences in labor-augmenting technological progress turn out to be the main quantitative force on the technology side behind the post-war US structural transformation, and that this force is well captured by Cobb-Douglas production functions with equal capital shares but different TFP processes.

6.5.2 The Importance of Changes in Income and Relative Prices

Since the theoretical literature has emphasized the effects that result from changes in income and relative prices, it is natural to ask what the data say about these two effects. There are two natural and complementary approaches to this question. In the spirit of

[35] In order to avoid confusion, we stress that these capital shares refer to value added, and not to final expenditure. The capital shares for final expenditure at the sector level can be found in a related paper, Valentinyi and Herrendorf (2008).

our earlier analysis, one approach starts with a stand-in household and uses aggregate data to infer the relative importance of the two different mechanisms. The second approach uses data on individual households to estimate properties of preferences and then assesses the implications for aggregate behavior. In the interest of space, we will focus on the first approach, though we will briefly mention some results from the analysis of micro data. We discuss two recent contributions: Dennis and Iscan (2009) and Herrendorf et al. (2009). The former studies the forces leading to the movement of activity out of agriculture in the United States over the last two centuries, whereas the latter focuses specifically on the reallocation of activity across all three sectors in the United States since 1947. We describe each in turn.

6.5.2.1 The Movement Out of Agriculture in the US Since 1800

Dennis and Iscan (2009) seek to assess the relative importance of income effects, relative TFP growth and capital deepening on the movement of labor out of agriculture in the US over the last two centuries. Their framework is very similar to our benchmark model with the exception of three details. First, they have only two sectors, agriculture and non-agriculture. Second, they assume that all investment comes from the non-agricultural sector. Third, they do not impose that the capital share is the same in both sectors. Initially, Dennis and Iscan write the utility function as the two-sector analog of our utility function, but in their empirical analysis they also allow for the possibility that the subsistence term $\bar{c}_a$ changes over time. Given our earlier discussion, we note that while this general specification is not consistent with generalized balanced growth, it captures the basic forces that the theoretical literature has emphasized.

Dennis and Iscan (2009) derive an equilibrium relationship that expresses the share of labor devoted to agriculture as a function of three factors, which in turn reflect income effects through the subsistence term, relative productivity effects via differential growth rates of TFP, and capital deepening effects. Expressed in terms of our notation, this equilibrium relationship is[36]:

$$1 - n_{at} = \frac{1 - s_a(c_{at})}{1 + p_R(A_{at}, A_{nt}) s_k(k_{at}, k_{nt}) s_X(c_{nt}, X_t)}, \tag{6.47}$$

where:

$$s_a(c_{at}) = \frac{\bar{c}_a}{c_{at}}, \qquad p_R(A_{at}, A_{nt}) = \frac{\omega_a}{\omega_n}\left(\frac{A_{nt}}{A_{at}}\right)^{1-\varepsilon},$$

$$s_k(k_{at}, k_{nt}) = \left(\frac{1-\theta_a}{1-\theta_n}\right)^{\varepsilon}\left(\frac{k_{nt}^{\theta_n}}{k_{at}^{\theta_a}}\right)^{1-\varepsilon}, \qquad s_X(c_{nt}, X_t) = \frac{X_t}{c_{nt} + X_t}.$$

[36] We use the index n for the non-agricultural sector.

The term $1 - s_a(c_{at})$ captures the income effect that operates through the subsistence term $\bar{c}_a$. The terms $p_R(A_{at}, A_{nt})$ and $s_k(k_{at}, k_{nt})$ capture the relative price effects that arise from differential technological progress and capital deepening, respectively, while the term $s_X(c_{nt}, X_t)$ captures the effects associated with changes in the investment rate.

Dennis and Iscan (2009) calibrate the key parameters of the model (elasticity of substitution, subsistence terms, preference weights, and capital shares) and then assess the extent to which Equation (6.47) holds in the data. In particular, they substitute actual values into the right-hand side of Equation (6.47), solve for the implied share of labor allocated to agriculture, and compare this to the actual series from the data. To assess the importance of the different factors, they carry out the same exercise but only allow one of the factors to change over time.

The main findings of Dennis and Iscan (2009) are as follows. First, the model does a reasonable job of capturing the time series changes in the employment share of agriculture since 1800. If the value of $\bar{c}_a$ is held fixed throughout, the model somewhat under predicts the employment share for agriculture in the 1800s, but does fine in the post-1950 period. A small time trend in $\bar{c}_a$ over the period 1800–1950 yields a better fit over the entire period. Second, prior to 1950 the income effect is the dominant factor in accounting for the movement of employment out of agriculture, whereas the relative productivity effect is working in the opposite direction. Only in the post-1950 period do the effects of relative productivity and capital deepening play even a modest role in accounting for the change in the employment share of agriculture. They also consider various extensions to their analysis, such as incorporating trade, and they show that the results are robust to these extensions.

We want to stress three key implications of the results of Dennis and Iscan (2009). First, the fact that their model does a reasonable job of capturing the movement of labor out of agriculture over a long time period suggests that our benchmark model is sufficiently rich to capture some key features in the data. Second, the fact that a time-varying subsistence term, $\bar{c}_{at}$, improves the model's ability to account for the movement out of agriculture is notable, and suggests that a deeper theory of how income effects arise, may be warranted. Third, at least for the movement of labor out of agriculture in the United States, income effects are effectively the sole driving force behind this decline; even though the other factors play a role after 1950, this occurs when almost all of the decline in the employment share for agriculture has already happened.

It is also relevant to note some limitations of the analysis in Dennis and Iscan (2009). First, it only focuses on the movement of labor out of agriculture and does not address the issue of what forces shape the allocation of employment between manufacturing and services. Second, all of their results come from a calibration exercise, but there is little direct evidence on some of the key parameters they use for this exercise. Additionally, they connect their model to the data in a somewhat inconsistent fashion, in that they interpret their production functions as value added production functions, but when they look at consumption of agriculture they interpret it as consumption of final goods. In the next subsection, we discuss in detail why this is inconsistent. Third, they focus only on the changes in employment shares, and so do not address the issue of the discrepancy between value added shares and employment shares that we documented earlier. Nonetheless, we think that this paper makes an important contribution to the effort to identify the key economic forces behind structural transformation.

A related exercise was carried out by Buera and Kaboski (2009). Specifically, they assessed the ability of a calibrated version of our three-sector benchmark model to account for the broad patterns of structural transformation in the US from the 1800s to the present. One difficulty that they noted was the ability of the model to account for the acceleration in the nominal value added share of the service sector in the post-World War II period.

6.5.2.2 Structural Transformation in the US Since 1947

Herrendorf et al. (2009) offer a related but distinct approach to uncovering the importance of income and relative price effects in accounting for structural transformation. In contrast to Dennis and Iscan (2009), who considered the allocation of employment between agriculture and non-agriculture in the US since 1800, Herrendorf et al. (2009) consider the reallocation among consumption expenditure shares for all three sectors in the US since 1947. Specifically, starting with a stand-in household, they asked whether the utility function in (6.1) provides a good fit to the US data on expenditure shares in the post World War II period, and if so, what this implies for the values of the key parameters $\bar{c}_a, \bar{c}_s$, and ε, and the implied importance of income and relative price effects.

Although this seems to be a simple question, Herrendorf et al. (2009) argued that the question is not even properly specified. The reason for this is related to the difference between value added and final expenditure, which we have previously discussed. In particular, if one interprets the sectoral production functions as value added production functions then the arguments of the utility function necessarily represent the corresponding consumption of sectoral value added. In terms of our previous example of the purchase of a cotton shirt, this implies that the shirt is broken into three value added pieces, each of which the household values as they contribute to the three different categories of value added. Herrendorf et al. call this the value added approach. Alternatively, one may interpret the commodities in the utility function as final expenditure categories, as is typically done in household expenditure studies. The outputs of the production functions must then be viewed as final expenditure rather than value added. In terms of the purchase of a cotton shirt, the consumer simply derives utility from the shirt as a whole as it contributes to the single category of manufacturing consumption. Herrendorf et al. call this the final expenditure approach. It is important to note that there is no right or wrong in terms of these two approaches. From the perspective of preferences, these are simply two different ways of aggregating across the many characteristics that consumers value. As is true with any attempt to aggregate individual characteristics into broader groups, one can imagine examples where one approach seems preferable.

The choice of interpretation matters if the relative prices and quantities are not the same for the two different interpretations. In particular, even if the two different approaches display similar qualitative properties in terms of changes over time, differences in quantitative properties may have important implications for parameters of the utility function and the importance of income and relative price effects. Herrendorf et al. (2009) carry out the manipulations necessary to have consistent sets of data for the two approaches and they provide the following answers.

One possible outcome from this exercise is that one of the approaches provides a better fit to the data, in which case one might use this as evidence in support of one approach over the other. However, Herrendorf et al. (2009) found that for both approaches the preferences represented by (6.1) yield very good fits to the post-war US data on relative prices and expenditure shares. However, the two approaches yield very different parameter estimates for the utility functions and very different assessments of the relative importance of the effects of relative prices and income.

For the final expenditure approach, income effects are the dominant source of changes in expenditure shares, and the Stone-Geary utility function (6.28) of Kongsamut et al. (2001) provides a good fit to the data.[37] For the value added approach, it turns out that relative price effects are a much more important source of changes in expenditure shares. Moreover, the homothetic Leontief utility function $\min_{c_{at},c_{mt},c_{st}}\{\omega_a c_{at}, \omega_m c_{mt}, \omega_s c_{st}\}$, which results in $\varepsilon = \bar{c}_a = \bar{c}_s = 0$, provides a reasonable fit to the data. Interestingly, this utility function is a special case of the class of inelastic CES utility functions that Ngai and Pissarides (2007) considered.[38]

It is important to emphasize what these results mean. They are not an example of researchers obtaining different estimates for a given parameter from different data sets, suggesting that further work is needed to narrow down the set of possible values. Instead, the implication is that there are two different ways to interpret commodities in the utility function in multi-sector models. It turns out that being explicit about which interpretation is adopted is of critical importance, in that it has implications for what data is required to connect the model with the data, and as just shown, this has very important implications for implied preference parameters. Furthermore, note that the

[37] Many other papers have estimated linear expenditure systems implied by the Stone-Geary utility specification. A review of this literature is Blundell (1988).

[38] While Buera and Kaboski (2009) independently reached the conclusion that a low σ is required to match value added data, they also found that the benchmark model cannot account for the increase of the share of services in the last thirty years. Herrendorf et al. (2009) show that the reason for the different conclusions is that Buera and Kaboski (2009) assume that all investment is produced in manufacturing. This implies that they do not take into account that the investments produced in services have risen sharply since World War II.

two approaches are just two different aggregate representations of the same underlying economic data. The key message is that one cannot talk about the importance of income or relative price effects as drivers of structural transformation without specifying what representation of the data one is adopting. What shows up as income effects in one representation may manifest itself as relative price effects in the other representation. Different representations are connected via the complex input-output relationships in the economy. Herrendorf et al. (2009) show how one can construct the mapping between the two representations for a given input-output structure.

We stress two key results. First, the fact that the model is able to account for changes in expenditure shares for the US since 1947 is again support for the parsimonious model that we have adopted as our benchmark. Second, it highlights that empirical researchers working with multi-sector models must take care to be explicit about how commodities in utility functions are to be interpreted. Different interpretations have dramatically different implications for how the models are to be connected with the data and what the implied parameters of the utility function.

One of the limitations of this study is that it only focuses on the post-1947 period for the US, and this is a period in which the US has already experienced much of the reallocation out of agriculture. While it is of interest to extend this type of analysis to longer time periods and different countries, a key issue is data availability.[39]

6.6. EXTENSIONS OF THE BENCHMARK MODEL

In this section we discuss relaxing three features present in the analysis of the benchmark model. The first is the assumption that there is no international trade (closed economy). The second is the assumption that there is no cost of moving labor across sectors (perfect labor mobility). The third is the assumption that there are no costs of moving goods across sectors (zero transportation costs).

6.6.1 International Trade

Thus far, our theoretical analysis has taken place under the assumption of a closed economy. A key implication of being a closed economy is that the production of each of the four sectors must equal the corresponding household choices (either of investment or of one of the three consumption goods). The equality between sectoral productions and consumption/investment played a key role in generating the results concerning structural transformation that we obtained in the benchmark model. For example, in the model

[39] This is relevant for the analysis of Buera and Kaboski (2009). They carry out a calibration exercise for the US over a longer time period, but need to use different sources for relative prices in the pre-1947 period. Given that prices for value added consumption and final consumption are quite different in the post-1947 period and have very different implications for preference parameters, an issue arises with how to interpret results that use a mixture of prices.

of Ngai and Pissarides (2007), we saw that labor moved out of the consumption sector that had the highest productivity growth because of the household's desire to maintain the composition of its consumption allocation (inelastic demand). In the model of Kongsamut et al. (2001), technological progress was uniform across sectors, but labor moved out of agriculture because of the household's desire to change the composition of its consumption allocation toward manufactured goods and services (differences in income elasticities).

In this subsection we discuss the extent to which openness changes the results about structural transformation. We begin with the simple observation that the competitive equilibrium of a model in which all commodities are tradeable without costs will have a complete separation between the decisions of firms and households. This observation implies that in an open-economy version of our benchmark model without trade costs the production measures of structural transformation (i.e. employment and value added shares) would generically follow a different pattern than the consumption expenditure share. This is relevant because, as we have documented in Section 6.2, there is a discrepancy between production and consumption shares in some instances, most notably for the share of manufacturing in Korea.

Matsuyama (2009) was the first to analytically work out the idea of the previous paragraph for a simple two-country model. He abstracts from capital and considers a Stone-Geary utility function over the three consumption goods: food, manufactured goods, and services. He assumes that agricultural goods are an endowment whereas manufactured goods and services are produced with technologies that are linear in labor, and that agricultural and manufactured goods can be traded with the rest of the world at zero trade costs whereas services cannot be traded. Matsuyama shows two results for this simple model. First, if there is technological progress in manufacturing then the total manufacturing labor of both countries declines. Second, if one of the two countries experiences stronger technological progress in manufacturing than the other, then manufacturing labor in the first country may initially increase while manufacturing labor in the second country decreases unambiguously. Eventually, when technological progress in the manufacturing sector has been sufficiently strong, the share of manufacturing labor in the first country will decrease also. These results suggest that a hump-shaped relationship may occur in the country which experiences the stronger technological progress in manufacturing.

Yi and Zhang (2010) generalize the idea of Matsuyama to a two-country version of our benchmark model of structural transformation, in which all goods are produced with labor only. The assumption that agricultural and manufactured goods are tradeable without costs would then lead to the counterfactual implication that each country specializes in either agriculture or manufacturing. They therefore assume that each of the three sectors is the aggregate of a continuum of goods as in Eaton and Kortum (2002). Yi and Zhang (2010) simulate their model under the assumption that one country has higher productivity growth in manufacturing than the other country. They provide examples for which the country with the higher productivity growth in manufacturing experiences a hump shape in the shares of manufacturing employment and value added while the other country experiences a downward-sloping shape in the shares of manufacturing labor and value added.

From the empirical perspective it is of interest to ask whether there is evidence for the effects of openness on structural transformation, besides the hump shape of manufacturing employment and value added. One clear prediction of the models of Matsuyama (2009), and Yi and Zhang (2010) is that the labor shares of sectors that produce tradeable goods should differ across countries that have different sectoral productivities. In Section 6.2 we noted that there was some evidence of dispersion in sectoral labor shares across countries in the European Union and Japan, with Germany and Japan having unusually large share of manufacturing hours worked and Korea having an unusually large share of real manufactured value added. Betts et al. (2011), Sposi (2011), and Teignier (2012) study the role of international trade in Korea's industrialization. They find that international trade played a crucial role in the rapid rise in the manufacturing value added and employment shares. Teignier (2012) finds in addition that international trade could have played a much larger role if South Korea had not introduced agricultural protection policies.[40] While such a story may be consistent with various accounts regarding the importance of trade in the development of South Korea, it is hard to reconcile with the patterns we found in Section 6.2. Specifically, we found that South Korea did not display any distinctive behavior for the labor allocations.

We conclude that the effects of openness on structural transformation show up in a discrepancy between production and consumption in sectors that trade with the rest of the world. In the past, this applied to manufacturing, and to a lesser extent to agriculture. In recent years, however, there has been an increasing trend toward trade in services. An open question moving forward concerns the extent to which increased trade in services will influence the nature of structural transformation. For example, will increased trade in services hasten the movement of resources out of manufacturing in a country like the US which has relatively high productivity in many service industries, and is therefore thought to have a comparative advantage in services?

6.6.2 Labor Mobility

Our benchmark model assumed that labor was homogeneous and could be allocated across sectors without any labor mobility costs. There are several interesting issues that arise when there are labor mobility costs. In this subsection we discuss the most relevant ones.

[40] Swiecki (2013) builds a multi-country model of structural transformation in which sectoral allocations may be affected by country-specific distortions. He shows how this model influences our estimates of the gains from trade and the incentives for countries to adopt protectionist policies.

We begin with the paper by Lee and Wolpin (2006) about the large reallocation of labor from manufacturing to services in the United States over the period from 1968 to 2000. The goals of this paper are to measure the costs associated with sectoral labor reallocation and to assess the relative importance of labor demand and supply factors for sectoral labor reallocation, where labor demand factors are defined as changes in sectoral productivity and relative prices and labor supply factors are defined as changes in demographics, fertility, and educational attainment. To reach these goals, they develop a framework with a detailed labor market. To begin with, there are three occupational choices in each sector: blue collar, white collar, and pink collar (i.e. secretarial, clerical, etc.). Moreover, workers differ in their educational attainment and they can accumulate sector-specific and occupation-specific human capital while working. Lastly, there are various types of technological changes and the production functions have a constant elasticity of substitution between capital and labor.

Lee and Wolpin (2006) estimate their model using micro data. Their main findings are as follows: First, labor demand factors are the key driving forces behind the reallocation of labor across sectors. In contrast, labor supply factors do not play much of a role. This finding is consistent with the emphasis that our benchmark model puts on technological factors. Second, and in contrast to our benchmark model, the mobility costs associated with moving across sectors are large; for example, the monetary cost of changing sectors can be as large as 75% of annual earnings. Moreover, changing occupations within a sector is significantly less costly than changing sectors while maintaining the same occupation.

Lee and Wolpin (2006) carry out several counterfactuals regarding how changes in mobility costs would have affected the evolution of labor market outcomes. Interestingly, they find that if mobility costs had been zero, aggregate productivity would have been higher and the labor market histories of individual workers would have been different, but the evolution of sectoral employment shares and value added shares would not have changed much. The economics behind this result is that with lower mobility costs workers can better allocate their time to the sector in which their idiosyncratic productivity is highest. This raises aggregate productivity and changes the labor market histories of individual workers. However, since it leads to flows of workers in both directions, the effect on relative sectoral employment is relatively small. This result suggests that abstracting from mobility costs in our benchmark model does not have large quantitative effects on the sectoral employment allocation.

Lee and Wolpin (2006) also ask what would have happened if sectoral labor mobility had been more costly. They find that while there would have been little effect on trend changes in employment shares, the level of the employment share of services would have shifted upward. This result runs counter to the intuition that increased mobility costs will decrease the flow of workers into the expanding service sector. To understand this, it is important to realize that this intuition is based on how mobility costs affect the response to an unanticipated shock. In contrast, what matters for Lee and Wolpin's exercise are the choices that forward-looking new entrants make in the face of the trend that the service sector is becoming more attractive in comparison to the goods sector. If we increase the size of mobility costs, then more entrants move directly into the service sector, instead of first going to the manufacturing sector and later switching to the service sector.

There is more evidence that the role of new entrants is crucial for the labor reallocation across sectors in the context of structural transformation. For example, Kim and Topel (1995) show that during Korea's rapid industrialization almost all of the changes in the sectoral employment shares of agriculture and manufacturing resulted from changes in the behavior of new entrants. As a result, the large decrease in the agricultural employment share and the large increase in the manufacturing employment share were accomplished with little reallocation of existing workers.[41] To the extent that new entrants are an important source of labor market flexibility one might conjecture that economies with different rates of growth in the labor force might experience different patterns of structural transformation. However, we are not aware of existing evidence that supports this conjecture.

While some mobility costs might reflect technological factors, it is also possible that policies, regulations, and institutional factors lead to the barriers to labor mobility. Examples include implicit or explicit firing costs levied on employers, subsidies to establishments in declining industries, entry barriers that make it costly for firms to start up new establishments, generous unemployment benefits or early retirement schemes that are offered to displaced workers, and direct restrictions on the mobility of workers.[42] There are many studies of these types of factors, but most of them make no reference to the process of structural transformation. The reason for this is that most job creation and destruction occurs within, rather than across, narrow industrial classifications, and so the main effects come from the reallocation of resources across establishments when jobs are created and destructed.

Three exceptions that study the effects of labor mobility costs in the context of structural transformation are Nickell et al. (2002), Messina (2006), and Hayashi and Prescott (2008). Nickell et al. (2002) examine the correlations between sectoral composition and various policy and institutional factors in a panel data set panel of 14 OECD countries and 5 one-digit industries during the period 1975–94. One of their findings is that countries with more stringent employment protection policies have larger industrial sectors, suggesting that employment protection policies might impede the reallocation of employment from manufacturing into services. Messina (2006) considers the role of entry barriers. One distinguishing feature of structural transformation in Europe is that, conditional on aggregate productivity (i.e. output/hour), Europe has a much lower

[41] Matsuyama (1992b) and Rogerson (2006) both present models of sectoral reallocation that have this property.

[42] China is a clear example of an economy that has direct restrictions on the mobility of workers. Dekle and Vandenbroucke (2012) show that these restrictions slowed the Chinese movement out of agriculture.

employment share for services than do other rich countries.[43] Messina argues that this is the result of higher entry barriers in Europe, including such factors as direct costs associated with licensing and indirect costs associated with zoning restrictions or regulations that restrict shopping hours, etc. Because the reallocation of workers into services requires additional entry of establishments into the service sector, these barriers retard the movement of economic activity into the service sector. Hayashi and Prescott (2008) study the movement of labor out of agriculture in Japan before World War II. They argue that the pre-war patriarchy that forced the son-designated-as-heir to stay in agriculture, effectively amounted to a barrier to the movement of labor out of agriculture. Using a standard neoclassical two-sector growth model, they show that the barrier-induced sectoral distortion and the implied lack of capital accumulation account well for the depressed output level of Japan's pre-war economy.

Although Lee and Wolpin (2006) incorporated a range of factors that make mobility costly for individual workers, their model still shares the feature of our benchmark model that all labor reallocation was voluntary from the perspective of the worker. A large literature has documented the large earnings losses that older workers face when they are displaced (see, for example, Jacobson et al. (1993)). To many policymakers and commentators, the reallocation of labor from manufacturing to services that is part of the process of structural transformation is synonymous with the displacement of older, high-tenure workers in the manufacturing sector and either unemployment or large losses in earnings. While the connection may seem clear-cut, direct evidence on this point is much less clear-cut. As noted by many authors, most job creation and destruction occurs within narrow industry classifications, and so is not directly related to the reallocation of activity across broad sectors.[44]

6.6.3 Goods Mobility

If openness matters for the process of structural transformation in some settings then it follows that the cost of moving goods may influence structural transformation as well through their effect on trade. More interesting is the possibility that transport costs might influence structural transformation in a closed economy setting. One simple idea in this literature stems from noting that while agriculture is predominantly rural, much of the activity outside of agriculture takes place in cities. It follows that food consumed by non-agricultural workers needs to be transported from rural to urban areas. If this is the case, then high costs of moving food from rural areas could exert a negative influence on the movement of labor out of agriculture.

[43] This was not apparent in Section 6.2 since we plotted the service share of hours worked versus per capita income rather than output per hour.

[44] See, for example, Davis and Haltiwanger (1992).

Herrendorf et al. (2012) study this idea in the context of the transport revolution in the US before the civil war, during which the construction of railroads reduced dramatically the transportation costs to the most fertile farm land in the Midwest. They build a model with two regions (Midwest and Northeast) and three sectors (agriculture, manufacturing, and services). Consistent with our benchmark model, their model also allows for both income effects via a subsistence term in the utility from agriculture, and productivity effects in terms of the factors that determine the allocation of labor to agriculture. They show that the reduction in transportation costs between the two regions leads to the settlement of the most fertile farm land in the Midwest, which is followed by a reduction in the agricultural labor force.

Adamopoulos (2011), and Gollin and Rogerson (2010) study this idea further in the context of a static model with agriculture and non-agriculture and different locations. Adamopoulos shows that transportation costs between locations can exert an important influence on the allocation of resources across locations and between agriculture and non-agriculture. Gollin and Rogerson carry out some numerical exercises to suggest that there is a strong interaction between increases in productivity and reductions in transportation costs in terms of their impact on labor moving out of agriculture.

6.7. APPLICATIONS OF STRUCTURAL TRANSFORMATION

In this section, we return to the question we posed in the introduction to this chapter: Does incorporating structural transformation into the standard growth model deliver new insights? In other words, is there a substantive payoff to working with versions of the growth model that account for structural transformation? We discuss several issues where changes in the sectoral composition of the economy matter have been shown to matter. We conclude that explicit modeling of structural transformation offers important additional insights into these cases.

6.7.1 Structural Transformation and Economic Development

Caselli (2005), and Restuccia et al. (2008) argue that the proximate cause of much of the large differences in living standards across countries is attributable to two simple facts: (1) developing countries are much less productive in agriculture relative to developed countries, and (2) developing countries devote much more of their labor to agriculture than do developed countries. These two facts suggest that in order to understand why developing countries are so poor it is of first-order importance to understand the forces that shape the allocation of resources between agriculture and the other sectors. A version of the growth model extended to incorporate structural transformation is the natural framework to be used in this context.

Work by Gollin et al. (2002, 2006) illustrates how low agricultural productivity can be the source of large cross-country differences in aggregate productivity. For ease of exposition we focus on the simpler presentation in the 2002 paper, which uses a two-sector version of our benchmark model, with the two sectors being agriculture and non-agriculture. They assume that the population is constant and normalize it to one. Preferences are such that there is a subsistence level $\bar{c}_a$ of agricultural consumption at which individuals are also satiated. The non-agricultural production function is essentially a Cobb-Douglas production function in capital and labor. In contrast, there are two agricultural production functions: a traditional and a modern one.[45] Both agricultural production functions are linear in labor, though the analysis would be unaffected by assuming a fixed quantity of land and decreasing returns to scale in labor. The traditional production is assumed to be the same across countries and to be sufficiently productive to exactly meet subsistence agricultural needs when all labor is allocated to it. The modern production function has a country-specific TFP parameter and it is the only production function that is subject to technological progress.

In this model, only the agricultural technology with the larger productivity will be used in equilibrium. Initially, this is the traditional technology. Since the modern technology is subject to technological progress, at some point the modern technology will replace the traditional technology as the only technology that will be used. The somewhat extreme structure of the model then yields a very simple solution method for determining the equilibrium. Total food production must be $\bar{c}_a$. As long as the traditional technology is used, this means that all labor will be in agriculture. When the modern technology starts to dominate the traditional technology, labor will start to flow from agriculture to non-agriculture. With the time series for labor allocations determined, the remainder of the model becomes a standard growth model with an exogenously given process for labor. The growth rate of labor in the non-agricultural sector is completely determined by the exogenous growth rate of labor productivity in the modern agricultural sector. Since all countries have the same output of agriculture, cross-country differences in aggregate output are entirely driven by differences in non-agricultural output.

Several implications follow. First, countries that use the modern technology in agriculture but have low productivity in it will have to devote more labor to agriculture. This leads to less labor in non-agriculture and hence to less aggregate output. Given the observed differences in the share of labor that is allocated to agriculture, Gollin et al. (2002) show that this mechanism can account for a large part of the cross-country differences in aggregate output. This is interesting because in their model the only difference across countries is the level of productivity of agriculture.

Second, assuming that productivity growth rates are constant over time, the model necessarily implies that transition dynamics will be long-lived, thereby addressing a point emphasized by King and Rebelo (1993), that in a standard, one-sector growth model transition to the steady state capital level is rapid.[46] This point does not carry over to

[45] Hansen and Prescott (2002) use a similar assumption but at the aggregate level.

[46] Chang and Hornstein (2011) make a related point about Korea. They show that two modifications of the one-sector growth model are essential to account for the long-lived transition dynamics since 1960,

the two-sector model because labor allocated to the non-agricultural sector only slowly converges to its asymptotic level. Third, the model implies that in a closed economy, setting advances in agricultural productivity are a precondition for growth. This view was a central argument of Schultz (1953), and figured prominently in later contributions by Johnston and Mellor (1961), Johnston and Kilby (1975), and Timmer (1988), among others. More recently, it has taken a central state in the writing of non-economists such as Diamond (1997).[47]

Laitner (2000) considers a similar framework as Gollin et al. (2002) but focuses on a different issue. He notes that in the time series data there is evidence of an increase in savings rates early in the industrialization process. Whereas some have argued that the increase in savings rate is the driving force behind the industrialization process, Laitner shows that, in a model of structural transformation, this apparent increase in savings rate is simply an artifact of how NIPA measures saving. Early in the development process most labor is employed in agriculture, and so most savings take the form of realized capital gains in the value of land, which is not recorded as savings by the NIPA. As labor moves out of agriculture and agriculture becomes a smaller part of aggregate output, this issue becomes less important quantitatively. Laitner argues that viewed from the perspective of his model of structural transformation, one should not attach any significance to the apparent increase in savings rates that occur in the early stages of development.

6.7.2 Structural Transformation and Regional Income Convergence

One of the dramatic secular changes in the US economy over the post World War II period is the convergence of incomes across regions; see, for example, Barro and Sala-i-Martin (1992). In the context of the standard one-sector neoclassical growth model, this convergence in incomes would be attributed to changes in either regional TFP or regional factor accumulation. Caselli and Coleman (2001) show that a model of structural transformation provides a richer understanding of the economic forces at work. The motivation for their analysis is provided by the fact that the convergence in regional incomes between the north and the south of the United States coincided with a dramatic narrowing of regional differences in the employment share in agriculture. They use a model that differs from our benchmark model along several dimensions. First, they consider a two-sector version of the model, with the two sectors being agriculture and non-agriculture. Second, they consider a two-region version of the model, where each region has the same structure as our model and there is free mobility of goods across regions. They assume that the technologies are such that the north has a comparative advantage in manufacturing and the south has a comparative advantage in agriculture. They focus on the special case in

during which Korea continued to accumulate capital. The first modification is to distinguish between agriculture and non-agriculture and to take into account that Korean agriculture used relatively little physical capital. The second modification is to model that the relative price of capital remained high during most of the transition dynamics.

[47] See Tiffin and Irz (2006) for a recent empirical assessment.

which the technologies in manufacturing are the same in both regions and the south has higher in TFP agriculture (for simplicity, they assume that the north has zero TFP in agriculture). Third, they assume that there are mobility costs in terms of sectoral reallocation of labor. Specifically, all workers begin in the agricultural sector, and they must pay a cost if they are to move to the non-agricultural sector. They interpret this mobility cost as the cost of acquiring skills that are needed in the non-agricultural sector and argue that it is necessary if one is to account for the secular changes in labor allocations and relative wages.

The basic economics of their analysis is the following: When the United States was relatively poor, more of its workers were engaged in agriculture, due to non-homothetic preferences which imply a large share for agricultural expenditures at low levels of income. Because the south had a comparative advantage in agriculture, it was doing relatively more of it. Because of mobility costs, wages were higher in non-agriculture. Putting these features together, incomes were lower in the south. Over time, production technology in non-agriculture advanced, leading to a decline in the share of workers in agriculture. They also posit that in addition, mobility costs decreased, therefore leading to convergence between agricultural and non-agricultural wages.[48]

6.7.3 Structural Transformation and Aggregate Productivity Trends

Our model of structural transformation allows for the possibility that different sectors have different levels, as well as growth rates of labor productivity. Herrendorf and Valentinyi (2012) provide evidence from the 1996 Benchmark Study of the Penn World Tables on sectoral TFP differences across countries. They find that there are large sectoral TFP differences relative to the United States not only in agriculture, but also in manufacturing, and that the sectoral TFP differences in these two sectors are much larger than in the service sector.[49] Aggregate labor productivity may then be affected by the sectoral composition of the economy. In particular, to the extent that different countries are at different stages of the process of structural transformation, sectoral reallocation associated with structural transformation could generate significant changes in aggregate productivity growth [Echevarria (1997)]. In principle, episodes of acceleration or slowdown in aggregate productivity growth may occur even if in each country sectoral productivities are growing at constant rates.

In a recent paper, Duarte and Restuccia (2010) have investigated the importance of these effects in a sample of 29 countries for the period of 1956–2004. They employed a somewhat simplified version of our benchmark model in which labor is the only factor

[48] In related work, Hnatkovska and Lahiri (2012) show that structural transformation importantly contributed to the narrowing of the urban-rural wage gap in India during 1983–2010.

[49] The result of Karádi and Koren (2012) suggests that cross-country productivity differences in services might be larger if the development accounting framework allows for producers to tradeoff transportation costs to city centers against land rents in the city center.

of production (and production functions are linear in labor). They assumed that each sector's labor productivity grows at a constant rate, but that level and growth rates differ across economies as dictated by the data.

The preference structure of Duarte and Restuccia (2010) assumes a period utility function which is a two-good version of (6.28):

$$C_t = \omega \log\left(c_{at} - \bar{c}_a\right) + \omega_n \log\left(c_{nt}\right).$$

c_{nt} stands for non-agricultural consumption and it is a CES aggregator of manufactured goods and services. Preference parameters are calibrated so as to match the behavior of the US economy and are assumed to be the same across countries. The initial productivity levels of all countries relative to the US are inferred from the model by requiring that it match the observed employment shares in the initial period. Inputting the sectoral productivity growth rates from the data, Duarte and Restuccia (2010) then simulate the model and compute the implied series for aggregate labor productivity.

Even though their model assumes constant productivity growth rates at the sectoral level of each country, it generates large movements in relative aggregate productivity across countries over time. Key to this finding is that differences in the levels and growth rates of labor productivity between rich and poor countries are larger in agriculture and services than in manufacturing. This implies that during the process of structural transformation, the reallocation of labor from agriculture to manufacturing led to a catch up of aggregate productivity relative to the US, and the reallocation from manufacturing to services leads to a falling behind of aggregate productivity relative to the US.

In related research, Bah and Brada (2009) study the countries from Central Europe which have recently entered the European Union. The point of departure of their analysis is the stylized fact that central planning during communist times resulted in over-agrarianism and over-industrialization, and the neglect of the service sector in these countries. Bah and Brada document that even today employment in the service sector is considerably smaller in Central Europe than in the core countries of the European Union. Moreover, they find that in all of these countries the service sector has lower TFP than the manufacturing sector. This implies that structural transformation into the service sector will lead to losses in GDP per capita, unless reforms are implemented that make the service sectors more productive.

6.7.4 Structural Transformation and Hours Worked

Following Prescott (2004), there is a sizeable literature that seeks to understand the large differences in hours worked that have emerged over time between the US and countries in continental Europe. In order to be able to compare hours worked across countries of different size, Prescott divided total hours worked by the working-age population. Prescott used the standard one-sector growth model to demonstrate that changes in labor taxes could account for much of the emerging difference.

Rogerson (2008) argued that a model of structural transformation provides additional insights into the evolution of hours. In particular, he compared the evolution of hours worked per working-age person in the US to those in an aggregate of five continental European economies (Belgium, France, Germany, Italy, Netherlands) since 1956. Whereas hours worked were about 5% higher in Europe in 1956, by 2003 they were more than 30% lower. Looking at the sectoral evolution of hours worked reveals an interesting pattern. During the period in which hours worked in these European economies fell by more than 35% relative to the US, one observes that the relative level of hours worked in the goods sector in Europe fell dramatically, whereas the relative level of hours worked in services remained relatively flat.[50] One might be tempted to conclude that the key to understanding the relative decline in hours worked in Europe lies in understanding the relative decline in hours worked in the goods sector. However, when one views the sectoral evolution of hours worked in the context of structural transformation, one is led to exactly the opposite conclusion. Specifically, in 1956, Europe was considerably behind the US in terms of development, and consistent with our earlier empirical analysis, had a larger share of hours in the goods sector and a smaller share in the service sector than the United States. By 2003, Europe has basically caught up with the United States in terms of productivity. Holding all else constant, one would expect that the sectoral hours worked distribution in Europe in 2003 would look similar to that in the United States. That is, the process of structural transformation leads us to expect that while hours in the goods sector in Europe should have decreased relative to the US, hours in the service sector in Europe should in fact have increased. Put somewhat differently, the issue of understanding why hours worked are so much lower in Europe reduces to the issue of understanding why the European service sector has failed to grow like its counterpart in the US. In fact, this dynamic was apparent in the hours plots in Figure 6.2.

In addition to simplifying the analysis by aggregating agriculture and manufacturing to one category and by abstracting from capital, Rogerson's model differs from our benchmark model along two key dimensions; he adds a labor supply decision and he allows for home production, which he assumes to be substitutable with the output of the service sector. His model combines both income and price effects to generate structural transformation. Taking changes in productivity and labor taxes as given, he calibrates the preference parameters so as to match the changes in the US economy between 1956 and 2003, including the change in time devoted to home production.[51] He then feeds in European values for productivity and taxes in both 1956 and 2003 and examines the ability of the model to account for aggregate and sectoral observations in Europe in 1956 and 2003. Overall, Rogerson finds that the model accounts well for the sectoral European labor allocations.

[50] Hours worked in a sector again is defined as total hours worked divided by the working-age population.

[51] See Aguiar and Hurst (2007), and Ramey and Francis (2009) for evidence on the decline of home production time in the US.

Rogerson assumes that the utility function is non-homothetic in that it has a subsistence level of goods consumption. This turns out to be important for understanding relative hours worked in Europe in the initial year of his study, 1956. At that time, Europe already had higher tax rates than the US, yet they had higher hours of work. The non-homotheticity acts like a negative income effect, and this effect is larger the lower is aggregate productivity. Given that Europe lagged the US in aggregate productivity in 1956, this effect serves to increase hours in Europe relative to the US. Additionally, because the model generates structural transformation, Europe devoted more labor to goods production than the US in 1956. Because there are fewer non-market substitutes for goods, this effect also serves to increase the amount of time devoted to market work.

In related work, Ngai and Pissarides (2008) add a home production sector to their earlier model of structural transformation that we have discussed above, Ngai and Pissarides (2007). They showed that over time the model with home production generates a shallow U-shaped curve for hours devoted to market work, and that it leads to the marketization of home production, i.e. the movement of time out of home production and into market production of services. Both of these patterns are found in the US data. The initial decrease in market work is associated with the movement of activity into services, which have better home-produced substitutes. But as time advances, a higher rate of growth in the productivity of market produced services relative to home-produced services leads to the movement of activity out of the home sector and into the market sector, which results in the increase in market hours.

Another dramatic trend in labor market outcomes has been the rise of female labor force participation. Several authors have argued that the process of structural transformation is an important factor in accounting for this change. The basic idea is that jobs in the goods sector (i.e. agriculture and manufacturing) and the service sector tend to have different weights on various dimensions of labor input. In particular, the goods sector places more emphasis on brawn while the service sector places more emphasis on brains. If men and women have different relative endowments of these two factors, then the movement of activity from one sector to the other could plausibly affect the desire of women to seek employment in the market sector. Fuchs (1968) noted this explanation for the rise of female labor force participation.[52]

Rendall (2010) builds a two-sector model in which she can quantitatively evaluate the difference between men and women and argues that structural transformation is an important quantitative factor in accounting for the rise of female labor force participation. In related work, Akbulut (2011) also argues that the rise of the service sector has been an important factor in accounting for the rise of female labor force participation in

[52] Galor and Weil (1996) also note the changing demands for brain and brawn, though not in the specific context of structural transformation. See also the papers by Goldin (1995, 2006) for additional analysis of the evolution of female labor force participation patterns.

the US, but the key reallocation in her model is the movement of labor out of home-produced services and into market produced services in response to a more rapid rate of technological progress in market services relative to home produced services. Finally, Ngai and Petrongolo (2013) argues that structural transformation can not only account for the decline in the gender gap in labor force participation, but also for the decline in gender wage gap. In particular, they point out that if structural transformation is driven by differential productivity growth, and women have comparative advantage in services, then the gender wage gap will decline as hours worked in services increase relative to goods.

Olivetti (2012) provides evidence from a large sample of developed and developing countries that connects the U-shaped profile for female labor force participation to structural transformation, extending the earlier work by Goldin (1995). Specifically, she finds that as countries develop, the share of women who work in agriculture relative to all working women decreases faster than the share of men who work in agriculture relative to all working men; the share of women who work in services increases faster than the share of men; and the share of women who work in manufacturing remains flatter than the share of men.

6.7.5 Structural Transformation and Business Cycles

There are many different ways in which theories of structural transformation and business cycles might overlap. One idea which frequently recurs is that some business cycles are the result of periods of greater reallocation of economic activity across sectors. To the extent that this reallocation of activity occurs at the broad sectoral level emphasized by models of structural transformation, structural transformation and business cycles could be intimately related.

Using the search model of Lucas and Prescott (1974) as a reference point, Lilien (1982) argued that if it takes time for labor to move from one sector to another, then periods of above average reallocation will also be periods of above average unemployment. He then argued that business cycles in the post-World War II US were characterized as periods of above average reallocation of labor among two-digit sectors, as measured by the variance in employment growth rates at the two-sector level. However, subsequent work by Abraham and Katz (1986) argued that Lilien's statistical finding about changes in the variance of sectoral growth rates could simply be due to the fact that sectors vary in their response to aggregate shocks, and that data on vacancies supported this latter explanation over the sectoral shifts explanation.

The idea of Lilien (1982) has experienced a recent resurgence in popularity in the face of the current recession, with various economists suggesting that mismatch is an important element of the current high level of unemployment, and that the decline of broad sectors, such as manufacturing and construction, is an important element of this mismatch. However, despite its popularity, recent empirical research by Sahin et al. (2011), and Herz and van Rens (2011) finds little evidence for this explanation.

We note that even if reallocation were concentrated during recessions, it would not follow that recessions are caused by the reallocation. Rather, it may be that recessions are caused by a second factor, and that the decisions that lead to reallocation are made in such a way that reallocation coincides with the recession. That is, for example, it may be that steel mills go out of business permanently during recessions, but this may simply reflect that the optimal timing of exit for a steel mill is during a downturn in economic activity. Rogerson (1991) argued that movement out of agriculture in the US has been concentrated during upturns in economic activity, whereas the movement of workers out of manufacturing has been concentrated during downturns.

Even if structural transformation is not the cause of business cycles, it may still exert an influence on business cycles. For example, to the extent that value added varies in volatility across sectors, the sectoral composition of aggregate output is a potentially important determinant of business cycle fluctuations. In what follows, we mention two examples of this idea.

The first example is Da Rocha and Restuccia (2006), who disaggregate the economy into agriculture and non-agriculture and document that indeed there are important differences between the two sectors. In particular, they find that the agricultural sector is more volatile than the rest of the economy, is not correlated with the rest of the economy, and has counter-cyclical employment. They show that this implies that countries with a larger agricultural sector have more volatile aggregate output and less volatile employment. Moreover, it implies that as structural transformations out of agriculture occur, business cycle properties across countries converge.

The second example of how the sectoral composition matters is due to Carvalho and Gabaix (2013), and Moro (2012). They disaggregate the economy into services and manufacturing, largely ignoring agriculture. They document that the volatility of services is lower than in manufacturing. Moro (2012) argues that the reason for this is that the share of intermediate inputs is larger in manufacturing than in services. Irrespective of why the volatilities differ between the two sectors, the implication is that the volatility of aggregate output declines as the share of services increases along the path of structural transformation. Carvalho and Gabaix (2013) find that this accounts for most of the great moderation and its recent undoing. In particular, the great moderation is due to a decreasing share of manufacturing between 1975 and 1985, and its recent undoing in the form of rising aggregate volatility is due to the increase of the size of the financial sector.

6.7.6 Structural Transformation and Wage Inequality

One of the dramatic secular changes in the US economy over the last 50 years has been the marked increase in wage inequality that is associated with the return to skill. In a recent paper, Buera and Kaboski (2012a) argue that this rising return to skill is intimately connected to the structural transformation of economic activity toward services. They document in time series data the same threshold behavior of value added in services that we have found above, that is, there is a threshold for per capita income at which one observes an acceleration in the increase in the value added share for services. Interestingly, at that threshold there is also an increase in the fraction of the workforce that becomes skilled and of the skill premium. In the context of the US they also document that the entire rise in the service sector's share in value added in the last fifty years is accounted for by growth in sub-sectors that have higher than average shares of skilled labor. They go on to build a model that links these patterns as the outcome of structural transformation that is driven by neutral productivity growth.

We previously described some general features of the framework that they use. Relative to our earlier discussion, the key modification in this paper is that there are two types of labor: skilled and unskilled. Skilled labor is specialized to a particular service, is costly to acquire, and is subject to an increasing cost curve. To capture the fact that home production is necessarily less specialized, they assume that skilled labor is equivalent to unskilled labor in home production. Services differ in complexity, where complexity captures both the amount of labor that is required to produce them and the relative productivity advantage of skilled labor in producing the service.

As the economy develops it produces services that are increasingly complex, thereby creating additional incentives for both market production of wants and skill accumulation. Because there is an upward-sloping supply curve for skilled workers, the skill premium is also increasing. The structure of their model is such that the relative advantage of skilled labor in producing more complex services only emerges beyond a critical threshold level of complexity, so that these patterns also emerge beyond a threshold. A key fact that this model is able to account for, that our benchmark model cannot, is that this model predicts that the share of services in nominal value added is flat below some threshold.

An important implication of this work is that adding the different roles of human capital in various activities is an important ingredient in understanding some key features of structural transformation.

6.7.7 Structural Transformation and Greenhouse Gas Emissions

For several decades, greenhouse gas emissions and climate change have been at the forefront of the environmental policy debate. Grossman and Krueger (1995) documented that there is a hump-shaped relationship between the level of development and the level of greenhouse-gas emissions—as economies develop, emissions first increase, but then reach a maximum and subsequently decrease. This hump-shaped relationship is called the Environmental Kuznets Curve. Stefanski (2013) documented that not only emission levels but also emission intensities exhibit an Environmental Kuznets Curve and that emission intensities peak before emission levels.[53]

Structural transformation is relevant in the context of the Environmental Kuznets Curve because emission intensities differ across sectors.[54] Stefanski (2013) presented a simple model of structural transformation to account for the two facts described above. His model has two sectors—agriculture and non-agriculture—with exogenous labor-augmenting technological progress that is constant and even across the two sectors, and non-homothetic preferences for the agricultural and the non-agricultural good. He made three additional assumptions that are critical in the current context: at any level of output pollution intensity is higher in the non-agricultural sector than in the agricultural sector; there is constant exogenous technological progress in pollution abatement; technological progress in pollution abatement is faster than labor-augmenting technological progress in the production of output. To simplify the model, Stefanski (2013) considered the extreme case in which the agricultural sector does not emit any pollutants. He showed that under these assumptions both the level and the intensity of pollution rise during the early stages of development as resources are reallocated from the less polluting agricultural sector to the more polluting non-agricultural sector. As the economy continues to develop, the higher rate of technological progress in pollution abatement in the non-agricultural sector eventually dominates, leading first to a decline in aggregate pollution intensity and then later to a decline in the pollution level. While the analysis of Stefanski (2013) focused on a two-sector model, we note that in a three-sector model in which the service sector has a lower level of pollution intensity than the manufacturing sector, these effects would presumably be strengthened, since the later stages of development would feature an additional force leading to declines in both the level and intensity of emissions.

6.8. CONCLUSION

Our goal in this chapter has been to summarize the basic facts about structural transformation, and to present simple versions of the growth model that serve as the benchmark models being used to organize our thinking about these facts. Much of the early literature has focused on trying to identify multi-sector versions of the growth model that can generate structural transformation while simultaneously generating balanced growth. While the search for specifications that can simultaneously yield structural transformation at the sectoral level and balanced growth have proven to be useful in organizing research, we believe that focusing on frameworks that yield exact balanced growth is probably overly restrictive. The literature should instead focus on building models that can quantitatively account for the properties of structural transformation and in the

[53] Emission intensity is defined as emissions per real GDP.

[54] Emission intensity by sector is defined as sectoral emission per real sectoral value added.

process assess the importance of various economic mechanisms. We use this concluding section to highlight what we view as important priorities for future research in this area.

While we have a substantial amount of data regarding the process of structural transformation in today's advanced economies, it would be good to know more about the nature of structural transformation in today's less developed economies. To what extent are they following different paths from today's developed economies? And if so, what are the factors that give rise to these differences?

Two economies of particular current interest in this regard are China and India, both because of their size and because they have been experiencing very rapid growth. What role does structural transformation play in these countries' growth? Dekle and Vandenbroucke (2012) studied structural transformation in China during 1978–2003. They found that differential sectoral productivity growth and the reduction of the relative size of the Chinese government caused most of the structural transformation, but that mobility frictions (like the *hukou* system) slowed the movement out of agriculture. Rubina (2012) has studied structural transformation in India during 1980–2005. Contrary to the patterns that we have documented above, she has found that TFP growth was fastest in services. Moreover, she has found that a three-sector model can account for changes in sectoral value added but not in employment shares.

The growth miracle episode in South Korea has also attracted recent attention, specifically as it relates to the issue of openness, structural transformation, and growth. Betts et al. (2011), Sposi (2011), and Teignier (2012) have studied structural transformation in South Korea during its growth miracle. They argue that international trade accelerated the transition out of agriculture into industry and services. Teignier (2012) argues in addition that international trade could have played an even larger role if South Korea had not simultaneously introduced agricultural protection policies.

Üngör (2011) has compared Latin America with East Asia. He has found that differences in sectoral productivity growth rates account well for the different sectoral reallocations in the two regions, and in particular for the fact that Latin America has moved much more slowly out of agriculture.

We think that more quantitative case studies of structural transformation in currently poor countries will help to sharpen our understanding of the forces behind structural transformation in such countries. Additionally, we think it will be useful to think about the factors that influence productivity growth. Virtually all of the literature on structural transformation takes productivity changes as given, and effectively considers the implications of the exogenously given paths for productivity on the process of structural transformation. But if the paths of productivity differ significantly across countries, then it is important to ask what factors are responsible for these differences? If the differences are more pronounced in particular sectors in particular countries, what are the factors that account for this? Is it policies that influence the diffusion of technology, or perhaps policies that generate misallocation of inputs across producers?

Moving forward, we also think it will be useful to refine the standard three-sector focus of the literature. As today's advanced economies are increasingly dominated by services, it will be important to distinguish between different activities within services. For example, education and health care are very different activities than retail trade, in that they both represent an investment and tend to use very different skill intensities for the labor that they employ. The work of Jorgenson and Timmer (2011), and Duarte and Restuccia (2012) is a first step in this direction. Using data from EU KLEMS, Jorgenson and Timmer (2011) document for the European Union, Japan, and the US that there is substantial heterogeneity among services. Personal, finance, and business services have low productivity growth and increasing shares in employment and GDP whereas distribution services have rapid productivity growth and constant shares. Using data from the International Comparison Program 2005, Duarte and Restuccia (2012) study the difference between traditional and non-traditional services on a large cross section of countries where traditional services comprise mostly non-market services such as domestic and household services, education, health, and housing and non-traditional services comprise communication and transport services, insurance and financial services, and recreational and cultural services. For traditional services, they find that the relative price increases and the real expenditure share decreases with income, whereas, for non-traditional services, they find the opposite. An important task for future work is to build models that are consistent with these facts and to explore to implications that these models have for structural transformation and for aggregate outcomes.

There are many issues that we have not addressed or only touched upon in passing. One such issue is the role that human capital plays in the process of structural transformation. Buera and Kaboski (2012a) emphasize the fact that effectively all of the growth in the service sector in the US in the post WW II period occurs in high skill services. While they emphasized the role of human capital in the movement of resources from the goods producing sector to the service sector, it is also plausible that human capital may be important in understanding the movement of workers between the agricultural sector and the non-agricultural sector. In fact, the work by Caselli and Coleman (2001) that we described earlier is one paper that emphasized the role that human capital plays in this part of the structural transformation process. Recent work by Herrendorf and Schoellman (2012) provides additional evidence on this point. Using the CPS, they document for the US that wages per hour are considerably higher in non-agriculture than in agriculture. They show that this is accounted for by two main facts: non-agricultural workers are positively selected in that they have more years of schooling; and the returns to schooling and experience are higher in non-agriculture. An open question is to what extent similar findings hold in poorer countries than the US. In a recent paper, Lagakos and Waugh (2013) argue that accounting for the heterogeneous quality of labor across sectors is important in understanding the fact that poor countries seem to be have particularly low labor productivity in agriculture.

Another issue that we have not addressed is the role of industrial policy, broadly conceived. Specifically, we have chosen to discipline the analysis by assuming sectoral production functions with constant returns to scale and by abstracting from spillovers or externalities. As a result, we have interpreted structural transformation as a feature of the efficient equilibrium path, implying that there is no meaningful role for government policy. While our model framework can be used to understand how particular policies might distort the allocation of resources across sectors, there is no positive prescription for policy.

There is a sizeable literature that discusses structural transformation when there are increasing returns to scale and the equilibrium path is inefficient; see Matsuyama (1992a) for an early example and Matsuyama (2008) for specific references. The typical assumption in this literature is that non-agricultural production is subject to increasing returns, which accrue at the sectoral level, perhaps as the result of learning by doing, and which are not taken into account by households and firms. Multiple steady states then arise naturally and initial conditions determine the equilibrium path, and in particular whether the economy ends up in a poverty trap, that is, a steady state with low GDP per capita and the majority of the labor force in agriculture. These types of models suggest that policy may provide the big push that lets the economy escape from its poverty trap and leads to industrialization and self-sustaining economic growth. We have not discussed this theoretical possibility in more detail above because the empirical evidence on the success of big-push policies in particular, and industrial policies more generally, is mixed at best. But more generally, the extent to which externalities, public goods, market power, or other factors associated with inefficient equilibrium outcomes shape the process of structural transformation remains largely unresolved.

APPENDIX A: DATA SOURCES AND SECTOR ASSIGNMENTS

Historical Data 1800–2008

- Data source: GDP per capita at international dollars
  - Data on GDP per capita at 1990 international dollars are from Maddison (2010) for all countries and most years. There are some years in the early 19th century for Belgium, Netherlands, Sweden, the United Kingdom, and the United States when there are data on value added and employment shares, but Maddison does not report data on GDP per capita. We calculated GDP per capita at international dollars for these years in the following way. From alternative sources, we first calculated real GDP per capita for the missing years, and for the first year for which Maddison's

data is available. We then calculated the growth rates between the missing years and the first year for which the Maddison data is available. Lastly, we combined the growth rates with Maddison's data to calculate the per capita GDP at international dollars for the missing years. Next, we list the data sources for these calculations.

1. *Belgium*. 1835–1845: Real GDP from Groningen Growth and Development Centre, Historical National Accounts Database 2009, and population from Maddison (2010).
2. *Netherlands*. 1807–1830: Real GDP per capita from Smits et al. (2007).
3. *Sweden*. 1800–1820: Real GDP per capita from Krantz and Schn (2007).
4. *United Kingdom*. 1800–1830: Real GDP per capita from Clark (2009).
5. *United States*. Louis Johnston and Samuel H. Williamson, "What Was the US GDP Then?" MeasuringWorth, 2011.

- Data source:Value added at current prices
  - *Belgium*. 1835–1990: Groningen Growth and Development Centre, Historical National Accounts Database 2009. 1991–2007: EU KLEMS 2009.
  - *Spain*. 1885–1940: Groningen Growth and Development Centre, Historical National Accounts Database 2009, 1953–2004: Groningen Growth and Development Centre 10-sector Database 2007.
  - *Finland*. 1860–2001: Groningen Growth and Development Centre, Historical National Accounts Database 2009.
  - *France*. 1815–1938: Groningen Growth and Development Centre, Historical National Accounts Database 2009, 1950–1960: Mitchell (2007) Table J2, 1970–2005: Groningen Growth and Development Centre 10-sector Database 2007.
  - *Japan*. 1885–1940: Groningen Growth and Development Centre, Historical National Accounts Database 2009, 1953–2004: Groningen Growth and Development Centre 10-sector Database 2007.
  - *Korea*. 1911–1940: Groningen Growth and Development Centre, Historical National Accounts Database 2009, 1953–2005: Groningen Growth and Development Centre 10-sector Database 2007.
  - *Netherlands*. 1807–1913: Smits et al. (2007), 1970–2005: Groningen Growth and Development Centre 10-sector database, August 2008.
  - *Sweden*. 1800–2000: Krantz and Schn (2007), 2000–2005: Groningen Growth and Development Centre 10-sector Database, August 2008.
  - *United Kingdom*. 1801, 1941–1851: Broadberry et al. (2011) Table 8–9, 1811–1831, 1860–1910, 1950: Mitchell (2007) Table J2, 1920–1938: Feinstein (1972) Table 9, 1960–2005: Groningen Growth and Development Centre 10-sector Database 2007.

  - *United States*. 1800–1900: Agriculture and Manufacturing, Gallman (1960), Services, Gallman and Weiss (1969), 1909–1918: King (1930), 1919–1928: Kuznets et al. (1941), 1929–1946: Carter et al. eds (2006) Table Ca35–53, 1947–2008:Value Added by Industry, Gross Domestic Product by Industry Accounts, Bureau of Economic Analysis.
- Data source: Employment
  - *Belgium*. 1846–1961: Mitchell (2007) Table B1, 1970–2007: EU KLEMS 2009.
  - *Spain*. 1860–1964: Mitchell (2007) Table B1, 1970–2007: EU KLEMS 2009.
  - *Finland*. 1805–1960: Mitchell (2007) Table B1, 1970–2007: EU KLEMS 2009.
  - *France*. 1856–1968: Mitchell (2007) Table B1, 1970–2007: EU KLEMS 2009.
  - *Korea*. 1953–2005: Groningen Growth and Development Centre 10-sector Database 2007.
  - *Netherlands*. 1807–1913: Smits et al. (2007), 1920–1947: Mitchell (2007) Table B1, 1970–2005: Groningen Growth and Development Centre 10-sector Database 2008.
  - *Sweden*. 1850–2000: Krantz and Schn (2007), 2000–2005: Groningen Growth and Development Centre 10-sector Database 2008.
  - *United Kingdom*. 1801, 1813–1820 average assigned to 1817, 1851: Broadberry et al. (2011) Table 1 and Table 12, 1841: Mitchell (2007) Table B1, 1861–1938: Feinstein (1972) Table 59–60, 1948–2005: Groningen Growth and Development Centre 10-sector Database 2007.
  - *United States*. 1840–1920: Carter et al., eds (2006) Table Ba814–830, 1929–2008: NIPA Table 6.8 Persons Engaged in Production, Bureau of Economic Analysis.
- Sector assignments
  1. Agriculture corresponds to the sum of International Standard Industrial Classification (ISIC) sections A–B. If ISIC classification was not available, we assigned industries to agriculture if the source table heading said "Agriculture" or "Agriculture, forestry and fishing."
  2. Manufacturing corresponds to the sum of ISIC sections C, D, F and includes mining, manufacturing, and construction. If ISIC classification was not available, we assigned industries to manufacturing if the source table heading said "Mining" or "Extractive industries" or "Manufacturing" or "Construction".
  3. Services correspond to the sum of ISIC sections E, G–P and include utilities; wholesale; retail trade; hotels and restaurants; transport; storage and communication; finance; insurance; real estate; business services; and community social and personal services. If ISIC classification was not available, we assigned industries

to services if the source table heading said "Commerce" or "Finance" or "Trade" or "Transport" or "Communication" or "Services."

EU KLEMS 2009

- Data sources (EU KLEMS series code in brackets)
  1. Employment
     - Total hours worked by persons engaged in millions (H_EMP).
  2. Value added
     - Gross value added at current basic prices (VA).
- Sector assignment
  1. Agriculture corresponds to the sum of International Standard Industrial Classification (ISIC) sections A–B.
  2. Manufacturing corresponds to the sum of ISIC sections C, D, F and includes mining, manufacturing, and construction.
  3. Services correspond to the sum of ISIC sections E, G–P and include utilities; wholesale; retail trade; hotels and restaurants; transport; storage and communication; finance; insurance; real estate; business services; and community social and personal services.

World Development Indicators 2010

- Data sources (WDI series code in brackets)
  1. Employment
     - Employment in agriculture (% of total employment) (SL.AGR.EMPL.ZS).
     - Employment in industry (% of total employment) (SL.IND.EMPL.ZS).
     - Employment in services (% of total employment) (SL.SRV.EMPL.ZS).
  2. Value added
     - Agriculture, value added as % of GDP (NV.AGR.TOTL.ZS).
     - Industry, value added as % of GDP (NV.IND.TOTL.ZS).
     - Services, etc., value added as % of GDP (NV.SRV.TETC.ZS).
- Oil production
  1. Oil rents as % of GDP, (NY.GDP.PETR.RT.ZS).
- Sector assignment
  1. Agriculture corresponds to the sum of ISIC divisions 1–5 and includes forestry, hunting, and fishing; as well as the cultivation of crops and livestock production.

2. Manufacturing corresponds to the category "Industry" in the WDI, which is the sum of ISIC divisions 10–45 and includes mining, manufacturing, construction, electricity, water, and gas.
3. Services correspond to the sum of ISIC divisions 50–99 and include value added in wholesale and retail trade (including hotels and restaurants); transport and government, financial, professional, and personal services (such as education); health care; and real estate services. They also include imputed bank service charges, import duties, and statistical discrepancies, as well as discrepancies arising from rescaling.

National Accounts of the United Nations Statistics Division

- Data sources
  1. Gross value added by economic activity at current prices in national currency.
- Sector assignment
  1. Agriculture corresponds to ISIC sections A–B.
  2. Manufacturing corresponds to the sum of ISIC sections C–F and includes mining, manufacturing, utilities, and construction.
  3. Services correspond to the sum of ISIC sections G–P and include wholesale; retail trade; hotels and restaurants; transport; storage and communication; finance; insurance; real estate; business services; and community social and personal services.

Historical Consumption Shares UK and US

- Data source: GDP per capita at international dollars at 1990 international dollars are from Maddison (2010)
- Data source: US Consumption share in current prices
  - ○ 1900–1928: Carter et al. eds (2006).
  - ○ 1929–2008: BEA.
- Data source: UK Consumption share in current prices
  - ○ 1900–1964: Feinstein (1972).
  - ○ 1965–2008: Office of National Statistics (ONS).

Penn World Tables

- Data source: PWT6.3 (PWT series code in brackets)
  1. Real Gross Domestic Product per Capita Relative to the United States (G-K method, current price) (y).
  2. Real GDP per capita in constant prices: Chain series (rgdpch).
  3. Real GDP per worker in constant prices: Chain series (rgdpwok).
  4. Population (pop).

- Data source: PWT benchmark 1980
  - Sector assignment.
    1. Agriculture corresponds to the sum of PWT80 items 1–50.
    2. Manufacturing corresponds to the sum of PWT80 items 51–54, 56–58, 63–66, 68–78, 81–83, 91–93, 95–97, 103–108, 112-113, 118-122.
    3. Services correspond to the sum of PWT items 55, 59–62, 67, 79-80, 84–90, 94, 98–102, 109–111, 114–118, 123–125.
- Data source: PWT benchmark 1985
  - Sector assignment
    1. Agriculture corresponds to the sum of PWT80 items 1–41.
    2. Manufacturing corresponds to the sum of PWT80 items 42–47, 49–51, 56–61, 63–68, 70–72, 75–77, 82–84, 86–87, 94–97, 101, 107–109.
    3. Services correspond to the sum of PWT items 48, 52–55, 62, 69, 73–74, 78–81, 85, 88–93, 98–100, 102–106.
- Data source: PWT benchmark 1996
  - Sector assignment
    1. Agriculture corresponds to bread and cereals; meat, fish, milk, cheese and eggs; oils and fats; fruit, vegetables and potatoes; other food; non-alcoholic beverages; alcoholic beverages.
    2. Manufacturing corresponds to tobacco; clothing including repairs; footwear including repairs; fuel and power; furniture; floor coverings and repairs; other household goods including household textiles; household appliances and repairs; personal transportation equipment.
    3. Services correspond to gross rent and water charges; medical and health services; operation of transportation equipment; purchased transport services; communication; recreation and culture; education; restaurants, cafes and hotels; other goods and services.

OECD Consumption Expenditure Data

- Data source:
  - Final consumption expenditure of households, national currency, current prices, OECD National Accounts Statistics. This data set includes the final consumption expenditure of households broken down by the COICOP (Classification of Individual Consumption According to Purpose) classification and by durability.
- Sector assignment (COICOP codes in brackets)
  1. Food: "Food and non-alcoholic beverages" (P31CP010).
  2. Manufactured goods: "Durable goods" plus "Semi-durable goods" plus "Non-durable goods" minus "Food and non-alcoholic beverages" (P311B+P312B+ P313B- P31CP010).
  3. Services: Services (P314B).

- Construction of the data for E7 countries (Austria, Denmark, Finland, France, Italy, Netherlands, United Kingdom) for the period 1980–2009. Consumption expenditure data are from the National Accounts of Eurostat both in local currency and euro. Then, for each year and each country, a conversion rate between local currency and euro was calculated by dividing total consumption expenditures in local currency with total consumption expenditures in euros. The three expenditure items expressed in local currency were converted into euros using this conversion rate, and then they were aggregated.

Real GDP per capita at 1990 International Dollars

- Prior to 1970, the data on GDP per capita at 1990 international dollars are from Maddison (2010) for all years and countries if it was available.
- After 1970, we constructed real GDP per capita at 1990 international dollars in the following ways. The data on GDP per capita at 1990 international dollars for the United States were taken from Maddison (2010). The real GDP per capita of the United States was multiplied by the data on real GDP per capita relative to the United States to calculate the real GDP per capita at 1990 international dollars for each country and each year.

APPENDIX B: PANEL REGRESSIONS

To get a balanced panel, we only include countries with data over the entire period 1970–2007. In addition, we restrict the sample in three ways: and we exclude countries in which the average ratio of oil rent to GDP exceeds 20% during 1970–2007[55]; we exclude countries with average populations of fewer than a million during 1970–2007; and we exclude the former communist countries. The reason for these exclusion criteria is that the sector composition in these countries may be distorted. This leaves 103 countries.

ACKNOWLEDGMENTS

We thank Francesco Caselli, Rachel Ngai, Todd Schoellman, and Gustavo Ventura for commenting on earlier versions of this manuscript, and Christopher Herrington for excellent research assistance. For financial support, Herrendorf thanks the Spanish Ministry of Education (Grants ECO2009-11165 and ECO2012-31358); Rogerson thanks both the NSF and the Korea Science Foundation (WCU-R33-10005); and Valentinyi thanks the Hungarian Scientific Research Fund (OTKA) (Project K-105660-ny).

[55] The oil-rent-to-GDP ratio is taken from the WDI.

REFERENCES

Abraham, Katharine, Katz, Larry, 1986. Cyclical unemployment: sectoral shifts or cyclical unemployment? Journal of Political Economy 94, 507–522.

Acemoglu, Daron, Guerrieri, Veronica, 2008. Capital deepening and non-balanced economic growth. Journal of Political Economy 116, 467–498.

Adamopoulos, Tasso, 2011. Transportation costs, agricultural productivity, and cross-country income differences. International Economic Review 52, 489–521.

Aguiar, Mark, Hurst, Eric, 2007. Measuring leisure: the allocation of time over five decades. Quarterly Journal of Economics 122, 969–1006.

Akbulut, Rahsan, 2011. Sectoral changes and the increase in women's labor force participation. Macroeconomic Dynamics 15, 240–264.

Alvarez-Cuadrado, Francisco, Poschke, Markus, 2011. Structural change out of agriculture: labor push versus labor pull. American Economic Journal: Macroeconomics 3, 127–158.

Alvarez-Cuadrado, Francisco, Van Long, Ngo, Poschke, Markus, 2012. Capital-Labor Substitution, Structural Change, and Growth. Manuscript. McGill University, Montreal.

Bah, El-Hadj, 2008. Structural Transformation in Developed and Developing Countries. Manuscript. Arizona State University, Tempe, AZ.

Bah, El-hadj, Brada, Josef C., 2009. Total factor productivity growth, structural change and convergence in transition economies. Comparative Economic Studies 51, 421–446.

Bar, Michael, Leukhina, Oksana, 2010. Demographic transition and industrial revolution: a macroeconomic investigation. Review of Economic Dynamics 13, 424–451.

Barro, Robert J., Sala-i-Martin, Xavier, 1992. Convergence. Journal of Political Economy 100, 223–251.

Baumol, William J., 1967. Macroeconomics of unbalanced growth: the anatomy of the urban crisis. American Economic Review 57, 415–426.

Betts, Caroline M., Giri, Rahul, Verma, Rubina, 2011. Trade, Reform, and Structural Transformation in South Korea. Manuscript. University of Southern California.

Blundell, Richard, 1988. Consumer behaviour: theory and evidence—a survey. Economic Journal 98, 16–65.

Boppart, Timo, 2011. Structural Change and the Kaldor Facts in a Growth Model with Relative Price Effects and Non-Gorman Preferences. Working Paper 2, University of Zürich.

Broadberry, Stephen, Campbell, Bruce M.S., van Leeuwen, Bas, 2011. The Sectoral Distribution of the Labour Force and Labour Productivity in Britain. Manuscript. London School of Economics.

Buera, Francisco J., Kaboski, Joseph P., 2009. Can traditional theories of structural change fit the data? Journal of the European Economic Association 7, 469–477.

Buera, Francisco J., Kaboski, Joseph P., 2012. The rise of the service economy. American Economic Review 102, 2540–2569.

Buera, Francisco J., Kaboski, Joseph P., 2012. Scale and origins of structural change. Journal of Economic Theory 147, 684–712.

Carter, Susan B., Gartner, Scott Sigmund, Haines, Michael R., Olmstead, Alan L., Sutch, Richard, Wright, Gavin (Eds.), 2006. Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition. Cambridge University Press, New York.

Carvalho, Vasco M., Gabaix, Xavier, 2013. The great diversification and its undoing. American Economic Review 103, 1697–1727.

Caselli, Francesco, 2005. Accounting for cross-country income differences. In: Aghion, Philippe, Durlauf, Steven (Eds.), Handbook of Economic Growth, vol. 1A. North Holland, Amsterdam and New York, pp. 679–742 (Chapter 9).

Caselli, Francesco, Coleman, Wilbur John, 2001. The U.S. structural transformation and regional convergence: a reinterpretation. Journal of Political Economy 109, 584–616.

Chang, Yongsung, Hornstein, Andreas, 2011. Transition Dynamics in the Neoclassical Growth Model: The Case of South Korea. Working Paper 11–04, Federal Reserve Bank of Richmond, Richmond.

Chenery, Hollis B., 1960. Patterns of industrial growth. American Economic Review 50, 624–653.

Clark, Colin, 1957. The Conditions of Economic Progress, third ed. Macmillan, London.

Clark, Gregory, 2009. The Macroeconomic Aggregates for England, 1209–2008. Manuscript. University of California, Davis.

Da Rocha, Jos Maria, Restuccia, Diego, 2006. The role of agriculture in aggregate business cycles. Review of Economic Dynamics 9 (3), 455–482.

Davis, Steven J., Haltiwanger, John C., 1992. Gross job creation, gross job destruction, and employment reallocation. Quarterly Journal of Economics 107, 819–63.

Dekle, Robert, Vandenbroucke, Guillaume, 2012. A quantitative analysis of China's structural transformation. Journal of Economic Dynamics and Control 36, 119–135.

Dennis, Benjamin N., Iscan, Talan B., 2009. Engle versus Baumol: accounting for structural change using two centuries of U.S. data. Explorations in Economics History 46, 186–202.

Diamond, Jared M., 1997. Guns, Germs, and Steel: The Fates of Human Societies. W.W. Norton, New York.

Duarte, Margarida, Restuccia, Diego, 2010. The role of the structural transformation in aggregate productivity. Quarterly Journal of Economics 125, 129–173.

Duarte, Margarida, Restuccia, Diego, 2012. Relative Prices and Sectoral Productivity. Manuscript. University of Toronto.

Eaton, Jonathan, Kortum, Samuel, 2002. Thechology, geography, and trade. Econometrica 70, 1741–1879.

Echevarria, Cristina, 1997. Changes in sectoral composition associated with economic growth. International Economic Review 38, 431–452.

Feinstein, Charles H., 1972. National Income, Expenditure and Output of the United Kingdom 1855–1965. Cambridge University Press, Cambridge.

Foellmi, Reto, Zweimüller, Josef, 2008. Structural change, Engel's consumption cycles, and Kaldor's facts of economic growth. Journal of Monetary Economics 55, 1317–1328.

Fuchs, Victor, 1968. The Service Economy. Columbia University Press, New York.

Gallman, Robert E., 1960. The United States commodity output, 1839–1899. In: Parker, William N., (Ed.), Trends in the American Economy in the Nineteenth Century, NBER Studies in Income and Wealth. Princeton University Press, Princeton.

Gallman, Robert E., Weiss, Thomas J., 1969. Production and productivity in the service industries. In: Fuchs, Victor R., (Ed.), The Service Industries in the Nineteenth Century, NBER Studies in Income and Wealth. Columbia University Press, New York.

Galor, Oded, Weil, David N., 1996. The gender gap, fertility, and growth. American Economic Review 86, 374–387.

Goldin, Claudia, 1995. The U-shaped female labor force function in economic development and economic history. In: Paul Schultz, T. (Ed.), Investment in Women's Human Capital and Economic Development. The University of Chicago Press, Chicago, IL.

Goldin, Claudia, 2006. Ely lecture: the quiet revolution that transformed women's employment, education, and family. American Economic Review Papers and Proceedings 96, 1–21.

Gollin, Douglas, Rogerson, Richard, 2010. Agriculture, Roads and Economic Development in Uganda. Working Paper.

Gollin, Douglas, Parente, Stephen L., Rogerson, Richard, 2002. The role of agriculture in development. American Economic Review, Papers and Proceedings 92, 160–164.

Gollin, Douglas, Parente, Stephen L., Rogerson, Richard, 2006. The food problem and the evolution of international income levels. Journal of Monetary Economics 54, 1230–1255.

Greenwood, Jeremy, Seshadri, Ananth, 2005. Technological progress and economic transformation. In: Aghion, Philippe, Durlauf, Steven (Eds.), Handbook of Economic Growth, vol. 1B. North Holland, Amsterdam and New York, pp. 1225–1273 (Chapter 19).

Greenwood, Jeremy, Hercowitz, Zvi, Krusell, Per, 1997. Long-run implication of investment-specific technological change. American Economic Review 87, 342–362.

Grossman, Gene M., Krueger, Alan B., 1995. Economic growth and the environment. Quarterly Journal of Economics 110, 353–377.

Hall, Robert E., Jones, Charles I., 2007. The value of life and the rise in health spending. The Quarterly Journal of Economics 122, 39–72.

Hansen, Gary .D., Prescott, Edward C., 2002. Malthus to solow. The American Economic Review 92 (4), 1205–1217.

Hayashi, Fumio, Prescott, Edward C., 2008. The depressing effect of agricultural institutions on the prewar Japanese economy. Journal of Political Economy 116, 573–632.

Herrendorf, Berthold, Schoellman, Todd, 2012. Why is Measured Labor Productivity so Low in Agriculture? Manuscript. Arizona State University.

Herrendorf, Berthold, Valentinyi, Ákos, 2012. Which sectors make poor countries so unproductive? Journal of the European Economic Association 10, 323–341.

Herrendorf, Berthold, Schmitz Jr., James, Teixeira, Arilton, 2012. The role of transportation in U.S. economic development: 1840–1860. International Economic Review 53, 693–715.

Herrendorf, Berthold, Herrington, Christopher, Valentinyi, Ákos, 2013. Sectoral Technology and Structural Transformation. Manuscript. Arizona State University.

Herrendorf, Berthold, Rogerson, Richard, Valentinyi, Ákos, 2009. Two perspectives on preferences and structural transformation. American Economic Review.

Herz, Benedikt, van Rens, Thijs, 2011. Structural Unemployment. Manuscript.

Hnatkovska, Viktoria, Lahiri, Amartya, 2012. Structural Transformation and the Rural–Urban Divide. Manuscript. University of British Columbia.

Jacobson, Louis S., LaLonde, Robert J., Sullivan, Daniel G., 1993. Earnings losses of displaced workers. American Economic Review 83, 685–709.

Johnston, Bruce F., Kilby, Peter, 1975. Agriculture and Structural Transformation: Economic Strategies in Late-Developing Countries. Oxford University Press.

Johnston, Bruce F., Mellor, John W., 1961. The role of agriculture in economic development. American Economic Review 51, 566–593.

Jorgenson, Dale W., Timmer, Marcel P., 2011. Structural change in advanced nations: a new set of stylised facts. Scandinavian Journal of Economics 113, 1–29.

Karádi, Péter, Miklós Koren, 2012. Cattle, Steaks and Restaurants: Development Accounting when Space Matters. Central European University, Mimeo.

Kim, Dae-Il, Topel, Robert H., 1995. Labor Markets and economic growth: lessons from Korea's industrialization, 1970–1990. In: Freeman, Richard B., Katz, Lawrence F. (Eds.), Differences and Changes in Wage Structure. University of Chicago Press for NBER, Chicago, pp. 227–264.

King, Wilford Isbell, 1930. Industrial origin of total realized income. In: King, Wilford Isbell (Ed.), The National Income and Its Purchasing Power. National Bureau of Economic Research.

King, Robert G., Rebelo, Sergio, 1993. Transitional dynamics and economic growth in the neoclassical model. American Economic Review 83, 908–931.

Kongsamut, Piyabha, Rebelo, Sergio, Xie, Danyang, 2001. Beyond balanced growth. Review of Economic Studies 68, 869–882.

Krantz, Olle, Schn, Lennart, 2007. Swedish Historical National Accounts 1800–2000 Lund Studies in Economic History. Lund University, Lund, <http://www.ekh.lu.se/database/lu-madd/NationalAccounts/default.htm>.

Kuznets, Simon, 1966. Modern Economic Growth. Yale University Press, New Haven.

Kuznets, Simon, 1973. Modern economic growth: findings and reflections. Amercian Economic Review 63, 247–258.

Kuznets, Simon, Epstein, Lillian, Jenks, Elizabeth, 1941. Distribution by industrial source. In: Kuznets, Simon (Ed.), National Income and Its Composition, 1919–1938, vol. 1. National Bureau of Economic Research.

Lagakos, David, Waugh, Mike, 2013. Selection, agriculture, and cross-country productivity differences. American Economic Review 103, 948–980.

Laitner, John, 2000. Structural change and economic growth. Review of Economic Studies 67, 545–561.

Lawver, Daniel, 2011. Measuring Quality Increases in the Medical Sector. Manuscript. Arizona State University.

Lee, Donghoon, Wolpin, Kenneth I., 2006. Intersectoral labor mobility and the growth of the service sector. Econometrica 74, 1–46.

Lilien, David, 1982. Sectoral shifts and cyclical unemployment. Journal of Political Economy 90, 777–793.

Lucas, Robert, Prescott, Edward, 1974. Equilibrium search and unemployment. Journal of Economic Theory 7, 188–209.

Maddison, Angus, 2010. Statistics on world population, GDP and per capita GDP, 1-2008 AD, University of Groningen, Groningen.

Matsuyama, Kiminori, 1992a. Agricultural productivity, comparative advantage, and economic growth. Journal of Economic Theory 58, 317–334.

Matsuyama, Kiminori, 1992b. A Simple model of sectoral adjustment. Review of Economic Studies 59, 375–388.

Matsuyama, Kiminori, 2008. Structural change. In: Durlauf, Steven N., Blume, Lawrence E. (Eds.), The New Palgrave Dictionary of Economics, second ed. Palgrave Macmillan.

Matsuyama, Kiminori, 2009. Structural change in an interdependent world: a global view of manufacturing decline. Journal of the European Economic Association 7 (2–3), 478–486.

Messina, Julian, 2006. The role of product market regulations in the process of structural change. European Economic Review 50, 1863–1890.

Mitchell, Brian, 2007. International Historical Statistics. Palgrave Macmillan, London.

Moro, Alessio, 2012. The structural transformation between manufacturing and services and the decline in the U.S. GDP volatility. Review of Economic Dynamics 15, 402–415.

Muellbauer, John, 1975. Aggregation, income distribution and consumer demand. Review of Economic Studies 62, 526–543.

Muellbauer, John, 1976. Community preferences and the representative consumer. Econometrica 44, 979–999.

Murphy, Kevin M., Shleifer, Andrei, Vishny, Robert W., 1989. Income distribution, market size, and industrialization. Quarterly Journal of Economics 104, 537–564.

Ngai, L. Rachel, Petrongolo, Barbara, 2013. Gender Gaps and the Rise of the Service Economy. Manuscript. London School of Economics.

Ngai, L. Rachel, Pissarides, Chrisopher A., 2007. Structural change in a multisector model of growth. American Economic Review 97, 429–443.

Ngai, L. Rachel, Pissarides, Chrisopher A., 2008. Trends in hours and economic growth. Review of Economic Dynamics 11, 239–256.

Nickell, Stephen, Redding, Stephen, Swaffield, Joanna, 2002. Educational Attainment, Labour Market Institutions, and the Structure of Production. Working Paper. Centre for Economic Performance, London School of Economics, London, UK.

Olivetti, Claudia, 2012. The Female Labor Force and Long-Run Development: The American Experience in Comparative Perspective. Manuscript. Boston University, Boston.

O'Mahony, Mary, Timmer, Marcel P., 2009. Output, input, productivity measures at the industry level: the EU KLEMS database. Economic Journal, 119OMahony-Timmer-2009, F374–F403.

Prescott, Edward C., 2004. Why do Americans work so much more than Europeans? Federal Reserve Bank of Minneapolis Quarterly Review 28, 2–13.

Ramey, Valery A., Francis, Neville, 2009. A century of work and leisure. American Economic Journal: Macroeconomics 1, 189–224.

Ray, Debraj, 2010. Uneven growth: a framework for research in development economics. Journal of Economic Perspectives 24, 45–60.

Rendall, Michelle, 2010. The Rise of the Service Sector and Female Market Work: Europe vs. US. Technical Report, University of Zurich, Manuscript.

Restuccia, Diego, Yang, Dennis Tao, Zhu, Xiaodong, 2008. Agriculture and aggregate productivity: a quantitative cross-country analysis. Journal of Monetary Economics 55, 234–250.

Rogerson, Richard, 1991. Sectoral shifts and cyclical fluctuations. Revista de Analisis Economico 6, 37–46.

Rogerson, Richard, 2006. Understanding differences in hours worked. Review of Economic Dynamics 9, 365–409.

Rogerson, Richard, 2008. Structural transformation and the deterioration of European labor market outcomes. Journal of Political Economy 116, 235–259.

Rubina, Verma, 2012. Can total factor productivity explain value added growth in services. Journal of Development Economics 99, 163–187.

Sahin, Aysegul, Song, Joseph, Topa, Giorgio, Violante, Gianluca, 2011. Measuring Mismatch in the US Labor Market. Manuscript.

Schultz, Theodore W., 1953. The Economic Organization of Agriculture. McGraw-Hill, New York.

Smits, Jan-Pieter, Horlings, Edwin, van Zanden, Jan Luiten, 2007. Dutch GNP and its Components 1800-1913 Lund Studies in Economic History. University of Groningen, Groningen, <http://nationalaccounts.niwi.knaw.nl>.

Sposi, Michael, 2011. Evolving Comparative Advantage, Structural Change, and the Composition of Trade. Manuscript. University of Iowa.

Stefanski, Radek, 2013. On the Mechanics of the Green Solow Model. Manuscript. Laval University.

Swiecki, Tomasz, Intersectoral Distortions, Structural Change and the Welfare Gains from Trade. Manuscript. Princeton University 2013.

Syrquin, Moshe, 1988. Patterns of structural change. In: Chenery, Hollis, Srinivasan, T.N. (Eds.), Handbook of Development Economics, vol. 1. North Holland, Amsterdam and New York, pp. 203–273 (Chapter 7).

Teignier, Marc, 2012. The Role of Trade in Structural Transformation. Manuscript. Universidad de Alicante.

Tiffin, Richard, Irz, Xavier, 2006. Is agriculture the engine of growth? Agricultural Economics 35, 79–89.

Timmer, Peter C., 1988. The Agricultural transformation. In: Handbook of Development Economics, vol. 1. North Holland, Amsterdam and New York, pp. 275–331 (Chapter 8).

Timmer, Marcel P., Inklaar, Robert, O'Mahony, Mary, van Ark, Bart, 2010. Economic Growth in Europe: A Comparative Industry Perspective. Cambridge University Press.

Üngör, Murat, 2011. Productivity Growth and Labor Reallocation: Latin America versus East Asia. Manuscript. Central Bank of Turkey, Istanbul.

Uzawa, Hirofumi, 1963. On a two-sector model of economic growth II. Review of Economic Studies 30, 105–118.

Valentinyi, Ákos, Herrendorf, Berthold, 2008. Measuring factor income shares at the sectoral level. Review of Economic Dynamics 11, 820–835.

Yi, Kei-Mu, Zhang, Jing, 2010. Structural Transformation in an Open Economy. Manuscript. University of Michigan.

CHAPTER SEVEN

The Chinese Growth Miracle

Yang Yao
China Center for Economic Research, National School of Development, Peking University, China

Abstract

This chapter provides a review of China's economic growth since 1978. Studying China's economic success may shed new light on the political economy of growth, the impacts of the ascent of large countries on the rest of the world, and the relationship between inequality and economic growth. The chapter starts with a review of the characteristics of China's economic growth and compares it with those of several similar economies. Then it shows how China's economic success has been created by innovative institutional arrangements as well as adherence to the policy advice prescribed by neoclassical economics. After that, the chapter describes China's export-led growth model and analyzes its causes and the structural imbalances associated with it. Lastly, the chapter presents data for income inequality in China and discusses how inequality may affect China's prospect of avoiding the middle-income trap.

Keywords

Chinese growth miracle, Political economy of growth, Export-led growth, Global imbalances, Middle-income trap

JEL Classification Codes

O1, O5, F3, F4, D3

7.1. INTRODUCTION

The economic ascent of China since the end of the 1970s provides an interesting and challenging case for the study of economic growth. China is certainly not the only success story in recent history; the four East Asian Tigers achieved comparable, if not better, records of economic growth in their fast-growing periods; and Brazil, a large country, did almost as well as China between 1950 and 1980. Nor is China likely to be the last success story; India has been following China closely. However, studying China may offer new insights into the economics of growth in several areas, particularly those related to the political economy of growth, the rise of large countries, and the relationship between inequality and growth.

In the area of political economy, China provides an experimental site for the study of authoritarian regimes. Like several other countries during their periods of fast growth, China has been under an authoritarian regime. While the consensus in the literature is that democracy, and for that matter, authoritarianism, is neither sufficient nor necessary for

Handbook of Economic Growth, Volume 2B
ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53540-5.00007-0

economic growth, there is an emerging interest in studying the variations among democratic and authoritarian regimes. In the case of China, its sheer size renders centralized absolute rule impossible. In the last several decades, the Chinese regime has developed various unique institutions that have helped incentivize local officials as well as held the country together. Studying those institutions may provide clues for why some authoritarian regimes are more successful than others, and from there one may draw some general implications for economic growth at large.

Related to the political economy of growth, the Chinese economy has been characterized by continuous deregulation through reforms. Two distinctive features have emerged from this process. One is that unlike other transition countries, China has managed the transition from a planning economy to a mixed economy having not only avoided major economic disruptions, but also maintained high economic growth. The other is that while China's deregulation has been gradual, it has not suffered from the pitfall of entrenching interest-group politics that has plagued deregulation in many other countries (Murphy et al. 1992). The key to understanding China's success may lie in the many contingent institutions that have been created as transitory institutions bridging the old and new regulatory regimes. These institutions are not perfect, but bring enough changes and at the same time cushion the shocks imposed upon the stakeholders of the old regime. Studying those contingent institutions will enrich our understanding of institutions and how they impact on economic performance.

As for the rise of large countries, China provides an example for their impacts on the rest of the world in the 21st century. There have been precedents of the rise of large countries; the rise of the United States at the end of the 19th century is the most prominent example. However, several factors make the rise of large countries in this century much different from its historical precedents. The most obvious is that the supply of fossil energy is becoming an issue today while it was not a hundred years ago, yet economic growth is still largely based on fossil energy. Even if technological progress could solve the supply problem, increasing greenhouse gas emission still calls into question whether fossil energy-based economic growth can be sustainable. The world order is also quite different today than a hundred years ago. While economic growth was confined to a small number of countries in the 19th century, today it is a global phenomenon. The "fallacy of composition" then begs the question whether export-led growth of large countries squeezes the space of growth of other countries; the rules applying to small open economies may not be readily applicable to large countries. This is no more evident in the round of global imbalances that started in the early 2000s. Like many precedents, this round of imbalances has led to a major crisis. Unlike in the past rounds, though, some large developing countries, noticeably China, have joined and altered the global production chain and become surplus countries this time. It thus becomes a question whether the world can absorb the rise of large countries in this century.

Against this background, it could be a fruitful exercise for economists to study China's export-led growth (ELG) model. This growth model is causing tensions in the world.

In the first decade of the 21st century, China's current account surplus has risen by an unprecedented rate and has become a mirror image of the rising current account deficit in the United States. Various theories, ranging from as simple as manipulated exchange rates to more sophisticated ones accounting for the role of finance, have been proposed in the literature to explain global imbalances. However, few of them study China as a stand-alone case, yet such an exercise can make several contributions to the theory of economic growth.

Historically, almost all large countries had current account surplus in their high-growth periods. To the extent that a current account deficit is unsustainable, it is natural to expect a fast-growing country to run current account surplus than deficit. Though it still begs intellectual exploration as to why large current account surpluses could be persistent. Short-term causes may not be good explanations because persistent current account surpluses have reoccurred many times in history. To understand the nature of global imbalances, researchers have to study long-term structural factors that have shaped the growth trajectories of emerging large countries. In this regard, China provides a contemporary sample for serious studies.

China's structural imbalances can be summarized in three puzzles. First, there has been a secular decline of the share of household income in GDP since the mid-1990s. Second, the national saving rate has increased steadily; in particular, corporate savings have increased as fast as household savings. Third, China has become a large net exporter of capital while its return to capital is higher than in most other countries.

The starting point to understand those three puzzles is to look at the long-term structural factors. Among them, the double transition, namely, abrupt demographic transition and large-scale movement of workers from the countryside to the city, is fundamental. Due to its strict family planning policy, China's demographic transition has been tremendously accelerated compared with similar developing countries such as India. In the meantime, fast industrialization has brought millions of people out of the countryside. A direct consequence of this double transition is unprecedented growth of an already large industrial labor force, which ought to impact on China's growth model. In a sense, the shock brought by China's export-led growth to the world can be traced back to the surge of China's industrial labor force. Here is where the large-country effect kicks in. Notwithstanding its fast growth, China is barely a middle-income country and there are tremendous regional disparities in the level of income. Therefore, structural change will continue and move inland. In addition, China will still enjoy demographic dividends in the next 20 years although their size will decline. As a result, China's episode of fast growth may be longer than its predecessors.

Fast growth may well place China on the surplus side of the global imbalances. The life-cycle theory predicts that a country's national saving rate is proportional to its growth rate; on the other hand, the growth rate of investment would be constrained by diminishing marginal returns to capital. Therefore, a country with high growth rates would be more likely to run current account surplus.

Studying the long-term structural factors by no means, though, preempts the study of short-term factors. China's currency peg is always an issue of hot debates. While it remains an empirically controversial question as to whether the peg has led to China's burgeoning current account surplus, a somewhat neglected question is why the fast growth of current account surplus has not led to serious real appreciation of the Chinese yuan. Similar phenomena also happened in Japan, Germany, and the Four Asian Tigers during their fast-growing periods. Finding out the commonalities among those economies may provide new insights into the course of economic development as well as into the relationship between economic catching-up and the real exchange rate formulated by the Balassa-Samuelson effect.

In the third area, a study of China can contribute to the understanding of the relationship between inequality and economic growth. Cross-country studies have generally established a negative correlation between inequality and economic growth and there is a large body of literature on the mechanisms. However, those mechanisms are not sensitive to the stage of economic development and most of them suggest a perpetuated poverty trap. This is not consistent with the Kuznets Curve that shows growth can go hand in hand with rising inequality in the initial stage of economic growth. A more sensible approach, thus, is to take into account the stage of economic development and study the different roles of inequality in different stages of economic development. Once again, China provides a contemporary case in this regard.

Inequality has risen quickly along with fast economic growth in China; worsening income distribution seems to have not stalled economic growth in the country. The question is whether this seemingly harmonious relationship can continue in the future. Pertinent to this question is the so-called middle-income trap, namely, a situation in which a country stops its catch-up process after its per-capita income has reached the middle-income level. Will China follow some of the countries—notably some Latin American countries and the Soviet Bloc—to lose growth as it moves into the higher-middle-income group defined by the World Bank? China shares the characteristics of both the Latin American countries in their fast-growing periods and the Soviet Bloc before it fell apart in 1989. The Latin American countries were characterized with high degrees of inequality, and the Soviet Bloc, apart from its rigid political regime, suffered from an investment-driven growth model. China has both; as a matter of fact, they are interrelated in the country. Studying China may shed new insights into the understanding of the inequality-growth nexus.

This chapter is aimed at providing a synthesis of the recent literature on China's economic growth, with more space devoted to the three areas above. Among these three areas, though, more attention will be paid to the second because it is a relatively new area. There are excellent review papers, particularly Xu (2011) and Brandt et al. (2011), dealing with the role of institutions in China's recent economic ascent. This chapter will not repeat what those two papers have already said. On the other hand, the economic growth and development literature has provided theories and evidence for the negative impacts of income inequality on growth and there are intensive studies on income inequality in China. This chapter will present some evidence for China's rising income inequality and then spend more space discussing its implications for the middle-income trap.

The rest of the chapter is arranged as follows. In Section 7.2 below, we will first introduce China's economic growth since the 1950s and compare it with some other major miracle economies. Then we will provide a review of three major features of China's economic growth, namely, economic transition, structural change, and export-led growth. Based on the experiences of some key predecessors (Brazil, Korea, and Japan) this section will also try to extrapolate China's economic growth toward 2020. Sections 7.3 and 7.4 provide a review of the theories and explanations of China's high economic performance. Section 7.3 focuses on the more conventional set of explanations that resort to initial conditions, sound government policies, and correct development strategies. Section 7.4 moves on to discuss the political-economy explanations. While some of the political-economy issues (e.g. fiscal decentralization) are relatively well understood in the literature, many others are still under-researched but have the potential to generate useful results for general economics. This includes the nature of second-best institutions; interplays between the bureaucracy and the economy; and the reemerging debate of the role of the state. Section 7.5 is devoted to explaining China's export-led growth model and discussing its sustainability. On the causes of this model, this section will emphasize the role of China's double transition. On the sustainability of the model, this section will deal with two sets of issues. One is whether the ELG has led China to a trade trap, namely, a state in which China is trapped in exporting low value-added products. The other is whether China's export expansion will be eventually checked by the fallacy of composition. Section 7.6 discusses China's structural imbalance problem in the context of global imbalances. After presenting the three puzzles manifesting the problem, this section will discuss the major causes of the imbalance. Both the long-term and structural causes and short-term government policies will be discussed. In the end, implications will be drawn as to the understanding of the global imbalances. Section 7.7 presents evidence of worsening income distribution and discusses its implications for China's long-run economic growth. In particular, this section will discuss the possibility of China falling into a middle-income trap. Lastly, Section 7.8 points out the areas that are open to further research.

7.2. ECONOMIC GROWTH IN CHINA: ACHIEVEMENTS AND FEATURES

Since the economic opening in the late 1970s, China has undergone a profound transition from a centrally planned economy to a mixed economy; in the meantime, it has managed to achieve the growth records of high-performing countries (regions). China therefore shares commonalities with both high-performing economies and transition economies. On the other hand, China's growth also has its distinctive features due to its historical past and some of the policies that are still implemented today.

7.2.1 China's Economic Growth in a Historical and Comparative Perspective

In his recent book *Why the West Rules—for Now*, Morris (2011) constructs a social development index for the East and the West from 14,000 BCE to 2000 CE. For the most part, the West led the East, except between 500 CE and 1800 CE when the East took the lead. Since 1800, the Industrial Revolution has led to the Great Divergence which separated the West, characterized by a civilization based on power-driven industrial expansion, from the East that has remained agrarian for most of the last 200 years. Table 7.1, adopted from Maddison (2001), provides a comparison of China and the world in terms of population and GDP between 1700 and 2015 (estimated). Thanks to crops (corn, potato, and sweet potato) brought from the New World, China's population soared from 1700. By 1900, it had almost tripled. Despite the wars in the first half of the 20th century, population growth accelerated although the fastest period of growth wasn't until the wars ended in 1949. Between 1950 and 2001, China's population more than doubled. However, this fast growth was dwarfed by global population growth, making China's share of the world population drop from 37% in 1820, to 21% in 2001. In 1820, China's share of the world GDP was almost as large as its share of population, making China a middle-income country by the standards of that time. Since then, China began a secular decline and by 1950,

Table 7.1 China in comparison with the world: 1700–2015

	1700	1820	1900	1950	2001	2015
			Population (mil.)			
China	138	381	400	547	1275	1387
World	603	1042	1564	2521	6149	7154
China in world (%)	23	37	26	22	21	19
			GDP (bil., 1990 international dollar)			
China	83	229	218	240	4570	11463
World	371	696	1973	5326	37148	57947
China in world (%)	22	33	11	5	12	20
			Per-capita GDP (1990 international dollar)			
China	600	600	545	439	3583	8265
World	615	668	1262	2110	6041	7154
China in world	0.98	0.90	0.43	0.21	0.59	1.16

Source: Maddison (2001).

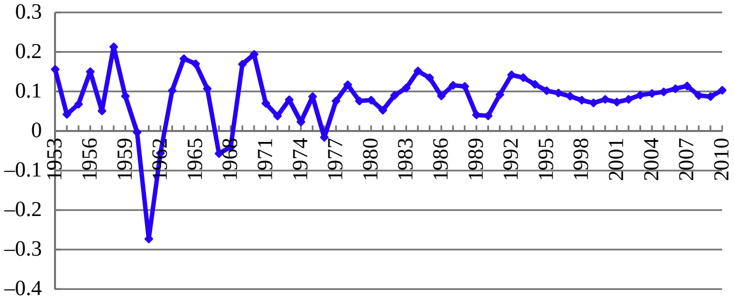


Figure 7.1 China's growth rates: 1953–2010. *Source: NBS at* *www.stats.gov.cn*.

its share of world GDP was a miserable 5%. With its per-capita GDP standing at 21% of the world average, China was definitely one of the poorest countries in the world at the time.

To Morris, who uses a century or longer time periods as his observation unit, China's decline between 1800 and 1950 was hardly something that should be pondered upon. However, this time period has become a reference point for China's economic growth ever since. Viewed against the decline of this period of time, China's economic ascent since 1950 can be seen as a long-run recovery to its position in the early 1800s. When we look at the period from 1950 to 2010 under a microscope, though, we find that growth was highly uneven across time, to say the least. Figure 7.1 provides data of China's real growth rates from 1953 to 2010 based on official statistics.[1] Clearly, we can divide the 58 years into two periods, one before 1978, and the other after. The year 1978 was the year when China began its path to economic reform and opening. Before that year, the average growth rate was 6.5%; afterwards, the average growth rate was 9.5%. The average growth rate between 1953 and 1977 was not low, but it was probably exaggerated by the artificially high prices assigned to heavy industrial products, the bulk of industrial output at the time. More importantly, there were large fluctuations during that period. There were three dramatic cycles with several years of high growth followed by one or several years of negative growth. The first cycle began in the early 1950s and ended in the Great Famine of 1959-1962 when a decline of 28% was registered for one year (1961). The second cycle was caused by the Cultural Revolution and was almost a replay of the first cycle although the trough was shallower. The third cycle was not as abrupt as the first two, but still ended with a negative growth rate in 1976. There were also ups and downs

[1] There are doubts about the reliability of China's official statistics issued by the National Bureau of Statistics (NBS), and there are studies providing adjusted statistics (e.g. Rawski, 2001;Young, 2003). However, the NBS is the only source that provides consistent time-series and spatially comparable data. As a result, most of the data used in this chapter come from the NBS although the Penn World Table (PWT) and the World Development Index (WDI) of the World Bank are consulted when international comparisons are made. The official data is treated with some caution in the text.

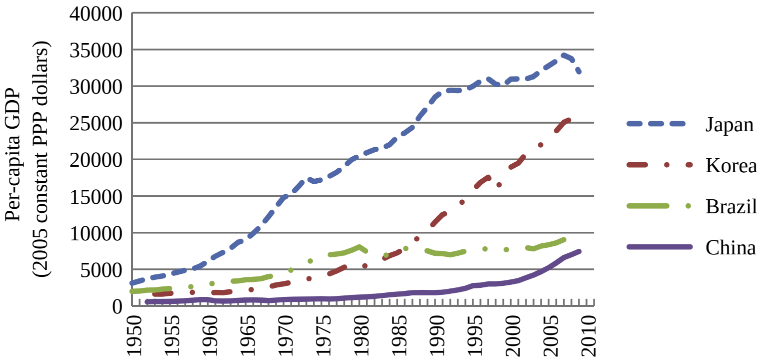


Figure 7.2 Per-capita GDP in four countries: 1950–2009. *Source: PWT 7.0.*

in the period 1978–2010, but there was never a year of negative growth. Clearly, reform and opening was the key to explaining the different records of performance before and after 1978. Although there have been problems with the official data,[2] the improvement in living standards is evident in almost every corner of the country. One indicator is the growth of automobile sales. In China in 2001, only 2.35 million cars were sold; 10 years later, that figure was 18.5 million, the highest in the world.[3]

It is a worthy exercise to compare China with other high-performing economies that have emerged since World War II. Here we choose three large countries, Brazil, Japan, and Korea, for the comparison.[4] Using the Penn World Table (PWT) data, Figure 7.2 plots China's per-capita GDP with those of the three countries since 1950.[5] Among the four countries, Korea is the only country that has maintained continuous high growth rates. China only began to take off around 1980. Japan had high growth before 1990, but the rate has since considerably decelerated. However, despite the so-called "lost 20 years,"

[2] The growth rates of 1998-2001 were probably fabricated by the government (Rawski, 2001). China was severely hit by the 1997 Asian financial crisis and the economy went into deflation in the several years after. The economy did not resume high growth probably until China joined the World Trade Organization in December 2001. However, there were also under-reports in the economy, especially in the service sector where informal employment is prevalent. This has forced the NBS to revise China's GDP figures twice in recent years, in 2005 and in 2009. The 2005 revision increased China's 2004 GDP by 17%, most of which came from the service sector.

[3] China Association of Automobile Manufacturers at http://www.caam.org.cn/xiehuidongtai/20120112/1605066975.html, January 12, 2012.

[4] Japan was not a newcomer; its industrial foundation was laid down before the war. We include it because China's current position in the global economy bears many similarities with Japan's in the 1980s, and many commentators use today's Japan as a reference for China's future, especially when it comes to the role of long-term factors such as demography in determining China's future growth.

[5] The PWT reports two series of GDP data for China. One is close to China's official data and the other adjusts China's initial level of per-capita GDP and arrives in lower growth rates for subsequent years. Here we use the second series of data (China Version 2) which starts in 1952. Korean data starts in 1953.

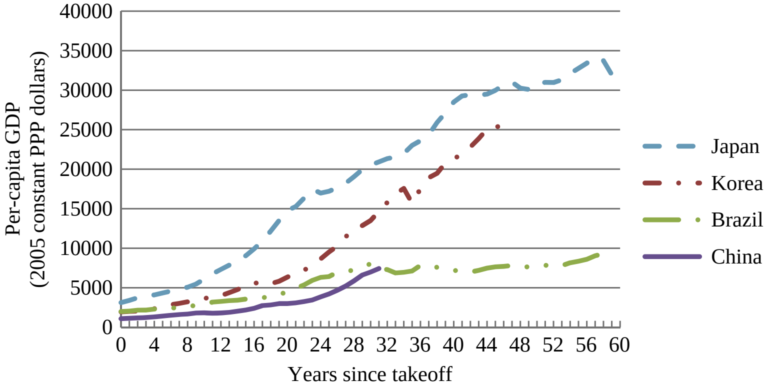


Figure 7.3 Growth since takeoff in four countries. *Source: PWT 7.0.*

Japan's per-capita income in internationally comparable terms has kept growing since 1990, primarily because Japan's domestic price level has almost kept unchanged while the price levels in other countries have increased. Brazil experienced high growth before 1980, but real income has declined since then and did not begin to grow again until 1995. In 2009, China's per-capita GDP was $7431 (2005 PPP constant prices), between Brazil's income levels of 1978 and 1979, between Korea's income levels of 1984 and 1985, or between Japan's income levels of 1962 and 1963.

Since the four countries began to experience high economic growth in different periods, a sensible approach is to normalize the comparison by the years since an economy started high economic growth. This is done in Figure 7.3. For Brazil and Japan, 1950 is chosen as the starting year, primarily because 1950 is the earliest year for which PWT provides data; for Korea and China, 1963 and 1978 are chosen as the respective starting years. The year of 1978 is chosen for China because China started the reform and opening policy in that year. The year 1963 was chosen for Korea because the Korean economy did not take off until after General Park Chung-hee got power through a military coup in 1962. Then we compare the average growth rates of the four countries in their first 30 years of fast growth. Between 1950 and 1980, Brazil achieved a remarkable average rate of growth of 7.8%, and Japan achieved 7.7%; between 1963 and 1993, the Korean economy grew by a marvelous rate of 8.7% per annum. As a comparison, China registered an average rate of growth of 7.8% between 1978 and 2008. That is, China has been a high performer, but certainly not better than other high performers.

It is tempting to extrapolate China's economic growth beyond 2010 using historical data. In this regard, the message delivered by Figure 7.3 is mixed. One sensible approach is to use the three other countries' growth records since they passed China's income level of 2009, to predict China's future growth. Then we have two extremes for the next 10 years since a country passed China's 2009 income level. On the one hand, Japan had an average

Table 7.2 Nominal GDP of China and the United States: 2009–2020

Year	US ($trillion)	China (¥trillion)	China ($trillion)	US/China	¥/$
2009	14.30	33.5	4.93	2.90	6.7
2010	14.84	39.8	5.88	2.52	6.6
2015	18.94	68.9	11.8	1.61	5.8
2020	24.17	119.3	23.68	1.02	5.0

Sources: For 2009 and 2010, data for the United States comes from the Bureau of Economic Analysis; data for China comes from NBS (www.stats.gov.cn). Figures for 2015 and 2020 are obtained under the assumptions made in the text.

growth rate of 9.6% between 1963 and 1973, and Korea had a rate of 9.0% between 1985 and 1995; on the other hand, Brazil's growth rate between 1979 and 1989 was only 3.5%. As a matter of fact, Brazil had many years of negative growth in the 1980s and 1990s and its average growth rate in the second 30 years, i.e. from 1980 to 2009, was only 2.3%. Clearly, Brazil, like many Latin American countries, fell into the middle-income trap in the 1980s and 1990s. Will China follow the track of Japan and Korea, or the track of Brazil?

Some studies cast optimistic predictions for China's future growth. The International Monetary Fund (IMF) predicted in its April 2011 World Economic Outlook that China would continue high growth rates and by 2016, would become the world's largest economy in PPP terms, taking 18% out of the world total.[6] Robert Feenstra believes that IMF has underestimated the size of China's real GDP because the prices it used were mostly from urban areas. Using the PWT data, he forecasts that China would become the world's largest economy by 2014 (Feenstra, 2011).

Realizing that the PPP figures are subject to difficulties in comparing the living costs across countries and across time, comparing countries by the current dollar is much easier, and in a sense provides more transparent figures. Keeping this in mind, Table 7.2 compares China and the United States for 2009 and 2010 and extrapolates the two countries' nominal GDP to 2015 and 2020 under the following assumptions: China grows by 8% per annum in real terms, the United States grows by 3% per annum in real terms; the two countries' inflation rates are 3.6% and 2%, respectively (averages between 2001 and 2010); and the yuan appreciates against the dollar by 3% per annum. Between 2009 and 2010, China narrowed its gap with the United States by a large margin, rising from barely about one third of the size of the United States to almost 40%. Under the above assumptions, the economy of the United States would be 61% larger than China's in 2015, and by 2020, the two economies would be almost the same size.

The assumption that China will grow by 8% per annum between 2011 and 2020 is a conservative estimation. The IMF have predicted that China's growth rates would exceed 9% till 2016. However, there are also many uncertainties about China's future growth. It might be incidental that Brazil fell into the middle-income trap right after it had

[6] http://www.imf.org/external/pubs/ft/weo/2011/01/weodata/index.aspx.

maintained high economic growth for 30 years, but China shared many characteristics with Brazil at that time, such as: an authoritarian government, high levels of inequality, and a large rural population. Therefore, it is not a sure thing that China could maintain high growth rates in the coming decade; the prediction presented in Table 7.2 is more indicative than reflective of inevitable outcomes.

7.2.2 Economic Transition and Growth

In the last 30 years, China has been both a developing country and a transition country. Starting in 1978, China began to move from a planning economy to a market economy. Compared with other transition countries, China's transition has not taken the toll of declining living standards; instead, it has maintained high growth rates while finishing most of the reforms. However, China today still bears some of the characteristics of a planning economy, noticeably, investment-driven growth, high shares of manufacturing in the national economy, a large sector of state-owned enterprises (SOEs), and a heavy presence of the government in the economy. On the other hand, China also shares many commonalties of other developing economies, especially its East Asian neighbors. Most significantly, it has followed its successful neighbors to adopt the ELG model. This subsection tackles the issues of transition and economic growth; the next two subsections deal with the issues related to structural change and the ELG model.

There have been many books and scholarly articles that provide excellent accounts of China's economic transition since 1978.[7] This subsection will not repeat those accounts, but will instead focus on the relationship between transition and economic growth, bearing in mind the question why China has managed high economic growth while moving from a command economy to a market economy. High growth is not granted when a command economy is transformed to a market economy; the experiences of the former Soviet Union and Eastern European countries have shown that economic decline could follow drastic transition. The question also bears ramifications for policy reforms in other developing countries. China's reforms did not follow a blueprint and can be best described by a meddling-through process. However, the direction of the reform was clearly toward a more market-based system, at least until the global financial crisis broke out in 2008. It would provide useful lessons for policy reforms in other developing countries once one understood how China has managed the direction right.

Before we get into the discussions, it is useful to briefly review where China stands today, after more than 30 years of reform. After finishing the rural reform that dismantled collective farming in 1984, China began to attack the two pillars of the command economy: state ownership and price controls. There were heated debates in the early 1980s on the sequence of reform. One school of thought, represented by Li, advocated a strategy to attack the first pillar, first based on the argument that a proper ownership structure is

[7] For a recent treatment, see Yao (2009) and other publications in the same series.

the foundation for a well-functioning market economy (Li, 2012). The opposite school of thought, represented by Wu, believed that ownership reform would be doomed to fail because the distortions in other parts of the command economy, particularly, the distortions in the price system, would provide wrong incentives to the enterprises (Wu, 2005). This school advocated simultaneous reforms in all fronts, putting specific emphasis on reforming the price system. The road that the reform actually took was a compromise. On ownership reform, contracts were introduced to incentivize SOE managers; on price reform, a dual-track price system was established in 1985. This system had two key components. On the one hand, prices of most consumer goods were liberalized; on the other, two tracks of prices were imposed on key inputs such as coal, oil, and steel, one still set by the government and the other set by the market. The government track was imposed on production and demand quotas while the market track was applied to outputs/purchases beyond the quotas. A producer of the key inputs could only sell its products in the market after it fulfilled its quota; likewise, an enterprise using those inputs had to buy them from the market once their demand quotas were used up. It is understandable that the market prices became much higher than the government prices and arbitrage would enrich those who had the privilege to get more quotas. However, the dual-track price system has avoided the hyperinflation that happened in the first several years after the big-bang reform in Russia and other transition countries.[8] Shortage was endemic in the command economy; a big-bang type liberalization would almost for sure cause hyperinflation. The dual-track price system dealt with shortage in two ways. By liberalizing the prices of consumer goods, it directly attacked shortage; by controlling the prices of key inputs, it slowed down the pace of price growth. By the early 1990s, the two tracks of prices converged, and finally the dual-track price system was confined to the annals of history by the unification of the official and market exchange rates in 1994. Immediately after that, massive privatization began following a 1995 government decision to only keep several hundred large SOEs in its hand. In the decade that followed, more than 90% of the SOEs were privatized. In the meantime, a vibrant private sector emerged and became dominant in urban employment (Figure 7.4); by 2008, more than two-thirds of urban workers (including migrant workers) were employed by the private sector. Accordingly, the shares of the SOE sector have dropped to less than 30% in the national GDP and corporate profits. China is best described as a mixed economy today.

However, there are also areas that have not been thoroughly reformed. Two of them have eminent impacts on the Chinese economy. One is that the financial sector is still tightly controlled by the government, and the other is that the government itself has not been transformed and has remained as a significant player in the economy. Many of the difficulties that the Chinese economy faces today can be traced back to those two unreformed areas. We will come back to them in Section 7.6.

[8] For a full account of the dual-track price system, see Chapter 9 in Yao (2009).

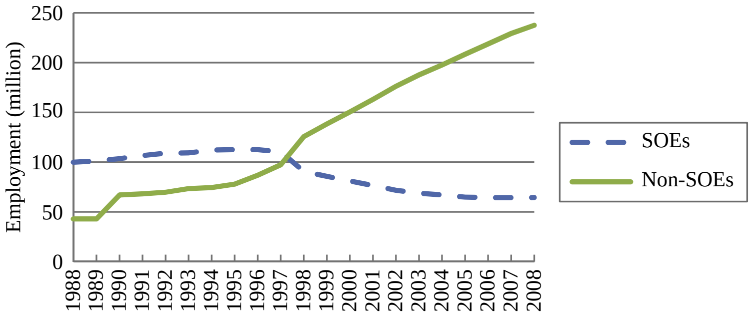


Figure 7.4 Urban employment: 1988–2008. *Source: China statistical yearbook of labor and social security, various years, the ministry of labor and social security.*

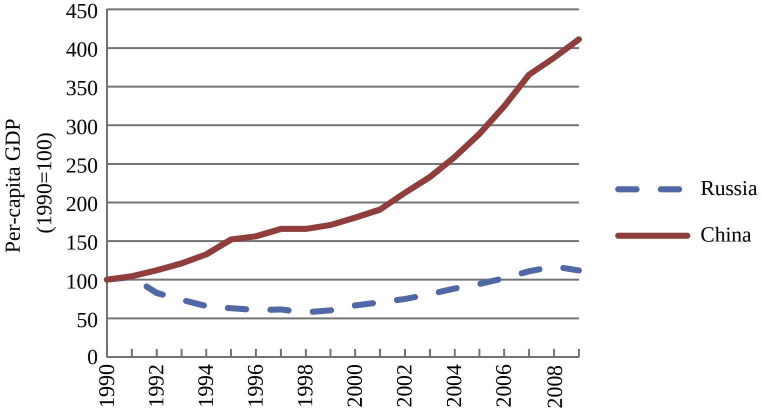


Figure 7.5 Economic performance of China and Russia in transition. *Source: PWT 7.0.*

In terms of the economic consequences, the Chinese transition has been far more a success than other transition countries. Figure 7.5 compares the economic growth in China and Russia between 1990 and 2009. In the first several years after the transition began in 1991, the living standard of the average Russian declined by nearly 50% and it was not until 2006 that the standard of living returned to its 1990 level. In contrast, the living standard of the average Chinese was quadrupled from 1990 to 2009. In 1990, the average Chinese had an income only 14% that of his or her Russian counterpart; in 2009, that had risen to 50%.

There are numerous studies in the literature that attempt to explain China and other transition countries' different records of economic performance in the process of transition. It is clear that the output drop in the former Soviet Union and Eastern European countries was the result of the disorganization caused by the big-bang reform (Blanchard and Kremer, 1997). The command economy has rigid albeit well-coordinated mechanisms to allocate resources among firms. The big-bang reform shattered those mechanisms and the market mechanism took time to reestablish. As a result, production slowed down, stopped, or even collapsed. However, several theories proposed in the literature show that structural and political factors made it impossible for the former Soviet Union and Eastern European countries to adopt a more gradual approach. Sachs and Woo (1994) argue that those countries were overindustrialized and created strong urban interests that resisted any gradual change to the state sector. In contrast, they believe that China had a weaker state sector which made reallocation of labor out of the sector possible. Qian et al. (2006a,b) explain China and Russia's different reform strategies using the contrast between the U-form organization in Russia and the M-form organization in China. The U-form organization in Russia delineated the management of the economy by line ministries that had to coordinate with each other to undertake a reform. This makes gradual and partial reforms impossible. In contrast, the M-form organization in China created relatively self-containing local units, so local experimentation was possible and minimized the costs of failed reforms. Boycko et al. (1997) emphasize the political imperatives that had driven the big-bang reform in Russia. Although the Communist government was gone in 1991, the old communist elites still controlled the economy and could come back with the resources they controlled. Massive privatization, thus, was believed by the liberal camp as a way to destroy the political base of the old communist elites.

While the above explanations are all well-founded, they sound too deterministic and have not paid enough attention to the human factors in the transition process. What if Gorbachev had opened the market and allowed people to participate in market transactions in the early 1980s? What if he had not started political reform but instead concentrated on economic reform? What if prices had not been liberalized overnight after the communist government fell apart? What if privatization had been conducted in a more orderly manner so disorganization could be avoided? To be sure, those questions have abstracted from the historical context; nevertheless, they are highly relevant for the policy reforms in other countries. For one thing, we do not expect that dramatic political changes happen often; policy reforms are, by nature, gradual in most countries. In this regard, the Chinese experience can provide several lessons. We will come back to these in Section 7.4 when we discuss the political-economy causes for China's economic success.

7.2.3 Uneven Structural Change

As a legacy of the command economy, the Chinese economy is heavily manufacturing centered, and as a result of that, labor movement from agriculture to the other sectors has been retarded. This is evident in Figures 7.6 and 7.7 that show the shares of the primary sector (agriculture and mining), secondary sector (manufacturing, construction, and transportation), and the tertiary sector (services) in national GDP and employment from 1952 to 2010, respectively.

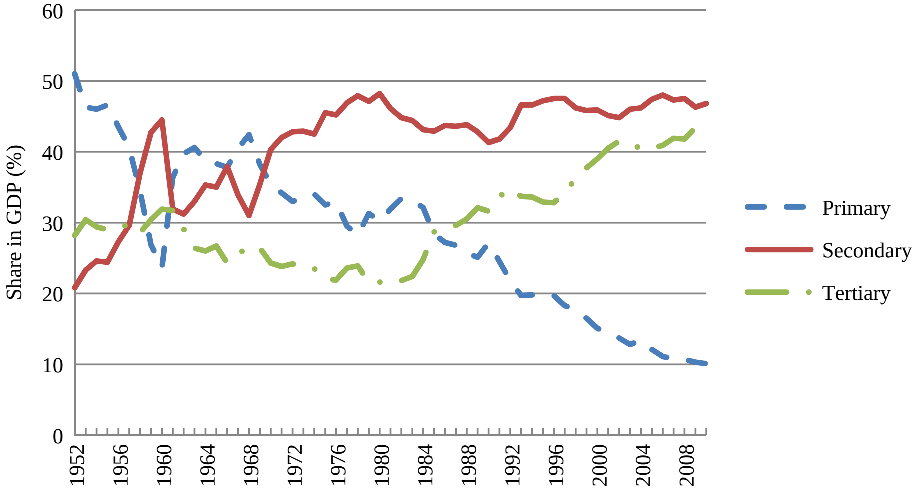


Figure 7.6 Shares of GDP of the three sectors. *Source: NBS at* www.stats.gov.cn.

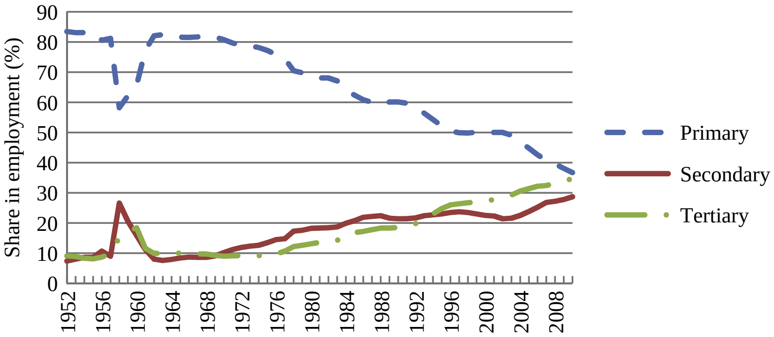


Figure 7.7 Shares of employment of the three sectors. *Source: NBS at* www.stats.gov.cn.

Except in the early 1980s, the GDP share of the primary sector declined over time, from 50% in 1952 to 10% in 2010.[9] The sector's share of employment has also declined, but with a slower pace. In 1952, it employed 84% of China's total labor force; by 2010, that number dropped only to 37%. This means that the productivity of the primary sector has declined relatively to the national average. In 1952, its relative productivity was 60%, i.e. 60% of the national average; in 2010, it declined to 27%. One of the causes for this

[9] There was a sudden drop in the share of the primary sector in both GDP and employment in the Great Leap Forward when 40 million workers moved from the countryside to the city. Half of them were sent back to the countryside after the Great Famine. In accordance, there was a surge of the GDP and employment shares of the secondary sector in the Great Leap Forward.

decline is that the household registration system, or the *hukou* system, has impeded labor movement from the countryside to the city so the countryside has been left with too many workers. However, the *hukou* has become much less a problem since 2003 when the Hu Jintao-Wen Jiabao government lifted most of the restrictions on labor movement.[10] Another cause is that there are a larger number of part-time farmers today than in the 1950s although part-time farming itself indicates that there have not been enough pulling forces from the other two sectors to draw farmers completely out of the countryside. This leads us to the third cause, namely, the dominant role of manufacturing in China's economic growth.

The share of the secondary sector in GDP rose before 1978; then it dropped in the 1980s, primarily because of the extraordinary growth in agriculture brought about by the rural reform. It rose again in the early 1990s and stabilized at around 47% in more recent years. The sector's share of employment has followed a somewhat different trajectory. It rose between 1962 and 1986, but more or less stabilized around 23% between 1986 and 2002. It even began a moderate decline in the later years of this period, signaling a sign that China would follow the conventional hump curve found for the manufacturing sector in other economies. However, this trend has been reversed since 2003 when the sector's share began to increase again. China's accession to the World Trade Organization (WTO) played a critical role for this reversed trend. Trade liberalization has lowered China's cost of trade by a large margin; as a result, China's comparative advantage in manufacturing is fully played out. It is worth noting that it was since 2003 that China began to harvest a large current account surplus. We will come back to this in Section 7.6.

The share of tertiary sector in GDP had declined until China entered the reform era although its share of employment began to increase in an earlier stage. The good news is that the sector has employed more people than the secondary sector since the mid-1990s. However, its share of GDP was still lower than that of the primary sector in 2010.

A comparison with other countries can give us a better understanding of where China's structural change stands today. Figure 7.8 shows the structural change of Korea. Two distinctions immediately emerge from the comparison of Figures 7.8 and 7.7. One is that the primary sector has been much larger in China than in Korea, and the other is that China's tertiary sector has been much smaller than in Korea. This is true even when the secondary sector hired about the same proportion of the labor force in the respective countries. The conclusion is that China has lagged behind in moving workers from agriculture to services. On the other hand, the share of the secondary sector in Korea began to decline in 1990. As we showed before, China's per-capita GDP in 2010 was equivalent to Korea's between 1984 and 1985. Using this as a reference, one may expect

[10] China's labor and population statistics have also been changed since then. Employment and resident status are now defined by the majority of time a person lived in a place in a year. If a person lives in a city for more than 180 days in a specific year, then he is counted as a resident in that city in that specific year. Accordingly, he is also counted as a worker in the urban sectors.

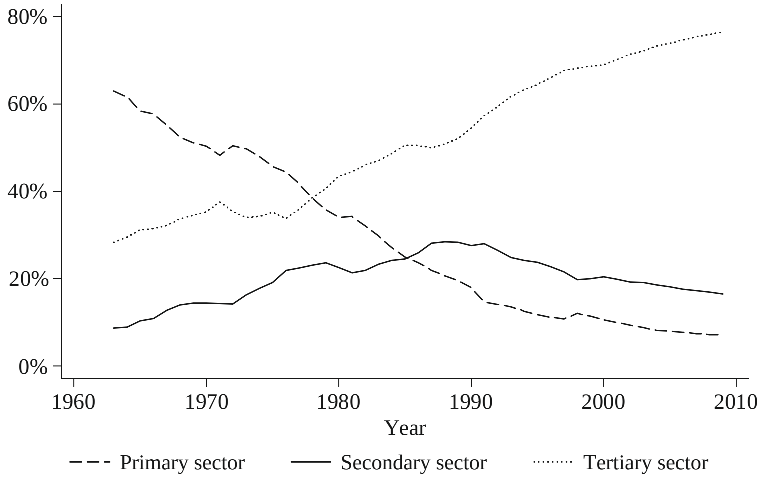


Figure 7.8 Employment shares in the Korean economy. *Source:* *Mao and Yao (forthcoming)*.

that the Chinese secondary sector would reach the turning point by 2015. However, whether this will happen critically depends on the ability of the service sector to absorb enough labor.

7.2.4 Export-led Growth

China began to adopt the ELG model in the early 1980s. At the end of the 1970s, when China began to reach out to the outside world, its leaders realized that the hail to China's achievement under the command economy was no more than self-glorification. In particular, China had been long taken over by its neighbors, including the Mainland's archrival, Taiwan. It was thus natural for China to adopt the model that had sent its neighbors to success. But unlike its neighbors, China is a large country, and its participation in the world system has different implications to the rest of the world. We will defer the discussion of those implications to Sections 7.5 and 7.6. Here, we would like to review some of the key features of China's ELG.

The first feature is that trade liberalization has been a key driver for China's export growth. As shown in Figure 7.9, there were clearly two periods in China's trade growth, one before and one after 2001, the year China joined the WTO. In the 10 years between 1991 and 2001, China's export grew by an average rate of 14.6% per annum; between 2002 and 2008, the rate increased to 27.3%. There was a large drop in 2009, but it was recovered in the next year. In 2010, China's export stood at five times what it was in 2001. Joining the WTO has moved China into a completely new trade regime in which its products are subject to uniformly low tariffs around the world. With its large reserves of labor and a sound industrial base, China could quickly tap into the benefits offered by this new regime.

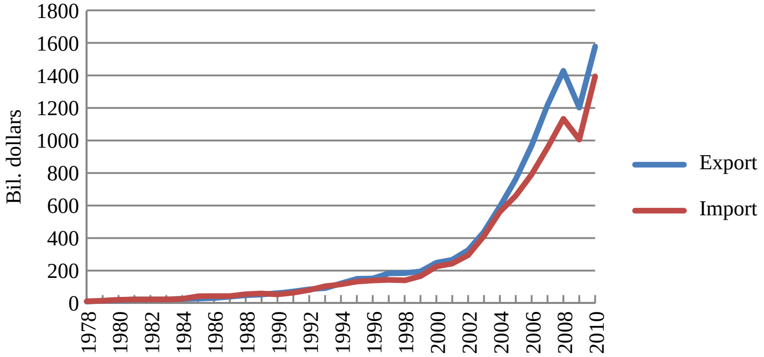


Figure 7.9 China's trade volumes (nominal dollars). *Source: NBS at www.stats.gov.cn.*

The second feature is that China's ELG did not lead to large trade imbalances until 2004 when China's trade surplus began to shoot up. As we showed in the previous subsection, there was a reindustrialization process starting around that time. This process is consistent with the explosive growth of export, but it is insufficient to explain why China should register huge trade surpluses. The financial crisis has slowed down China's export growth; as a result, China's surplus has also come down, from a peak of $295 billion in 2008 to $183 billion in 2010. It remains a question, though, whether this declining trend is cyclical or structural.

The third feature of China's ELG is that it relies heavily on processing trade. The share of processing export in total export reached 60% in the late 1990s, but has since dropped, and by 2010 it was barely above 50%. One of the salient features of processing trade is that by definition it creates trade surplus. As a matter of fact, China's entire trade surplus has been more than contributed to by processing trade since the early 1990s; that is, China's normal trade has been running deficits in all the years. China's processing export has very small per-unit value-added; the huge trade surplus created by processing export is mainly due to its large quantity.[11] With low value-added, it has often been questioned whether processing export is a sound strategy for China to hold on to.

Lastly, contrary to the conventional notion that China exports too much, China's export share in world total has been kept in line with its GDP share in world total, as evidently shown in Figure 7.10. In current dollars, China's share in world GDP declined before 1990; it was only 1.62% in that year. Its share of export has increased since data began to be recorded in 1970. By the mid-1990s, the export share caught up with the GDP share; both were about 2%. Between 2002 and 2008, the export share overtook the GDP share by an average of 1% point, reflecting the extraordinary growth of export

[11] An often-cited case is the iPhone. While one iPhone only costs Apple $6.5 to assemble in China, out of a total production cost of $178.96, China's total export of iPhones to the United States contributed $1.9 billion to the US's trade deficit with China in 2009 (Xing and Detert, 2010).

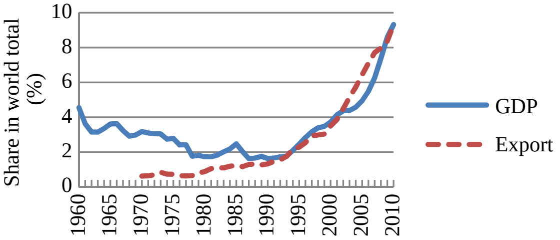


Figure 7.10 China's shares in world GDP and export. *Notes:* The two series are calculated using current nominal dollars. *Source: WDI at http://data.worldbank.org/data-catalog/world-development-indicators.*

in this period. In 2009 and 2010, though, the two shares were almost the same; in 2010, both were 9.3% of their respective world total. The correlation coefficient between the logarithms of the two series of shares for the period 1990 to 2010 was 0.98. Therefore, in a gravity model, China's share of world export can be almost perfectly explained by its share of world GDP. That is, China's export is nothing abnormal in terms of standard trade empirics.[12] The extraordinary growth of export between 2002 and 2008, though, may be the result of the one-shot improvement of China's trade regime.

What is the contribution of export to the Chinese economy? Due to its heavy reliance on processing export, export's contribution to China's overall GDP, including net export and the value-added created by backward and forward linkages, is very low, perhaps in the range of 10–12% (Lau et al. 2007). However, because export has been growing faster than GDP, the contribution of its growth to GDP growth is large. For example, in the period 2002–2008, export grew by an average of 27.3%, so its contribution to GDP growth was 2.73–3.28% points. That is, about 30% of China's GDP growth in that time period can be attributed to the growth of export.

In summary, China's growth experience since 1950 can be viewed from both a historical and a contemporary perspective. From the historical point of view, China's economic ascent can be seen as regression to its position in the world in the early 1800s, albeit by imitating the industrialization process initiated by the West. In this process, China has benefited from the advantage of backwardness that Morris (2011) uses as one of the structural factors to explain the ups and downs in the East and West in the last 15 millenniums. From the contemporary point of view, however, the advantage of backwardness does not guarantee fast catch-up; after all, catch-up has been the exception rather than the rule after World War II. The fates of human societies may be governed by some hidden rules in the long run; yet in the short run, human decisions play a more significant role in determining whether a society moves in the direction of ascent or toward the abyss of decline. In the next two sections, we will provide a review of selected theories and empirical evidence that have tried to explain why China has managed its economic ascent.

[12] In a standard gravity model, both GDP and export are in real terms. Since China's inflation rates have been in line with the world average, using real GDP and export should not change the result very much.

7.3. THE ECONOMICS OF THE MIRACLE

The Chinese growth miracle can be explained in two ways, firstly, using more conventional economic wisdom that attempts to answer the question: what has China done right?; and the other, resorting to political-economy factors to answer the question, why has China done right? This section focuses on the first explanation, and the next section focuses on the second. We will see in this section that what China has done right is mostly consistent with the neoclassical doctrines. In particular, China had more favorable initial conditions than, say, India, when its economy began to take off in 1978; China has maintained high saving rates and investment rates along the way of economic growth without compromising technological progress; it has also made significant improvements to human capital; lastly, the government has adopted a prudent fiscal policy and has maintained macroeconomic stability most of the time. Next, we look at China's initial conditions.

7.3.1 The Initial Conditions

In the classical Solow model, the steady state of an economy has nothing to do with its starting point. This is quite different from the models featuring technical or economy-wide non-convexities. These non-convexities often lead to multiple steady states with very different outcomes; depending on its starting point, an economy can reach different steady states. In empirical research, however, this theoretical distinction may not sound that important. In the Solow model, factors determining an economy's steady state, such as the saving rate, population growth rate, etc. are assumed constant over time. In this sense, they are part of an economy's initial conditions.

To begin our discussion of China's initial conditions in 1978, it would be helpful to put the country in an international context. Table7.3 then compares China and India around 1978. The first thing one notices is that China was a poorer country than India at the time. However, in terms of other human development indicators, China did a much

Table 7.3 Comparisons of China and India in 1978

	China	India
Per-capita GDP (constant $2000)	155	206
Adult literacy rate (%)	65.5	40.8
Tertiary school enrollment (% gross)	0.7	4.9
Life expectancy	66	54
Infant mortality rate (‰)	54.2	106.4
Share of manufacturing in GDP (%)	40.0	17.0
Share of manufacturing in employment (%)	17.3	13.0

Notes: China's literacy rate is for 1982 and India's literary rate is for 1981.
Sources: WDI at http://data.worldbank.org/data-catalog/world-development-indicators. India's share of manufacturing employment is from Valli and Saccone (2008), Table 3.

better job than India—China had a much higher adult literacy rate, a much longer life expectancy, and a much lower infant mortality rate. India scored higher than China only in tertiary school enrollment. Within the age group that was officially defined for tertiary education, 4.9% were enrolled for tertiary education in India whereas the figure was only 0.7% in China. In fact, it needed to wait until the early 2000s for China to catch up with India in this indicator. That is, China's approach to human development was targeted on ordinary people and people's basic needs, while India's approach was more elitist.

Some authors attribute China's better human development records, especially its relatively high levels of human capital achievement, to its historical and cultural roots. For example, Rawski (2011) believes that China's economic ascent is a consequence of its long-term accumulation of human resources in historical times before 1949. In particular, he emphasizes the roles of family farming, commercialization, closely knitted social organizations, and cultural beliefs in fostering human capital accumulation in China's historical times. This argument falls generally in line with the historian Kaoru Sugihara's notion of "industrious revolution" which he uses to describe the mechanism behind East Asia's long-term economic growth (Sugihara, 2003). In contrast with the West's Industrial Revolution that expanded economic production beyond human capacities, the East has undergone an industrious revolution that intensively explored human capacities for further economic growth. Sugihara emphasizes the limited natural resources and resulted small family farming as the most important cause for East Asia to undertake the industrious revolution instead of the Industrial Revolution.

While the thesis advanced by Sugihara and Rawski has much merit to recommend, it is also worth keeping in mind that the quality of human resources was much improved in the first 30 years of the new China although human destruction, manifested in particular by the Great Famine of 1959–1961 and the Cultural Revolution between 1966 and 1969, also happened. Only half of the eligible children went to elementary school in 1952; by 1978, 98% of them did.[13] Half of the adults were reported literate in the second census conducted in 1964; the ratio was increased to two-thirds by 1978, as Table 7.3 shows.

In addition to improvement of human development, a second favorable condition for China's economic take-off was equality. A thorough land reform and subsequent collectivization in the 1950s had equalized landholdings among farmers. The rural reform of 1978–1984 restored family farming, but in the meantime also institutionalized equal land distribution. In the city, the low-wage policy also considerably shrank income inequality. As a result, China was one of the most equalized countries in terms of income distribution in 1978. Socially, the Communist Revolution leveled out the Chinese society. Although there were political barriers preventing vertical mobility (the *hukou* system was one of the most notorious), the rural gentry class and urban capital owners were completely eradicated. There was no strong social or political group in the society except the Communist Party. We will see in Section 7.4 that equality can be one of the major reasons why the

[13] *Statistical Summaries: 60 Years of the New China.* Beijing: China Statistical Press, January 2010.

Chinese government has been made free-to-adopt growth-enhancing economic policies in the reform era.

A third favorable condition was a sound industrial base China had established in the first 30 years. Table 7.3 shows that the manufacturing sector in China was much larger than that in India. In terms of share in GDP, China's manufacturing sector took 40% whereas the figure was only 17% in India; in terms of share of employment, the gap was smaller, but still substantial with China's being 17.3% and India's being 13%.[14] The Chinese manufacturing sector was also relatively more productive than the Indian manufacturing sector. Labor productivity in the Chinese manufacturing sector was 2.3 times of the Chinese national average, whereas India's was only 1.3 times of the Indian national average.

China and India both had an old civilization and both achieved high levels of prosperity in historical times. India got independence in 1947, and China ended its half-century long internal turmoil in 1949. So how had China managed to achieve a generally better record of human development and a larger manufacturing sector than India by 1978? Table 7.3 has already hinted at the answer: China had suppressed people's income to speed up industrialization and the improvement of other human development indicators. This is no more evident in its pursuit of the heavy-industry development strategy (HIDS).

China was basically an agrarian country in the early 1950s. In 1952, industry accounted for 20% of the national GDP and only hired 6% of the country's total labor force (Lin et al. 2003).[15] Modeled on and aided by the Soviet Union, China began a drive of fast industrialization through the HIDS. The Chinese government adopted several measures to accelerate capital accumulation in the country. First, farmers were organized into communes and had to sell their products to the state under suppressed prices. It is estimated that over ¥200 billion were extracted from farmers during the period 1958–1978 (Wu, 2001). Second, urban wages were set to very low levels and to solve the problem of shortage, rationing was prevalent. Third, interest rates were set to under 5% per annum to save the costs of HIDS. Fourth, the value of the yuan was set high to reduce the costs of imports of technology and equipment. Fifth, high tariffs were instituted to protect domestic industry and to generate government revenue. Lastly, heavy industries received disproportionally large amounts of investment. Figure 7.11 shows the ratio of investment between heavy industries and light industries in the period 1952–1990. Using the average of the 1980s as a benchmark for a reasonable ratio, it is clear that the heavy industrial sector was overinvested in the period of command economy, especially in the 1960s.

[14] Manufacturing's share in employment is relatively small in all countries, with the highest less than 30% in most countries (see for example Korea in Section 7.2.4).

[15] China's industrial share of GDP in 1952 was higher than India's in 1978. So India's lower achievement in 1978 could be a result of its lower starting point in the 1950s. However, the share of manufacturing in India in the 2000s was about 30%, lower than what China had achieved by the end of 1970s.

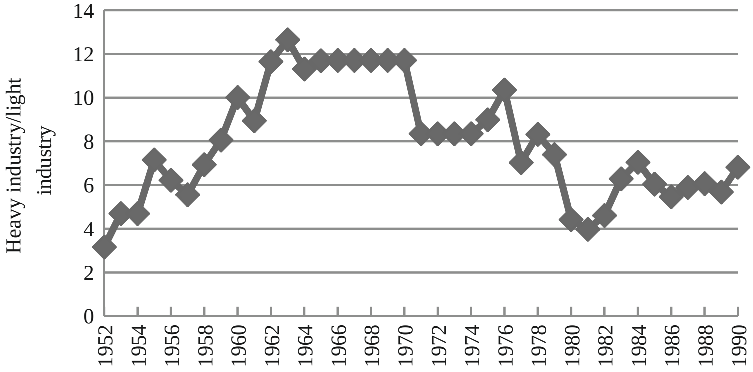


Figure 7.11 The investment ratio between heavy industry and light industry: 1952–1990. *Source: Yao and Zheng (2007)*.

In general, the above policies had done more harm than good to the Chinese economy.[16] However, the HIDS had also transformed China from an agrarian economy to one on the way to industrialization. This explains the differences between China and India in industrial development at the end of the 1970s. It is worth noting that industrial development is not just about the accumulation of physical capital, it is also about the accumulation of human capital. Although the physical capital accumulated in the era of command economy had soon become obsolete in the reform era—many companies established under the command economy went bankrupt or were sold to private companies in the 1990s—technicians and skilled workers trained in the command economy had played a critical role in the initial stage of the private sector development.

The Chinese experience fits into Gerschenkron (1962)'s theory of economic backwardness. To Gerschenkron, a backward economy can skip some of the development stages experienced by more advanced economies by adopting their advanced technologies and practices. In doing so, it would suppress consumption, invest more heavily, and rely more on state entities. To the extent that the accumulation of physical and human capital in the first 30 years laid a foundation for China's take-off in the second 30 years, the first phase was indispensable for China.

7.3.2 Savings, Investment, and Productivity Growth

It is well known that China's economic growth has relied heavily on investment. It is worth keeping in mind, though, that the investment rates—defined as the share of

[16] In a dynamic general equilibrium model featuring heavy industry's technical and pecuniary externalities, Yao and Zheng (2008) calibrate the optimal rate and length of capital subsidies to heavy industry for China's command economy. They also calibrate the actual rate of subsidies and find that it is 37.6%, 6.6% points higher than the optimal level. In addition, the HIDS was implemented for 25 years, 13 years longer than the optimal length.

capital formation in GDP—were already very high before 1978. They were 24.5%, 30.8%, 26.3%, and 33.0%, respectively, in the first to the fourth five-year plans.[17] This required remarkable savings for a country with very low per-capita income. It was made possible by heavy government extraction through the command system. As a comparison, the investment rates of India—a then richer country than China—were 15–18% in the 1960s and 1970s (Cheten et al. 2006). Clearly, China fitted into Gerschenkron's theory of backwardness more than India did. Moving into the reform era, China's high saving rates have continued and have increased dramatically since 2000. As it will be shown in Section 7.6, the national saving rate was increased to 52% by 2010. This high rate of savings was almost unprecedented in history (except in some countries during wartime periods and in a few years of the Soviet Union).

High savings have inevitably led to high investment. China relied heavily on investment for growth under the command economy; it remains as an investment-driven economy in present days; its investment rates reached 50% of GDP after 2009. While China's high saving rates have become an international topic in recent years, the sustainability of its investment-driven growth has already been called into question since the mid-1990s. For example, Kim and Lau (1996) have estimated the Solow residuals for East Asian economies for the period 1980–1995 and found that the contribution of technological progress, measured by the Solow residual, is negative in many economies. In particular, they have found that capital growth on average contributes to 96% of China's economic growth whereas technological progress contributes to −1.4%. Even in Taiwan and Korea whose exports have far more technological contents, technological progress is found to have only contributed 2.5% and 1.0% to the two economies' growth, respectively. Those results were often used to question the sustainability of the East Asian economies, and the subsequent Asian Financial Crisis seemed to have verified the doubt.

However, depending on the stage of economic development, an economy can grow without technological progress. In the framework of the Solow model, the growth rate of per-capita income is solely determined by the rate of technological progress when an economy reaches its steady state; however, growth can be driven purely by capital accumulation before an economy reaches that state. It is readily admitted that most developing economies have not reached their steady states; therefore, capital accumulation can be an important driver for economic growth.

In addition, more recent studies have found that technological progress is one of the growth drivers for the Chinese economy.[18] Even Alwyn Young, a sarcastic critique of China's official statistics, finds a positive growth rate for China's non-agricultural total factor productivity (TFP) in the period 1978–1998. After accounting for growth of labor

[17] The time spans of these four five-year plans were 1953–57, 1958–62, 1966–70, and 1971–75. The interim period of 1963–65 was a period of adjustment to the Great Famine. The investment rate in this period was 22.7% (Lin et al. 2003).

[18] For a brief review of the literature, see Zheng et al. (2009).

Table 7.4 TFP growth in 1978–2005 (%)

Growth rates	1978–1995	1995–2005
GDP	10.11	9.25
Capital	9.19	12.38
Labor	3.6	2.59
$\text{TFP}^{(0.5)}$	3.72	1.77
$\text{TFP}^{(0.4)}$	4.27	2.74

Notes: All data come from official sources. Labor has been adjusted by quality. $\text{TFP}^{(0.5)}$ assumes that capital's share is 0.5, and $\text{TFP}^{(0.4)}$ assumes that capital share is 0.4.
Source: Zheng et al. (2009), Table 1.

(largely due to increased labor force participation), the shift of labor out of agriculture, and rising educational levels, he finds that non-agricultural labor productivity grows by 2.6% and TFP grows by 1.4% per year. Table 7.4, adopted from Zheng et al. (2009), presents another set of calculation for China's TFP in the period 1978–2005. Between 1978 and 1995, TFP growth accounts for 42% of GDP growth; the figure declines to 30% between 1995 and 2005. This decline was largely caused by the accelerated growth of capital. As we pointed out earlier in this section, the Chinese economy experienced another wave of heavy-industry development in the early 2000s. Its effects may not be fully evident in a short period of time. Therefore, a study of more recent data is required to settle the issue.

It is noteworthy that none of the existing methods assessing technological progress has successfully accounted for the technological progress embedded in capital. It is obvious that a factory is not just adding capital to its pile of stock when it buys a new assembly line—it requires reengineering of its production process and a whole new team of workers, which are definitely part of the story of productivity growth. In this regard, it would be more illuminating to directly look at labor productivity that takes into account both TFP growth and the contribution of capital accumulation. Figure 7.12 presents data for the accumulative growth of labor productivity in both the manufacturing and service sectors for the period 1978–2009. Before 1990, the growth was minimal in both sectors. The growth since the early 1990s has been phenomenal. In the manufacturing sector, the output of one worker in 2009 was equivalent to the output of 12 workers in 1978. The growth in the service sector was less dramatic, but a worker in 2008 was still equivalent to more than four workers in 1978.

In addition to fast capital accumulation, China is experiencing fast structural change; the productivity gains from factor movements from low productivity sectors to high productivity sectors cannot be underestimated. Table 7.5, adopted from Brandt et al. (2008), presents the results of a simple growth decomposition between agriculture and non-agriculture for the period 1978–2004. In this whole period, labor reallocation from agriculture to non-agriculture contributed 24.6% to China's overall growth. In the first

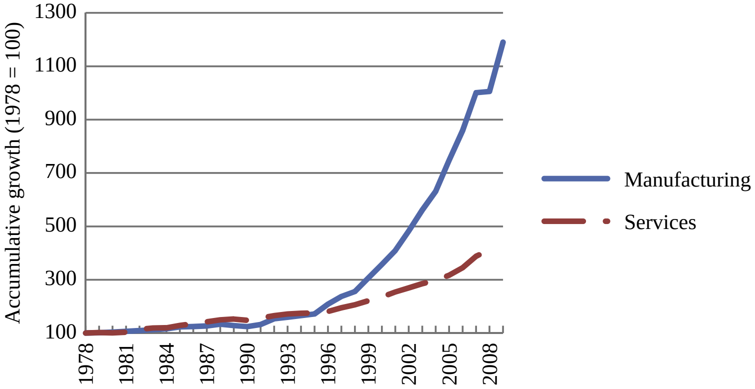


Figure 7.12 Growth of labor productivity in manufacturing and service sectors. *Source: Data before 2005 come from Lu and Liu (2007). Data after 2005 are provided by Feng Lu.*

Table 7.5 Growth decomposition: agriculture versus non-agriculture

	1978–2004	1978–1988	1988–2004
Aggregate	100.0	100.0	100.0
Output per worker in agriculture	27.3	23.8	19.4
Output per worker in non-agriculture	48.1	26.4	69.2
Reallocation	24.6	49.9	11.4

Source: Brandt et al. (2008), Table 17.2.

sub-period, 1978–1988, the contribution was much higher, reaching 50% of the country's overall growth. It decreased to 11.4% in the second sub-period, 1988–2004. The high contribution in the first sub-period was mainly caused by the extraordinary growth of rural industry in that period. The gap of productivity between non-agriculture and agriculture was much larger in the second sub-period than in the first sub-period, but the rate of labor movement was significantly lower in the second than in the first sub-period.

7.3.3 Human Capital Formation

As we showed in Section 7.2, one of China's achievements by 1978 was relatively higher levels of human development, including primary education. In the reform era, China has continued to improve its stock of human capital due to continuous government commitment and increased returns to education. In 1993, the government set the goal in its National Plan of Educational Reform and Development to increase government educational spending to 4% of the national GDP by the end of the 20th century. This target was missed at the turn of the century. In the new National Plan of Educational Reform and Development: 2010–2020, approved in 2010, the government pledged again to meet

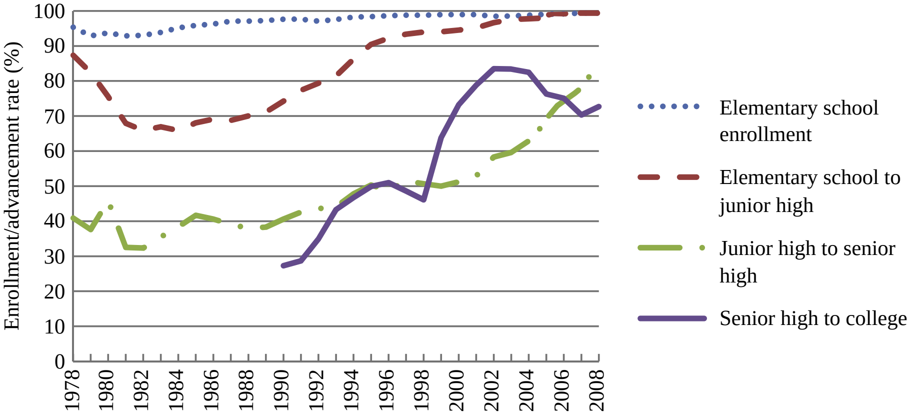


Figure 7.13 Enrollment rates and advancement rates: 1978–2008. *Notes:* In China, elementary schools have 6 years, and junior and senior high schools each have 3 years. There are two types of senior high schools, vocational and academic. Elementary school enrollment rate is defined as the ratio between the number of students in elementary schools and the number of children between 6 and 12 years old. Elementary school advancement rate is defined as the ratio between the number of students advancing to junior high schools and the number of elementary school graduates. Junior high school advancement rate is defined in the same way. High school advancement rate only accounts for students graduated from academic high schools. *Source: China Educational Statistical Yearbook 2009. Beijing: Renmin Education Press, November 2010.*

the target by 2012. It is noteworthy, though, that as a share of total government spending, China's government education spending is not low. In 2008, it was 13.8%, lower than the level in the United States, but higher than those in the United Kingdom, Japan, and the Nordic countries (Bai et al. 2010). As for the private returns to education, most studies (e.g. Zhu, 2011; Li et al. 2012) show that one more year of schooling on average increases a person's income by 8–9%. In particular, using their twins data set, Li et al. (2012) find that the return to vocational high school education is between 19.6 and 21.9%, the return to vocational college education is between 21.5 and 23.0%, and the return to college education is between 35.7 and 40.0%. That is, the return to college education is about 10% for one additional year in school, similar to those found for the US and Europe.[19]

Official statistics show that school enrollment and advancement rates have increased over the years. Figure 7.13 presents data for elementary school enrollment rates and the advancement rates of each level of school to the next level for the period 1978–2008.[20] While the general trend for the four indicators was improvement, there were also fluctuations. The most significant was the decline of the advancement rate of senior

[19] In contrast, they find that the return to academic high schools is only between 4.0 and 5.4%. That is, it is between 1.3 and 1.8% for one additional year in school (China's high schools have 3 years).

[20] Note that one should not infer the junior high school enrollment rate from the elementary school enrollment rate and its advancement rate because there are drop-outs in junior high schools.

Table 7.6 Higher education and R & D personnel and spending: 1991–2008 (1000)

	Undergraduates			Graduates			R & D		
Year	Enrollment	Admission	Graduation	Enrollment	Admission	Graduation	R& D activity personnel	Scientists and engineers	R& D spending in GDP (%)
1991	2044.0	620.0	614.0	88.1	29.7	32.5	228,600	132,100	
1992	2184.0	754.0	604.0	94.2	33.4	25.7	227,000	137,200	
1993	2536.0	924.0	571.0	106.8	42.1	28.2	245,200	137,200	
1994	2799.0	900.0	637.0	127.9	50.9	28.0	257,600	153,900	
1995	2906.0	926.0	805.0	145.4	51.1	31.9	262,500	155,400	0.6
1996	3021.0	966.0	839.0	163.3	59.4	39.7	290,300	168,800	0.6
1997	3174.0	1000.0	829.0	176.4	63.7	46.5	288,600	166,800	0.6
1998	3409.0	1084.0	830.0	198.9	72.5	47.1	281,400	149,000	0.7
1999	4134.0	1597.0	847.6	233.5	92.2	54.7	290,600	159,500	0.8
2000	5560.9	2206.1	949.8	301.2	128.5	58.8	322,400	204,600	1.00
2001	7190.7	2682.8	1036.3	393.3	165.2	67.8	314,100	207,200	1.07
2002	9033.6	3205.0	1337.3	501.0	202.6	80.8	322,200	217,200	1.23
2003	11,086.0	3822.0	1877.0	651.3	268.9	111.1	328,400	225,500	1.31
2004	13,335.0	4473.4	2391.2	819.9	326.3	150.8	348,200	225,200	1.44
2005	15,617.8	5044.6	3068.0	978.6	364.8	189.7	381,475	256,063	1.32
2006	17,388.0	5461.0	3775.0	1104.7	397.9	255.9	413,200	279,800	1.39
2007	18,849.0	5659.2	4477.9	1195.0	418.6	311.8	454,400	312,900	1.40
2008	20,210.2	6076.6	5119.5	1283.0	446.4	344.8	496,700	343,500	1.47

Sources: Data other than R& D spending come from *Statistical Summaries: 60 Years of the New China* (Beijing: China Statistical Press, January 2010). Data for R& D spending come from *China Statistical Yearbook*, 2000, 2005, 2010, and *China Science Technological Statistical Yearbook 2010* (Beijing: China Statistical Press, October 2010).

academic high schools to colleges. In the early 2000s, more than 80% of senior academic high school graduates went on to college, but the rate dropped to barely above 70% in 2008. There was a large wave of college expansion in the early 2000s (Table 7.6). One consequence of this expansion was that the starting salaries of college graduates were suppressed.[21] This may explain the drop of high school advancement in the period. The effect might be larger for rural students because the cost of college education is relatively much more substantial to them than to their urban peers.

Figure 7.13 shows that most children have advanced to junior high level in recent years. Figure 7.14 then shows the enrollment rates of high schools and colleges for the period 1992–2009. The high-school enrollment rate had a dramatic increase since 2003 after the stagnation in the late 1990s, reaching 80% by 2009. The college enrollment rate has been increasing steadily and reached 22% in 2009. The government projected in its National Plan of Educational Reform and Development: 2010–2020 that the college enrollment rate would grow to 40% by 2020. Most of the growth, though, would be contributed by slower growth of population.

Compared with primary and secondary education, China's higher education has been advancing by a much faster pace. Table 7.6 provides data for college and graduate school admission, enrollment, and graduation for the period 1991–2008. In the period, college admission and enrollment increased by a factor of 10-fold; graduate school admission and enrollment were increased by a faster pace of 15-fold. In 2008, Chinese universities produced 5 million bachelors and 344,000 masters and PhDs. This fast growth of high

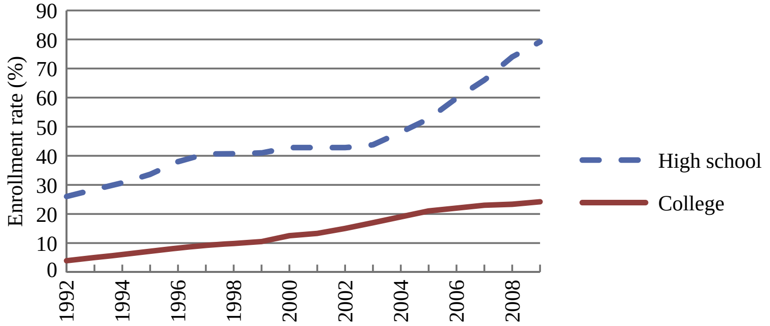


Figure 7.14 High school and college enrollment rates: 1992–2009. *Notes:* The high school enrollment rate is defined as the ratio between the number of registered high school (both vocational and academic) students and the population between 15 and 17 years old; the college enrollment rate is defined as the ratio between the number of registered college students and the population between 18 and 22 years old. *Source: China educational statistical yearbook 2009. Beijing: Renmin Education Press, November 2010.*

[21] Wu and Zhao (2010) use two waves of surveys (2002 and 2005) to find that the college expansion suppresses the starting salaries of college graduates by 10.5%.

education has been a strong source for the growth of China's R&D personnel, which more than doubled in the period (columns 7 and 8 in Table 7.6). Accordingly, R&D spending as share of GDP was increased from 0.6% in 1995 to 1.47% in 2008. The government's 12th five-year plan has set the goal to increase the share to 2.2% by 2015. This will then raise China to the rank of developed countries.

The true challenge facing China is how to increase the human capital of the 140 million migrant workers whose educational achievements are mostly at or below junior high. They are the bulk of China's workforce. Most of them will not drop back into full-time schooling; on-job training and part-time schooling are the only choices to increase their human capital. However, the government has not paid enough attention to them. According to the National Plan of Educational Reform and Development: 2010–2020, most government educational resources will be devoted to strengthening formal education. This will prepare China for the years beyond 2020, but will basically ignore the current labor force. Not only is it wasteful, but also it entails risks for China's ambition to upgrade its technological capacities in the next 10 years.

7.3.4 Macroeconomic Stability

One of the regularities coming out of the empirical literature of economic growth is that macroeconomic instability is detrimental to economic growth. This was why John Williamson put macroeconomic stability as the first of the ten policy recommendations that he summarized as the Washington Consensus for the restructuring of the Latin American economies after the deadly sovereign debt crisis in the 1980s (Williamson, 1990). Compared with other developing countries, China has done a good job in maintaining a stable macroeconomic environment. This has been a rare achievement if one also considers the fact that China has more or less finished the transition from the command economy to a market economy, a process that has uniformly caused hyperinflation in the other transition countries. There have been several rounds of business cycles since 1978; some of them have led to high inflation rates by the standard of the recent Chinese history, but most of them were mild compared with other developing countries.

Figure 7.15 presents China's inflation rates between 1978 and 2010. Four periods of higher inflation rates can be identified in the figure: the early years of reform, 1988–1989, 1992–1995, and 2007–2008. The first two periods of inflation were caused by attempts to reform prices. Despite the dual-track price system, the price level still increased in the 1980s. The inflation rate reached 18% in 1988 and caused widespread panic among the population. It was also one of the driving forces behind the 1989 student movement. After economic reform was resumed in 1992,[22] an investment frenzy began in the country and led to a sharp surge of the price level. In 1994, the inflation rate reached 24%, the highest in the People's Republic history. Then it declined very fast and after the Asian

[22] The reform was stalled after the 1989 student movement. In the spring of 1992, China's paramount leader Deng Xiaoping paid a visit to the south and called the party to resume reform.

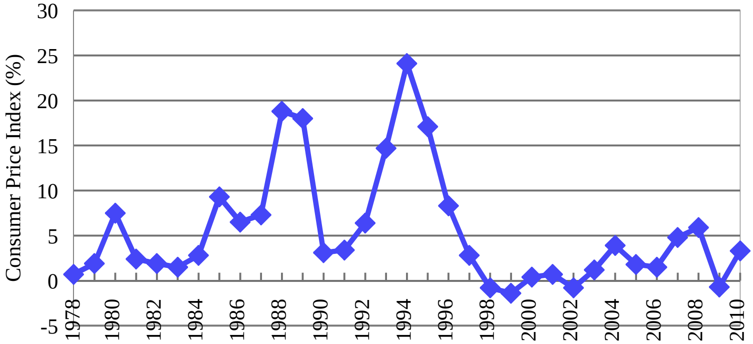


Figure 7.15 China's consumer price indexes: 1978–2010. *Source: NBS at* www.stats.gov.cn.

Financial Crisis became negative. Deflation continued until 2003 when China entered a new round of price growth. This time it was mainly caused by China's large current account surpluses. However, compared with the previous three waves of price growth, this wave has been much milder, indicating more sophisticated macroeconomic management by the authorities.

China's macroeconomic stability has been helped by a generally fiscally prudent government at the central level. Figure 7.16 presents data for the central government's deficits in the period 1995–2010 and its debts in the period 2005–2010.[23] The central government incurred the highest deficit rates around 2000. In that year, 45% of its expenditure was

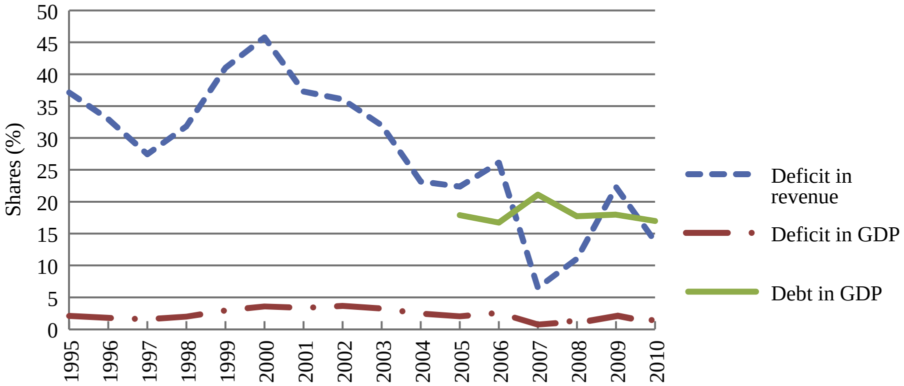


Figure 7.16 Central government deficits and debts. *Notes:* Deficits are calculated as the difference between the central government's expenditure (its own spending and its transfer to local governments) and its revenue. Debts are the accumulative net debts at the year end, including both domestic and foreign debts. *Sources: NBS, China statistical yearbook, various years.*

[23] The Ministry of Finance began to release data on the accumulated net debt in 2005. Before that year, only data on new debts and repayments were released. Because the data for earlier years are incomplete, it is not possible to convert data before 2005 to those consistent with the data reported since that year.

financed by debt. This was a result of the government's response to the Asian Financial Crisis. There was a large wave of infrastructural building after the crisis; most of China's highways were built in that time period. The deficit-revenue ratio declined substantially to 6.5% in 2007. Then the global financial crisis led to another wave of infrastructural building although it was much milder than the first wave, where government expenditure is concerned. When we put the deficit against the national GDP, we find that the deficit-GDP ratio was substantial in the first wave but then dropped to around or under 2% since 2007. On the other hand, the debt-GDP ratio was between 17 and 18% since 2005, except in 2007.

Local governments' fiscal situation has been more troublesome. One of the problems is that no systematic data exist to gauge the size of local deficits and debts. The Budget Law requires that local governments balance their budgets. But local governments can tap into China's weak financial system through government-sponsored financial companies. Despite the reform in the banking sector, many banks, particularly the city commercial banks, still lend local governments' soft loans because they do not want to offend their powerful local hosts. In addition, local governments have large amounts of assets, especially land, in their hand to collateral their borrowings, so banks often believe that it is safe to lend to local governments. By the end of 2010, the outlet of local government debts was ¥10.7 trillion, or 26% of GDP (Wu, 2012). Seventy-nine percent of the debts were loans from commercial banks. In light of the weak fiscal discipline of local governments, many people are worried about the risks of large bad loans coming from local government debts. However, there are reasons to be more optimistic than what China experienced at the turn of the century.[24] For one thing, local government debts are mostly collateralized by land and infrastructure, so banks could get most of their values back in case their loans were defaulted.

In summary, there is no secret to China's economic success from a purely economic perspective because it has adopted the right policies frequently recommended by neoclassical theory and empirics. This conclusion has a strong bearing on the debate surrounding the so-called China Model.[25] The review in this section has shown that at least on the economic front, China has not created a new growth model; rather, it has converged with the common model recommended by economists for developing countries

[24] Nonperforming loans were 25% of GDP at the time (Lardy, 1998), most of which were accumulated by local governments and SOEs over the reform period.

[25] In popular and policy spheres, the debate is often framed in the contrast between the Beijing Consensus and the Washington Consensus. The first consensus is believed to feature state capitalism and an authoritarian state, and the second consensus is believed to feature free market and a democratic state (e.g. Bremmer, 2011). However, both are quite different from their original formulations. The Beijing Consensus was proposed by Joshua Ramo in 2004 (Ramo, 2004), mostly to describe how China had managed to maintain social stability with high speed of economic growth. The Washington Consensus was proposed by John Williamson in 1990 (Williamson, 1990) as a summary of ten policy recommendations for the structural adjustment in Latin America following the sovereign debt crisis.

that emphasizes high saving rates, human capital accumulation, technological progress, macroeconomic stability, and above all, a well-functioning market that protects property rights and encourages entry and innovation. The real challenge to explain China's growth miracle, therefore, is why China has adopted the right economic policies. This obliges us to turn to the political economy of China's economic growth.

7.4. THE POLITICAL ECONOMY OF THE MIRACLE

There have been many political-economy theories proposed to explain China's success. To cover all of them is beyond the scope of this chapter. Instead, this section provides a selective review of those that either are actively pursued by contemporary researchers or bear implications for the current debates in the field of economic growth. Specifically, we will deal with four topics: fiscal decentralization; promotion within the bureaucracy; institutions and institutional change; and the role of the state.

7.4.1 Fiscal Decentralization

One of the puzzling features of the Chinese regime is that the country has one of the most decentralized fiscal arrangements in the world despite its one-party political system. Xu (2011) calls the Chinese regime a regionally decentralized authoritarian (RDA) regime. He provides an excellent review for the historical roots and implications of this regime. This subsection is not intended to repeat his review. Instead, we will first describe the extent of fiscal decentralization in China, then move on to a discussion of its implications for economic growth, and finally conclude the subsection by pointing out what seems like contradictions in the Chinese RDA regime.

To begin with, we realize that even in the era of command economy, the Chinese system was not totally centralized. There were two waves of decentralization before 1978, one during the Great Leap Forward, and the other during the Cultural Revolution. The first wave ended up with a disaster, but the second wave paved the way to institutionalized decentralization in the reform era. In the early 1980s, a fiscal contracting system was established between the central government and each provincial government. The central government negotiated with each province a separate revenue sharing contract and revised it on a yearly basis. This practice was then mimicked by provinces for their fiscal relationships with subordinate cities, and again by cities with subordinate counties. While the system had a large and positive effect on local economic growth, two consequences had rendered it unsustainable. One was that it was an irregular system, subject to constant changes in almost every year; and the other was that the central government's share of revenue dropped to barely above 20% in the later years, despite its growth in the early years. As a result, a fiscal consolidation reform was conceived in 1993 and put into effect in 1994. This reform had established a tax and revenue sharing system that bears similarities with the American federal system. Three categories of taxes were defined.

They are central taxes, local taxes, and shared taxes. A new tax, the value-added tax (VAT), was introduced as a shared tax. It has been the largest tax since it was instituted. Two consequences have emerged from the reform. First, fiscal federalism was instituted in a politically highly centralized country. This mixed system was more a result of historical imperatives than of a well-designed master plan. As we will see later, however, it seems to have resolved the conundrum faced by many large developing countries regarding the central-local relationship. Second, the reform has consolidated fiscal power to the center, first through VAT of which the center takes 75%, and later through both VAT, and corporate and personal income taxes of which the center takes 50%.[26]

Figure 7.17 presents the shares of the central and local governments in total government revenue between 1976 and 2010. By the end of the Cultural Revolution, the central government's share of revenue was only 12%. It increased to 40% in 1984, but then began to drop again. It jumped to more than 50% in the first year of the tax reform and has since been more or less stabilized. However, the central government's share of expenditure has followed a completely different pattern, as shown in Figure 7.18. It has been declining since 1984, and by 2010 it dropped to below 20%. So, who has financed local governments' burden of expenditure that is way above their revenue capacities? The answer is central government transfer and extra-budgetary income such as revenue from selling land. Central government transfer has been equivalent to more than 70% of its

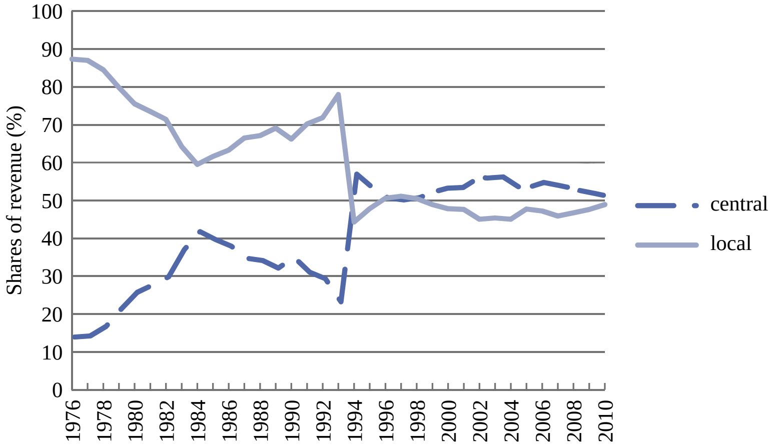


Figure 7.17 Shares of central and local government in total government revenue. *Source: NBS at www.stats.gov.cn*.

[26] This is a rough description of the sharing rule. In practice, it is more complicated. For example, the central government returns part of its VAT revenue to local governments according to a formula that takes into account their base year (1993) figures and their growth rates of VAT in each year.

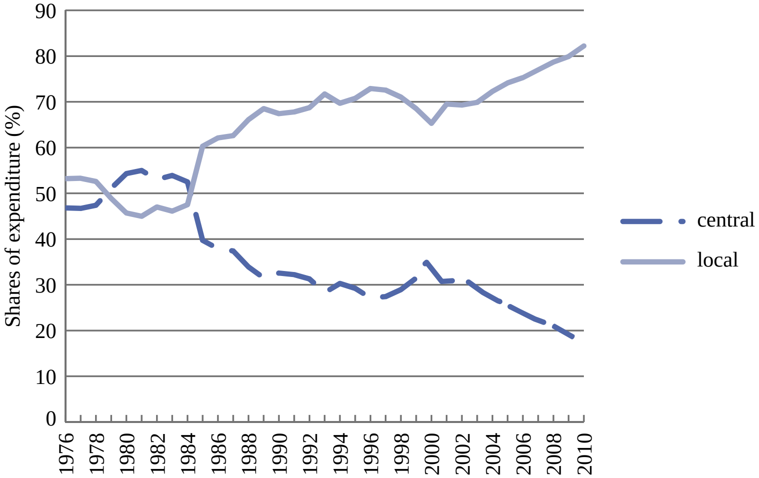


Figure 7.18 Shares of central and local governments' expenditure. *Source: NBS at* *www.stats.gov.cn*.

revenue.[27] This raises the question: why does the central government not leave more revenue to the provincial governments in the first place? The answer is that fiscal transfer is an important leverage that the central government takes on provincial governments. Together with political control, this serves as an important mechanism to allow the central government to implement national goals.

How has fiscal decentralization helped China's economic growth? Xu (2011) has provided an extensive review to answer this question. Not to repeat what he has said, here we highlight three factors. First and foremost, fiscal decentralization has incentivized local government officials to develop the local economy. Unlike fiscal decentralization in other countries that mostly focus on the expenditure side, China's fiscal decentralization has been conducted on both the expenditure and revenue sides.[28] This gives local governments the incentive not just to compete for expenditure handed down from the center, but also to maintain a continuous stream of local revenue. To do that, local governments have to create favorable local conditions to attract businesses and keep them there. Qian and Weingast (1997) believe that fiscal decentralization has created a credible commitment for the government not to grab from firms; and for that, they call China's fiscal federalism market-preserving federalism.

The second, and often neglected role of fiscal decentralization, is that it has led to a reduction of enterprises' tax burdens. Government revenue as a share of GDP dropped

[27] The sum of central government's own spending and its transfer to local governments is larger than its revenue. It covers the gap by issuing public debts.

[28] For example, in India, a country of federal system, local governments were responsible for 58% of expenditure with 38% of revenue in 2003 (Rao and Singh, 2004).

from 31% in 1978 to less than 20% in 1993 (Wang and Hu, 2001). It has increased to about 25% of GDP in recent years, but still quite below the levels under the command economy.

Third, fiscal decentralization has facilitated the reform process. According to Qian et al. (2006a,b), the M-form structure has allowed local experimentation and lowered the cost of reform. Yao (2009) describes China's reform process as one comprised of continuous interactions between local experimentation and ideological adjustment in the center. China's transition has happened without drastic political changes; to move forward, it requires changing the belief system of the ruling communist party. The party, of course, is not ironclad, but comprised of different factions whose political convictions can be quite different from each other. To persuade the hardliners inside the party, the reform-oriented factions have to show that the reform would really help the party and China as a whole. Local experiments serve exactly this purpose.

Several empirical studies support a positive relationship between fiscal decentralization and regional economic growth. For example, Lin and Liu (2000) and Jin et al. (2005) find a positive relationship between the ratio of locally retained revenue to total local revenue and local economic growth.

It is noteworthy, however, that in theory the net effect of decentralization on economic growth is not determined. While it boosts local incentives, fiscal decentralization could also put off economic growth in regions that are doing less well because of reduced central transfers and limited sharing of some key public goods across regions. It may also create regional barriers for cross-regional trade, and foster corruption. In the case of China, Cai and Treisman (2007) have put forward strong counterarguments to the theory that decentralization has contributed to the Chinese growth miracle. Contrary to the results of Lin and Liu (2000) and Jin et al. (2005), Zhang and Zou (1998) find a negative relationship between the two measures for the period 1980–1992.[29] Studies of other countries also return mixed results. In fact, according to Rodriguez-Pose and Ezcurra (2011) who provide a review of the recent within-country and cross-country studies, there are more studies finding a negative relationship between decentralization and economic growth than studies finding a positive relationship. So why has decentralization generally helped economic growth in China?

7.4.2 The Promotion Tournament

In explaining the diverse performances of fiscal decentralization in the world, some recent studies (e.g. Blanchard and Shleifer, 2001; Enikolopov and Zhuravskaya, 2007; Rodriguez-Pose and Ezcurra, 2011) have directed attention to the political institutions that accompany fiscal decentralization. In particular, Enikolopov and Zhuravskaya (2007)

[29] All the three studies may suffer from the problem that the retention ratio was endogenously determined. This may explain why different authors arrive at different conclusions when they study different periods. More credible studies should find a more exogenous measure for decentralization.

find in a cross-country study that the strength of the national parties significantly improves the performance of fiscal decentralization. However, administrative subordination (i.e. appointing local politicians rather than electing them) does not improve the results of fiscal decentralization**.** The first result falls in line with the Chinese reality. It is also consistent with Cai and Treisman (2007)'s argument that China's economic success has to be explained by the growth-enhancing policies adopted by the central government. However, the second result is against China because subnational leaders are generally appointed in the country. To understand the Chinese case, we need to have a close look at how political institutions are interwoven with fiscal decentralization in the country. This leads us to study the Chinese bureaucracy and the promotion tournament embedded in it.

China's civil servant system can be dated back to 1500 years ago when *keju*, a civil examination system, was established. The exam was mainly on the Confucian classics and as a result, the bureaucracy has been deeply influenced by the Confucian doctrines. Although *keju* was abolished in the early 1900s, its core ideas have been preserved. Among them, two have strong implications for contemporary China.[30] The first is an elitist view of the bureaucracy, namely, to qualify as a government official, one has to be learned, capable, and virtuous. The second is a reciprocity view toward governance, namely, people are the subjects to be governed, and in return, government officials should take care of the people.[31] That is, the Chinese bureaucracy has a strong flavor of meritocracy. In practice, it has two significant characteristics. First, government officials are selected from a long process in which they have to show that they are both capable and willing to serve the people and the party. For a young man who is determined to move to the very top of the hierarchy, he has to be prepared that it will take 20–30 years of hard work and a lot of luck for him to do so. Second, government officials are expected to take proactive moves to improve people's welfare. This requirement is quite different from the accountability that a democracy imposes on its officials. Instead of holding government officials accountable to the law, the Chinese regime emphasizes the responsibility that government officials hold toward the people. There are balancing institutions, such as the legislative and law, but their roles are supplementary; the Chinese regime is clearly dominated by the bureaucracy.

Within the bureaucracy, the Chinese Communist Party (CCP) serves as both the controller and the selectorate (Besley and Kudamatsu, 2008). As the controller, the CCP sets the agenda and direction of the bureaucracy; as the selectorate, the CCP selects elites

[30] For a formal treatment of the modern Confucianism and its implications for contemporary China, See Bell (2010).

[31] These two views fit into what Robert Dahl calls "the guardianship view of the state" (Dahl, 1991). Dahl provides convincing arguments as to why guardianship is not consistent with a liberal view of the society. In particular, he argues that the guardian, a single virtuous ruler or a group of technocrats, cannot obtain enough information to take care of the individual needs of ordinary people. This critique has a strong bearing for the Chinese meritocracy, whose problems will be discussed in details in Section 7.6.

and determines their promotion inside the bureaucracy. A system has been developed to evaluate and promote officials. The party has a committee corresponding to each level of the government. Within each committee, the department of organization is in charge of the personnel in the jurisdiction of the corresponding government. Because the number of positions shrinks quickly when one moves up the hierarchical ladder, local officials are effectively engaged in an elimination tournament. Although the criteria of promotion are multi-dimensional, encompassing all the major concerns of the central government, such as economic growth, tax revenue, employment, social stability, and so on, what actually determine an official's promotion invariably lie in two areas, namely, the ability to promote economic growth and the ability to solve the most urgent problem faced by the party. In light of the multi-tasking theory, this comes as no surprise: both are easy to measure and their effects are immediate.

Empirical studies have supported the role of economic growth in the promotion tournament. Li and Zhou (2005) is the first study to show the link between economic growth and promotion. They study provincial party secretaries and governors and find that those officials' chances of getting promoted increase by 15% over the mean if their provinces' average growth rate in their tenures is one standard deviation higher than the average. However, their results are challenged by other studies. For example, Wang and Xu (2008) find that provincial party secretaries and governors who are later promoted to the central government do not significantly outperform others; the provincial leaders who come from and then go back to the central government even underperform the average. One of the problems of studying the provincial leaders is that their promotion can be highly influenced by the center's political preferences and political lineages. For example, Opper and Brehm (2007) construct an index of local leaders' political connections to the members of the political bureau and find that it has a strong predicting power for their promotion whereas economic growth plays a highly insignificant role. In addition, economic growth may not be a sufficient statistic for the leaders' personal abilities because local conditions, very diverse in China, may contribute heavily to local growth. One of the regularities observed for the promotion tournament is that top leaders in the CCP central committee are invariantly promoted from coastal provinces and the two powerful cities, Beijing and Shanghai; even if they originally did not work there, they would be first moved there to prepare for promotion. Therefore, using economic growth rates to predict their promotion may only pick up the effects created by the promotion process itself.

Yao and Zhang (2011) improve the literature by studying city party secretaries and mayors in 241 cities of 18 provinces for the period 1994–2008. They utilize the leaders who were shuffled between cities to construct a large connected sample of cities and leaders. Shuffling serves to make leaders comparable across borders. Without shuffling, leaders' abilities are bundled together with the cities' local conditions. Although leaders having served the same city can be compared because they share the same city fixed effect, a comparison across cities is impossible. Within the connected sample, comparisons can be made. From a leader who was moved from one city to another, one can identify the fixed effects of the two cities. Deducting the two fixed effects from the performances of the leaders having only stayed in one city, one can compare them across cities. Based on their connected sample, Yao and Zhang are able to rank all the leaders and find that the variation among the leaders is a significant contributing factor to the variation of economic growth within the sample cities. In addition, they find that leader ranking is a significant predictor for promotion: the leader of the highest ranking is 30% more likely to get promoted than the leader of the lowest ranking. However, this positive correlation is only found for mayors, not for party secretaries. This is consistent with the different roles they assume in the bureaucracy: the party secretary is in charge of the personnel, political stability, and other less economically related issues, whereas the mayor is in charge of the daily operation of the government, for which economic growth is one of the top priorities.

In summary, the CCP, through a meritocratic bureaucracy, has introduced strong career incentives to the rank of local officials so their conducts are molded to generally promote economic growth. In so doing, the CCP itself has transformed from a political party to the selectorate of the Chinese meritocracy as well as the controller of the country. What is left out is why the CCP has changed its course. In addition, the reliance on a meritocratic bureaucracy does not preempt the role of institutions. Today's China is quite different from the country 30 years ago; much of the difference is due to institutional change as well as income growth. This is the topic of the next subsection.

7.4.3 Institutions and Institutional Change

The thesis that institutions matter for economic growth is widely accepted by economists although there are some disapproving arguments.[32] To the extent that institutions are everywhere and provide incentive structures to agents, the thesis is almost tautological. The real question is why and when growth-promoting institutions are adopted in some countries but not in other countries. To a large extent, this can be boiled down to studying the efficiency hypothesis formally formulated by North and Thomas (1973): institutions evolve to explore economic gains. Following this line, Yao (2004) formally shows in the framework of implementation theory that the efficiency hypothesis does not hold under a well-behaved political process without side payments. In reality, though, the political process can be perturbed and cross-group transactions are commonly used to buy support. For example, agents may engage in a Coase bargaining to obtain the institution that maximizes the social output. This is what Acemoglu calls "the political Coase theorem" (Acemoglu, 2003). However, as convincingly argued by Acemoglu et al. (2006), the political Coase theorem rarely holds in reality because the political dynamics often does not allow for the Coase bargaining.

[32] For example, Glaeser et al. (2004) find that poor countries get out of poverty through good policies, often pursued by dictators, and subsequently improve their political institutions.

The above concise review of the literature highlights the significance of the Chinese transition for the theory of institutional change. The key to explaining China's largely peaceful yet efficient transition from the command economy to the market economy lies in two areas. One is the sense of crisis, and the other is the contingent institutions created in the process of transition. This does not mean that other factors are not important; rather, their significance is of second order. The sense of crisis served as the catalyst for the Chinese transition, and contingent institutions have helped draw different groups to the common agenda of reform.

In the mid-1970s, the CCP leadership faced two kinds of crises: one from the outside, and the other from inside. By the mid-1970s, it was clear to the CCP leadership that China had lost the race with not only the advanced capitalist countries, but also its developing neighbors, including its runaway province, Taiwan. To the old generation of leaders, this reminded them of the old saying that had rung in their ears for decades: "Lagging behind is to get bullied by others." This strong Darwinian belief became one of the impetuses pushing for change. Inside China, the CCP's legitimacy was withering away. The catastrophes of the Great Famine and the Cultural Revolution had depleted the CCP's revolutionary dividends and by the mid-1970s, its top leadership had a strong sense of crisis of legitimacy. The drop in agricultural output in 1976–1977 set the alarm that another famine would fall upon the country and led directly to the ensuring rural reform (Yang, 1998). With procedural legitimacy out of the question, the only choice left for the CCP was to gain legitimacy through performance, i.e. delivering tangible benefits to the population. Turning the party's gravity toward economic growth thus became a national consensus under the leadership of Deng Xiaoping. The reform movement was underway.

To go with the reform, however, there were still many hurdles to overcome. To avoid engendering its own rule, the CCP had to take a gradual approach to reform. This then created two problems in the transition period. One was the resistance of the social and political groups whose interests were tied to the old institutions, and the other was the incongruence of the CCP's own political institutions and the new economic institutions. To overcome those two hurdles, many contingent institutions were created. A contingent institution arises as a response to solve the most pressing issue facing the decision makers. For that it may have to compromise with the existing constraints governing institutional change, so it is often imperfect and will disappear or evolve when the constraints are lifted.

The dual-track price system (DPS) introduced in Section 7.2 is a prime example of a contingent institution. It was certainly not an optimal institution, but in addition to avoiding hyperinflation, it has also managed to evade the backlash of the groups with strong vested interests in the command economy. The way the big-bang approach adopted to attack this issue was fast privatization that was thought would eliminate the political bases of those interest groups (Boycko et al. 1997). In contrast, the dual-track approach provided limited protection to those interest groups, creating what Lau et al. (2000) has called a "reform without losers."

According to Lau et al. (2000), the efficiency of the DPS lies in its two features: one is that the quotas were strictly enforced and market resale of quotas was allowed. This is quite different from the similar reform of the Soviet Union studied by Murphy et al. (1992) that was not able to enforce the quotas. Because the quota prices were lower than the market prices, firms had no incentive to produce for quotas so the dual-track system would collapse. In the Chinese case, quotas were strictly enforced. In this case, administrative discipline helped China's DPS to succeed. On the other hand, allowing the resale of quotas eliminated the inefficiency stemming from the misallocation of quotas. However, quota resale was one of the early sources of corruption in China's reform era. The DPS therefore provides an example of corruption through "greasing the wheels."

The DPS disappeared in the early 1990s primarily because the market prices had converged with the quota prices. The market prices dropped because there were more and more firms supplying to the market. In particular, the township and village enterprises (TVEs) played a significant role. They were not covered by the government plan and had to rely on the market to obtain supplies and sell their products. Their growth greatly enhanced the market track. Yet they themselves were one of the contingent institutions. On legal terms, they were owned by the government, but in effect, they were jointly operated by individual entrepreneurs and the government. In fact, many of them were so-called "red hat" enterprises: they were established by entrepreneurs, but to avoid the uncertainty surrounding private firms, they were registered as township- or village-owned firms. Because of this legal ambiguity, property rights were not clearly delineated within the enterprises. By the standard theory of firm, therefore, TVEs could not have worked. Yet they flourished and contributed to 40% of the national industrial output growth at their highest point (Lin and Yao, 2001). In the 1990s, when private firms obtained a firm legal status, though, almost all the TVEs were privatized.

We can provide more examples of contingent institutions; the rural reform, SOE privatization, and the remuneration scheme for government officials all experienced stages of contingent institutions. Like the DPS, many of them created new forces demanding for less distorted institutions and as a result they disappeared in the end. One issue worth more exploration is the corruption created by contingent institutions. Yao (2004) shows that efficient institutional change is possible if side payments are allowed. Interpreting from this perspective, corruption is one kind of side payment that buys the support for reform. However, this does not tell us why economic growth has not been seriously undermined by rampant corruption. Figure 7.19 compares China with 88 other countries during the world during the period 2001–2009, in terms of corruption and economic growth. Two panels are shown in the figure. The left panel is a scatter diagram of the average growth rate of per-capita GDP against the mean corruption perception index (CPI) in the period. CPI is constructed by Transparency International by citizen surveys conducted each year. Larger values of CPI indicate cleaner governments. There is a weak negative correlation between CPIs and GDP growth rates. China is identified in the chart and is one of the

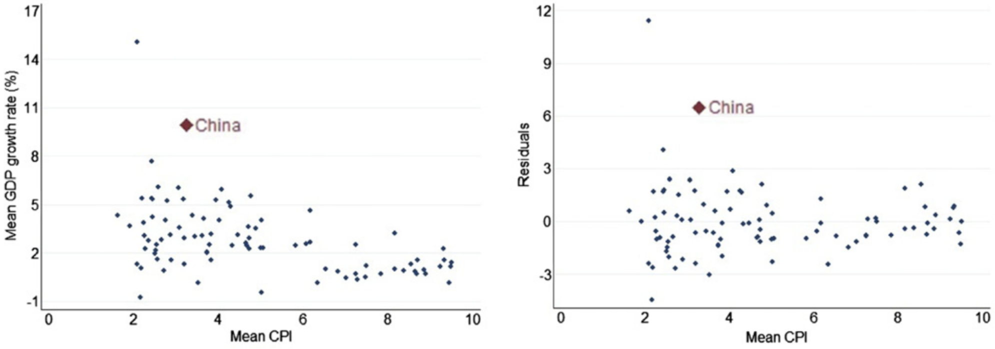


Figure 7.19 Corruption perception and GDP growth in the world: 2001–2009. *Notes:* The left panel depicts the mean CPI and the mean growth rate of per-capita GDP in the period 2001–2009. In the right panel, the mean growth rate of per-capita GDP is regressed on a constant and the per-capita GDP of 2001 (constant $2000), and the residuals are depicted. *Sources: GDP data are from WDI at http://data.worldbank.org/indicator; CPI (Corruption Perception Index) is from transparency international at http://www.transparency.org/. Higher values of CPI indicate cleaner governments.*

two outliers with low CPIs but high growth rates.[33] However, the negative correlation may be caused by the correlation of CPI with a country's initial income. To deal with this issue, the mean per-capita GDP growth rate is first regressed on a constant and the per-capita GDP in 2001, and then the residuals are plotted against the average CPIs again. Now the negative correlation vanishes, but China is still one of the few outliers at the lower end of the CPI. China's average CPI was 3.4, qualifying it as part of the most corrupt 25% of countries. But compared with both countries with cleaner governments, and countries with equally corrupt governments, China gained a much higher average growth rate. Why has corruption not become a serious impediment to China's economic growth?

The RDA regime certainly has helped mitigate the negative impacts of corruption. As the preceding review in this section has shown, the promotion tournament has a strong dose of meritocracy that aligns local leaders' interests with the pursuit of economic growth. In addition, competition among different localities has placed a check on the ability of local officials to grab from the business. This check is reinforced by China's manufacturing-based growth model. The political-economy theory of the resource curse (e.g. Bulte and Damania, 2008) asserts that resource abundance fosters a predatory state that suffocates growth. This thesis is built on two premises: the state has a monopoly on the extraction rights of natural resources, and resources, i.e., the prey of corruption, are not mobile. In a manufacturing-based economy, however, the preys are capital owners who can easily move to other places. Officials in different jurisdictions compete with

[33] The other country is Iraq. Its high average growth rate was probably a result of recovery from the war.

each other; they would be very happy to take the capital driven out by their more corrupt neighbors. Empirical evidence shows that local government officials pay more attention to manufacturing growth than the growth of agriculture and services (e.g. Yao and Zhang, 2011), primarily because manufacturing does not rely much on locality-specific inputs, whereas agriculture and services do. Depending on the composition of the local economy and its reliance on locally provided inputs, therefore, there could be an equilibrium in which government officials assume the role of both a helping hand and a grabbing hand.

A general lesson emerging from China's reform process is that developing countries may have to give up the pursuit for institutional purity and instead focus on institutional efficacy when they conduct policy reforms. Good institutions align agents' own interests with the societal interests. But agents take actions in a web of institutions, many of which can be detrimental to economic efficiency, yet cannot be easily overturned in a short period of time because they are deeply rooted in a country's history. Therefore, the new institutions have to adapt to the existing institutions. As a result, they may not be pure; but with a wise design, they can be effective in raising economic efficiency and creating forces supporting further reform.

7.4.4 The Role of the State

Because China has an authoritarian state, it is so easy to believe that authoritarianism is the key to understanding China's economic success. The preceding review in this chapter has proven that this belief, if not totally wrong, is a gross simplification of what has happened in China. This does not mean that the state has not played a role in China's economic growth; in fact, it has played an important role. What scholars need to study, however, is what this role is and why it has promoted economic growth.

The most prevailing view is that the Chinese state is a developmental state. For example, Lee (2008) summarizes three common features of the governments in China, Japan, and Korea: investing in capacities including human capital and technological progress; gradualism; and government intervention. Like the classical arguments for the developmental state, this summary believes that the three governments take conscious actions to pursue economic growth. Lin (2009) brings development strategy into the analysis. In his formulation, China's economic success in the reform era is a result of the government's conscious change of its development strategy. Before 1978, the Chinese government adopted the HIDS, which was not in line with China's comparative advantage in the labor-intensive industry; after 1978, the comparative advantage strategy has been adopted by which China has embarked on a path of growth featuring industrial upgrading that has followed China's improved factor endowments. In theory, the comparative advantage strategy is essentially equivalent to allowing the market to choose. Lin argues, however, that in reality there are many market failures so the government's conscious choices are required.

What the developmental state school has not fully explained is why and how the government is able to adopt growth-enhancing policies. In particular, it does not explain why some authoritarian governments are able to adopt these policies while others are not. In recent years, there is a small but growing literature trying to answer this question. Two papers in this field have direct bearings for China; both of them emphasize the institutionalization of the ruling party as a driving force for better economic performance in autocracies.

Besley and Kudamatsu (2008) build a principal-agent model to show that the selectorate is more likely to get rid of bad leaders if it has securer power in a divided society. They provide five case studies (including one on China) to illustrate their theoretical prediction. Institutionalization is embedded in the conditions they have identified for better economic performance. In the first place, institutionalization helps the selectorate—the ruling party—to secure its power. More importantly, the selectorate has to have a set of pre-agreed rules—a form of institutionalization—to get rid of bad leaders. If the leader is an absolute despot, there is almost no way to get rid of him except by resorting to violence. Gehlbach and Keefer (2008) provide more direct evidence for the role of party institutionalization. They find in a cross-country regression that autocracies performed better in terms of economic growth when their ruling parties had longer history. They interpret this finding as evidence for the positive role of party institutionalization. Specifically, their theoretical model takes within-party information sharing as the most distinctive feature of party institutionalization. Party members are informed of the behavior of the leader and can punish the latter by refusal. As a result, the leader becomes less predatory on party members who then become more likely to invest. Gehlbach and Keefer (2008) also use China as a case to illustrate their theory.

While their specific mechanisms can be debated, these two studies have rightly pointed out that institutionalization is one of the mechanisms that separate successful autocracies from failed ones, a theme that often emerges from the writings of political scientists studying comparative politics. In the case of China, institutionalization of the CCP has definitely been one of the key drivers for the Chinese government's growth-enhancing policies. In the Mao era, government decisions were haphazard, pretty much depending on Mao's personal preferences which changed frequently. When reform was started, one of the important changes that Deng Xiaoping brought to the party was institutionalization. Government decision making was streamlined and the decision rights were delineated. The standing committee of the political bureau of the central committee was established as the main decision-making body. Personal cult was eliminated and collective leadership has taken root. A mandatory retirement scheme was introduced and an implicit term limit was imposed on the top leaders. A succession rule was established to allow the next-generation leaders be selected by the current leaders and the retired leaders together. In addition, a trinity of power has taken shape to assign the three top jobs, the party secretary, the president, and the chairman of the military committee, to one person, so power is consolidated and the strife experienced in the Mao era can be avoided.[34] Ideologically, the CCP has waved farewell to its revolutionary past and has transformed itself into an all-people's party.[35]

The CCP has become more elitist in the process of transformation. This has been driven by both supply-side and demand-side factors. On the supply side, the CCP has intentionally targeted young people from elite universities (Li and Walder, 2001; Han, 2007; Bishop and Liu, 2008) and put more organizational efforts into sectors with potential high rents (Hu and Yao, 2012). On the demand side, people still want to join the party because the membership carries significant premiums in earnings and career advancement, especially in the sectors of high rents (Hu and Yao, 2012).[36] As a result, the CCP has defied the prediction of the market transition theory (Nee, 1989) and has doubled its membership to 68 million over the last 30 years.

In essence, the institutionalization thesis reiterates the role that checks and balances have played in more constitutionalized systems. However, democracies by design have checks and balances as their inherent institutional elements, yet their economic performances are as diverse as among autocracies. To answer the question why some autocracies have done better than others, one needs to go one step further to study the social and economic conditions that have shaped the autocracies in different places.

In this regard, political scientist Meredith Woo-Cumings' account of the Taiwanese and South Korean experiences provides illuminating ideas. She notes that the governments in Taiwan and South Korea could be relatively free to adopt economic policies that enhanced the two economies' long-term growth prospects in their early stage of economic

[34] Mao served as the chairman of the CCP. Before the Cultural Revolution, Liu Shaoqi served as the president of the country and Deng Xiaoping served as the party secretary who led the daily operation of the party. In the 10 years of Cultural Revolution, these two positions were suspended. The indeterminacy was one of the sources that caused distrust between Mao and his heir-apparent Lin Biao. In open occasions Lin strongly proposed that Mao assumed the presidency although his true wish was that Mao would allow him to take the position so his succession could be secured. Mao sensed that, and firmly rejected Lin's proposal. This also alerted Mao that Lin could be a rival and began to prosecute Lin's close aid, Chen Boda. Mao's move in turn alerted Lin who, together with his son, began a plot to overthrow Mao. His plan fell through; he died with his wife and son in an airplane crash in Mongolia fleeing to the Soviet Union.

[35] This was highlighted by the Three Representations announced in its 16th party congress held in 2002. Instead of representing the working class, the CCP now claims to represent the most advanced cultures, the most advanced productive forces, and the interests of the vast majority of the Chinese people.

[36] Li et al. (2007), however, reject the existence of the income premium. Party members may earn higher income only because they have higher abilities than non-party members. Li et al. (2007) use a unique data set of twins to deal with this identification issue and find that there is no party premium within the pair of twins. On the other hand, Hu and Yao (2012) use the China Household Income Panel Survey (CHIPs) data and find that party membership carries higher premiums in earnings and promotes career advancement in high-rent sectors than in low-rent sectors. Because they make the comparison among party members in different sectors, they can perform a quasi-difference-in-difference study to deal with the issue of self-selection in party membership. They also conduct a panel analysis for people who have data both before and after they joined the party.

development because the two societies were made relatively equal by the Japanese colonists between 1895 and 1945 (Woo-Cumings, 1997). On the one hand, Taiwan and Korea were designated as suppliers of agricultural goods in imperial Japan's version of the Great East Asian Commonwealth so urban industrialists were suppressed in those two places. On the other hand, the Japanese colonists intentionally restricted the growth of the landed class in both places because they feared that this class would become a breeding ground for nationalist sentiments and organized upheavals against their colonial rule. "This discontinuity had a powerful leveling effect, equalizing incomes more than in most developing countries and providing a fertile ground for instituting effective interventionist states, which were given a relatively free hand to forge a developmental coalition as they saw fit." (Woo-Cumings, 1997; p. 331).

He and Yao (2011) take Woo-Cumings' idea to study China. They build a repeated Stackelberg game featuring three actors, the government and two opposing groups of citizens, to study how social equality has induced the Chinese government to adopt growth-enhancing policies. In the model, the two groups of citizens compete with each other to produce the government, and once the government is produced, the other group can wage a revolution to overthrow it if its policy is not desirable for the group. He and Yao show that in the perfect Markov stationary equilibrium, equal political power of the two groups induces the government to treat them equally in social distribution, which in turn guarantees maximum social output. They call the government at this point a disinterested government, i.e. a government that turns blind to political identities, but instead maximizes the social output.

As Section 7.2 has shown, China was made an equal society when its economy took off at the end of the 1970s. This favorable condition has enabled the CCP to ignore the issue of redistribution for quite a long time. In the meantime, there has been no social group that has become strong enough to challenge the CCP's ruling position, so it does not need to waste resources to defend its rule. Then to maximize its own gains, it is rational for the CCP to adopt growth-enhancing policies because economic growth brings both legitimacy to its rule and tangible material benefits to its core members. However, there is no free lunch in the world. The CCP's growth-centered approach has contributed to enlarging income inequality in the country, which would potentially undermine its disinterestedness toward the society. Section 7.7 will discuss this issue in more detail.

In summary, the Chinese government has adopted growth-enhancing policies both because China had favorable initial social and political conditions and because its political system has introduced institutions that align officials' personal interests with the societal interests. It is worth noting that those institutions have not followed any blueprints; instead, they are contingent arrangements aimed at solving the most urgent issues in front of the decision makers. At the philosophical level, this is a result of the Chinese leaders' belief in pragmatism: there is no ultimate truth; what is going on today is a reasonable result as long as it improves the world (Bromley, 2009). There are downsides to this belief, but it has helped China go through the turbulent phase of economic reform.

7.5. EXPLAINING CHINA'S EXPORT-LED GROWTH MODEL

Section 7.2.4 introduced China's export-led growth (ELG) model and its great achievements in the 2000s. This section sets out to explain the driving forces behind China's ELG. The explanation revolves around the double transition of demography and labor reallocation and the reconfiguration of the East Asian manufacturing. Along the way, we will also show that China's export has not led to the fallacy of composition. Following that, we will show that Chinese export has been upgraded throughout the years avoiding the so-called trade trap.

7.5.1 The Double Transition and China's ELG Model

Demographic transition has been shown to be a significant contributor to East Asian economic growth (e.g. Bloom and Finlay, 2009). Since 1979, China has adopted a strict family planning policy. In the city, a family is only allowed to have one child; in the countryside, a so-called 1.5 children policy—meaning a family can only have one child if the first is a boy, but can have a second child if the first is a girl—is embraced. Because of this policy, China has experienced a dramatic demographic transition. Figure 7.20 compares China's age-dependency ratios—defined as the ratio between the dependent population (people aged under 15 and people over the age of 65) and the working population (people between the ages of 15 and 65)—and those of the United States and India for the period 1960–2010. In the 1960s and 1970s, China was broadly similar to India with its

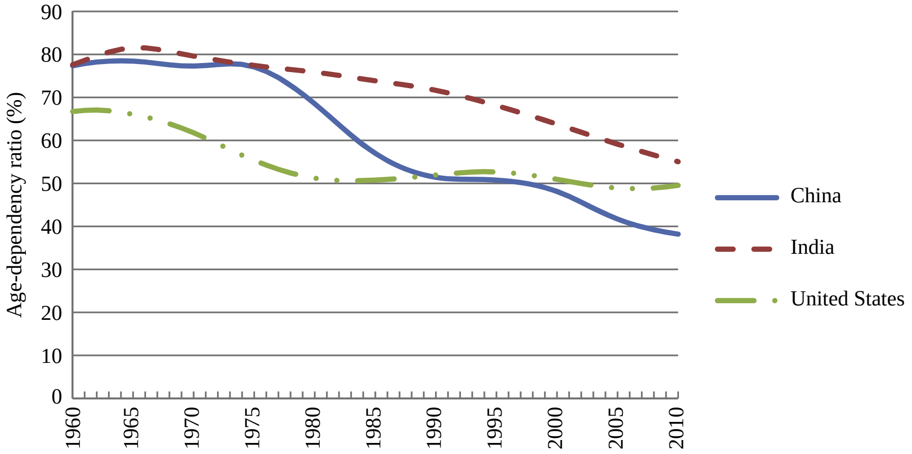


Figure 7.20 Age-dependency ratio in China, India, and the United States: 1960–2010. *Source: UNDP human development index at* http://hdr.undp.org/en/.

age-dependency ratio close to 80%. Since then, China's age-dependency ratio has been dropping by a much faster rate than India's. In 1990, it dropped below the American level of 50%, and by 2010, it reached 38%. In the same year, India's age-dependency ratio was 55%. On the other hand, the age-dependency ratio in the United States has been stabilized at around 50% since 1980.

Fast decline of the age-dependency ratio has two effects on the Chinese economy. One is to increase labor supply. Between 1990 and 2010, China's labor force increased by roughly 10 million each year on average; the decline of the age-dependency ratio contributed 4 million. The other effect is to increase the national saving rate. The life-cycle effect tends to increase the household saving rate. In addition, corporate profits increase because of increased labor supply, but Chinese companies do not pay dividends often so their savings increase. Lastly, the government saves a large portion of its revenue in addition to saving on social security. All three factors contribute to increasing the national saving rate.

Concurrent with the drastic demographic transition, fast industrialization has brought labor out of agriculture and reallocated them to industry and services. Figure 7.21 shows the accumulation of migrant workers between 1993 and 2009. Except in 1997 when the Asian Financial Crisis happened, the number of migrant workers increased every year. Between 1997 and 2009, 8.7 million migrant workers left the countryside each year. The countryside has a large reserve of labor; for a long time, much of this reserve fitted

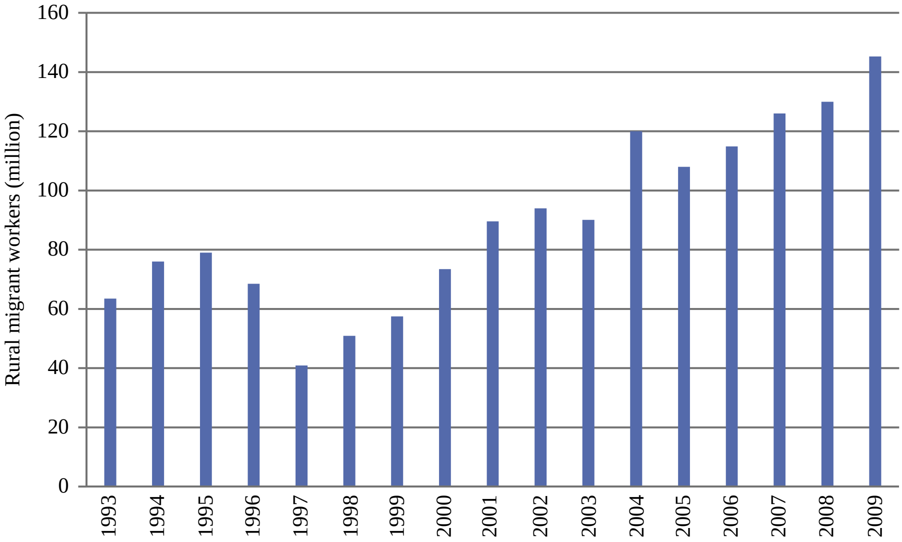


Figure 7.21 Migrant Workers: 1993–2009. *Source: China Yearbook of Labor Statistics, various years. Beijing: China Statistical Press.*

into Lewis' notion of surplus labor. As a result, the wage rate of migrant workers was suppressed, and the industry faced an almost flat curve of labor supply.

The double transition of demography and labor allocation is probably the most fundamental cause for China's ELG and its success. It has both a level effect and a growth effect to increase China's export. The level effect comes from a large labor supply, and the growth effect comes from two sources. One is increased savings which has allowed China to invest in its industrial capacity and upgrade its technology. The other is the differential rates of growth of labor productivity and the wage rate. Between 1990 and 2009, labor productivity in the manufacturing sector grew by an average rate of 13.6% (Figure 7.12), whereas the manufacturing wage only grew by an average rate of 7.4%. This means that China's unit labor cost declined by 66% in that time period. Because labor is the major non-tradable input, this large decline inevitably increased the competitiveness of Chinese exports.

There are signs that China is reaching the turning point of its double transition, though. As a matter of fact, its age-dependency ratio began to rise in 2010. By 2020, China's total labor force will probably begin to decline. On the other hand, wage rates have increased at quite a pace since 2009 (Knight et al. 2011). Many people began to speculate that China had passed the Lewis turning point—i.e. the point when labor supply turns from an infinite elasticity to a finite elasticity (e.g. Cai, 2010; Garnaut, 2010). However, wage increases alone cannot indicate whether an economy has passed the Lewis turning point; it could be a result of increased agricultural income.[37] At the macro-level, agriculture still employs 30% of China's total labor force, although its share in the national GDP is barely over 10%. Using individual data, Knight et al. (2011) estimate a probit model of migration and then compare the propensities of migration of migrant workers and farmers left behind. Extrapolating their results to the whole nation, they find that there were 70 million potential migrants in the countryside in 2007. With the pace between 1997 and 2009, it would take 8 years to fully absorb those migrants.[38]

The above evidence shows that China's double transition will come to an end between 2015 and 2020. Because of that, China's export growth will slow down. In addition, China's overall GDP growth will have to depend on the improvement of human capital.

[37] Lewis's (1954) original formulation of the surplus labor relies on the notion of institutional wage in agriculture. That is, a surplus labor does not increase agricultural output, but is paid by the institutional wage. However, the institutional wage, even if it exists, would be likely to increase as income increases. As a result, labor supply to industry would not be unlimited. Sen (1966) reformulates the notion of surplus labor. His premise is the existence of a constant shadow price of labor in a reasonable range. Within this range, farm households can adjust their labor supply to maintain a constant output when a member is moved out. As a result, industrial labor supply is flat at the constant shadow price.

[38] People left behind are much older than the current migrant workers. However, they can substitute for some of the younger workers already working in the city.

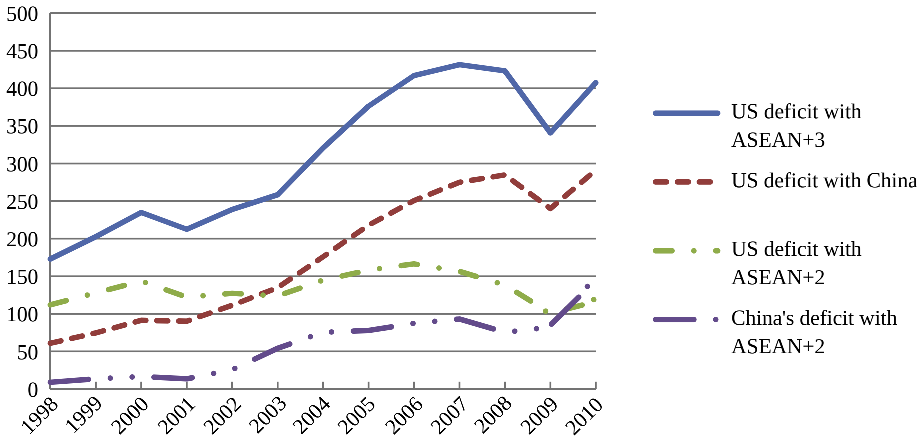


Figure 7.22 Trade balances between the US, Northeastern Asia and ASEAN countries ($billion). *Notes:* ASEAN+3 includes the 10 ASEAN countries and China, Japan, and Korea; ASEAN+2 excludes China. US-China trade balances are reported by the United States. *Source: United Nations, COMTRADE.*

7.5.2 The Reconfiguration of Economic Geography in East Asia

China began to become the world's factory after the Asian Financial Crisis; China's accession to the World Trade Organization (WTO) has accelerated this process by imposing a universal reduction of trade barriers for Chinese exports. The fragmentation of production allowed China to take advantage of its double transition and a relatively robust industrial base to take up the labor-intensive slices of the global value chain. As a result, the economic geography of East Asia has gone through a major reconfiguration. Instead of exporting directly to North America and Europe, China's neighbors provide China with raw materials, resources, and intermediate products; China then finishes the last phase of production and exports the final products to the world.

The four data series shown in Figure 7.22 provide a clue for the reconfiguration. The trade deficit of the US with the ASEAN+3—the ten ASEAN countries plus China, Japan, and Korea—has been increasing every year, except 2009. This increase has been closely matched by the growth of the US deficit with China, which approached $300 billion in 2008 and 2010. On the other hand, the US deficit with ASEAN+2—i.e. when China is excluded—has fluctuated and seldom passed $150 billion. All these can be contrasted with China's enlarging deficit with the ASEAN countries, Japan and Korea as a whole—from $8.8 billion in 1998, to $142 billion in 2010. Apparently, China was taking up the trade surplus of ASEAN+2 with the US, but these countries' losses were more than compensated for by their gains from China.

This conclusion has a strong implication for the fallacy of composition. In a recent paper, Blecker and Razmi (2010) show that China's exports to the United States has a significant crowding-out effect on other developing countries' exports to the United States and they take this as evidence for the existence of the fallacy of composition.

However, they have ignored the demand effect of China's exports for other developing countries. Although China has not become a direct source for consumer goods, it has increased other East Asian countries' exports of intermediate goods to the country (Park and Shin, 2009). In addition, evidence shows that China has played a positive role in increasing intra-ASEAN trade. Devadason (2011) documents inter-industrial trade among ASEAN-5 (Malaysia, Singapore, Thailand, Indonesia, and the Philippines) as well as between them and China. Adding export and import with China in a standard gravity model for intra-ASEAN-5 bilateral trade flows, she finds that a country's exports to another ASEAN-5 country increases by 0.18% or 0.22%, respectively, when its exports to China or imports from China increase by 1%.

In addition to bringing trade growth in East Asia, China has also made contributions to export growth in other parts of the world beyond the United States. As a matter of fact, Figure 7.23 shows that China has had a net trade deficit with the rest of the world other than the United States, in many years since 1998. The three schedules in the figure are obtained by subtracting China's trade surplus with the United States reported by China; reported by the United States; and adjusted by Fung et al. (2006), respectively, from China's total trade surplus.[39] Using Fung et al. (2006)'s adjustment as the best guess

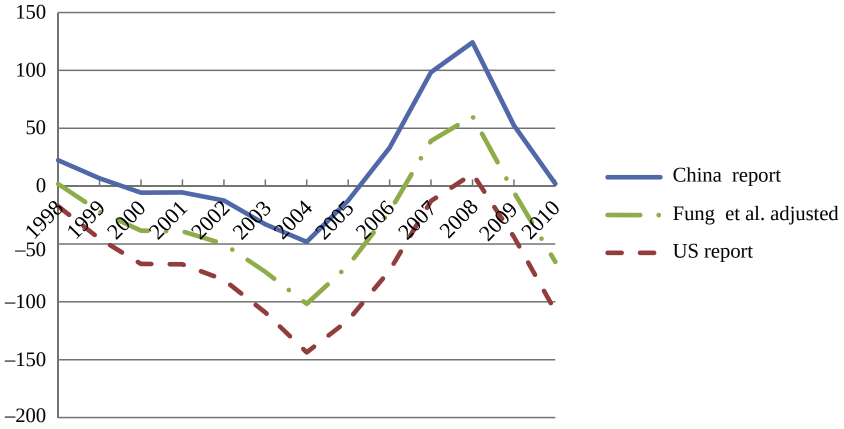


Figure 7.23 China's trade surplus with the world excluding the US. *Notes:* The three schedules are obtained by subtracting China's trade surplus with the United States reported by China; reported by the United States; and adjusted by Fung et al. (2006), respectively, from China's total trade surplus. Fung et al. (2006) only provide data till 2005. They show that Hong Kong's role as a rerouting destination of Chinese exports has declined over the years. For 2006-2010, the average rate of decline between 1998 and 2005 is used to extend Fung et al. (2006)'s adjustment. *Source: United Nations, COMTRADE.*

[39] There are large discrepancies between China-reported China-US-trade balances and US-reported China-US trade balances, mainly because of the trade flowing through Hong Kong. China-reported data do not include the surpluses generated by goods that it exports to Hong Kong but then are re-exported by Hong Kong to the United States, but US-reported data do. Therefore, US-reported deficits with China are much larger than China-reported surplus with the United States. Fung et al. (2006) make several adjustments by taking Hong Kong's re-export into consideration. The adjustment adopted in Figure 7.23

for China's surplus with the United States, we can see that China had deficits with the rest of the world in all the years between 1998 and 2010, except in 2007 and 2008. The fallacy of composition may only exist in a static setting; in a dynamic setting, the growth of export in one country would have a demanding effect for other countries. The true problem facing China is its concentrated surplus with the United States; both countries need to take action to moderate the imbalances between them.

7.5.3 Technological Upgrading of Chinese Exports

Because about half of China's export is low value-added processing trade, people are concerned whether China has fallen into a trade trap of low-end exports. Empirical evidence has shown, however, that China has made substantial progress in upgrading its exports. In the 1980s, the majority of Chinese exports were resources and agricultural goods; in the 1990s, garments became China's top export; and today, electronic products by far are the largest category of China's exports. Indeed, Rodrik (2006) shows that exports from China have been more sophisticated than exports from countries with similar income levels, and Schott (2008) finds that the structure of Chinese exports to the United States is similar to that of OECD countries' exports to the same country.

The domestic technological contents of Chinese exports—i.e. technological contents contributed by Chinese domestic producers—have not been improved at a linear pace, though. Using Rodrik's index of technological sophistication and the input-output table, Yao and Zhang (2008) calculate the domestic technological contents of the exports from China as a whole, Jiangsu province, and Guangdong province, respectively. In 1997, 91% of the technological contents of Chinese exports can be attributed to domestic production; the figure drops to 83% in 2002.[40] The decline is more significant in Jiangsu province, from 92% to 78%. In addition, the decline is more pronounced in the more technologically sophisticated sectors. However, Guangdong province is found to have experienced a V curve between 1992 and 2002. Its domestic technological contents decline from close to 90% in 1992 to barely above 50% in 1997, but bounce back to 80% in 2002. Guangdong is a pioneer in China's processing trade; this V-curve, thus, is a very encouraging sign.

Koopman et al. (2012) develop a new method to use the input-output table to calculate the domestic value-added of a country's exports when processing trade is pervasive. They find that the total domestic value-added in China's export is about 54% in both 1997 and 2002, but increases to 60.6% in 2007. There are large differences between normal exports and processing exports. For normal exports, domestic contents decline from 94.8% in 1997 to 84% in 2007; for processing exports, however, domestic contents increase from 21% in 1997 to 37.3% in 2007.

is based on US-reported data and takes into account mark-up and service fees charged by Hong Kong companies.

[40] The NBS revises the input-output table for the country and each province in every 5 years. Since 1997, the input-output table has contained 124 sectors. Before that, a simplified input-output table of 32 sectors was used.

The above evidence suggests that China's processing export has not trapped the country in low value-added and low-tech trade. The findings on Guangdong province and processing trade are particularly encouraging. It seems that there has been a learning-by-doing process going on in China's processing trade; Coopman et al.'s finding suggests that this process is even stronger in processing trade than in normal trade. In fact, processing companies are not all just characterized by the abundance of labor. For example, Faxconn, the largest processing company in the world, has become a technological leader in China. It was ranked the third in mainland China in terms of granted patents between 2005 and 2010; it was also ranked 13th in granted patents in the United States in 2010.[41]

One remaining issue is that there has been no study that separates domestic firms from foreign firms operating in China. China has been the second largest recipient of foreign direct investment (FDI) in the world. FDI accounts for about 6% of China's overall capital formation, half of China's export and almost all of China's trade surplus (Rosen, 2011). The improvements in processing exports may have all been done by foreign-invested companies including Faxconn, a Taiwanese company.

7.6. DOMESTIC AND GLOBAL IMBALANCES

China's ELG has brought growth and prosperity to the country; in the meantime, serious structural problems have emerged in the economy. In the literature, they are usually summarized under the title of structural imbalances. In the meantime, China's ELG has generated large amounts of trade surplus since 2004, and China's official foreign reserves had increased to an astonishing level of $3.2 trillion by the end of 2011. Not surprisingly, therefore, China has been in the center stage of the global imbalances. For the purpose of the review in this section, China's domestic and global imbalance problems can be summarized into the following three puzzles:

- Puzzle 1:The share of household income in national income has declined since 1996 although per-capita income has increased.
- Puzzle 2:The national saving rate has increased dramatically since 2000. The household saving rate has increased despite the declining share of household income in national income; the size of corporate savings has become as large as the size of household savings although the corporate saving rate has remained constant; and the government saving rate has remained higher than the household saving rate in most years.
- Puzzle 3: China has become a net international capital provider although returns to capital in the country are higher than in most other countries.

Below in Section 7.6.1 we will first provide evidence for these three puzzles, and then in the next several sections will review the explanations that have been put forward for them in the literature. It is noteworthy that all the three puzzles bear direct implications

[41] Foxconn official website at http://www.foxconn.com.cn/WisdomProperty.html.

for China's external imbalances; a thorough understanding of them can help us understand global imbalances as well.

7.6.1 Evidence of the Puzzles

Evidence of Puzzle 1 is shown in Figure 7.24, which presents the initial distribution of the gross national income (GNI). The share of household income increased in the first few years of the 1990s and reached 67% in 1996. It began to decline in that same year, though, and dropped to 50% in 2007. In the meantime, corporate and government income both increased to a quarter of the GNI.

Part of Puzzle 2 is shown in Figure 7.25, which presents China's GDP expenditures. Consumption as a share of GDP has declined since the early 1980s. While the early decline was probably a result of increased income, the decline since 2000 has been very abrupt and warrants close scrutiny. The other side of the story implied by this fast decline of the consumption share is that the national saving rate has increased dramatically from around 40% in the late 1990s to the astonishing level of 52% in 2010.

Figure 7.26 presents the composition of China's national savings in the period 1992–2008 for which the NBS's *Flow of Funds Table* provides data.[42] It provides evidence for the rest of Puzzle 2. The share of household savings in GDP experienced a shallow U-curve over the years. It was the highest in the early 1990s, reaching over 22% and then declined reaching its lowest of 16.2% in 2001. This was mainly caused by the

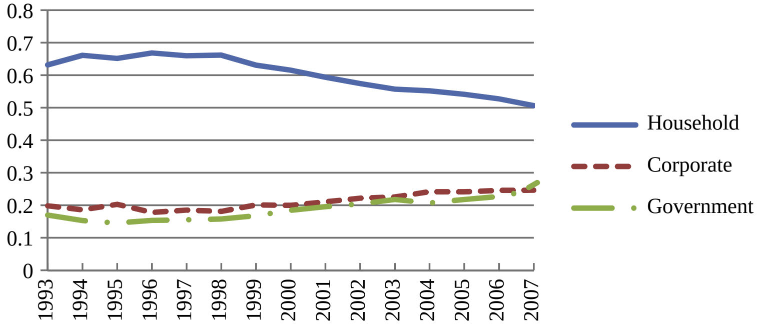


Figure 7.24 Shares of household, corporate, and government income in GNI. *Source:* *Bai and Qian (2009). (The data come from the NBS' Flow of Funds Table, whose latest release is for the year 2008. Bai and Qian (2009) provide adjustments to the official data and obtain more consistent figures.)*

[42] There was a redefinition of corporate savings in 2006. Before that year, all corporate income was counted as savings. Since that year, a substantial portion of corporate savings (50%, 60%, and 40% in 2006, 2007 and 2008, respectively) has been classified as consumption, which caused a significant drop of the national saving rate between 2005 and 2006.

privatization of SOEs that caused massive unemployment.[43] Since then, it has regained some ground and by 2010 climbed to 18.5%. It is noteworthy that as households' share in GDP has declined since the mid-1990s, so has households' share of disposable income. Households' contribution to the national savings has regained ground only because the household saving rate increased quite substantially, as shown in Figure 7.27. In 1992, a typical Chinese household saved one third of its disposable income. This was reduced to a quarter in 2000 and 2001, but then has entered a steady upward path until it reached 39.4% in 2008.

It is important to realize that corporate and government savings grew fast since the early 1990s (Kuijs, 2006). Between 1992 and 2005, before a redefinition of corporate savings was adopted, the share of corporate savings in GDP increased from 11.6% to 20.0%. Even after the redefinition that categorized a large portion of corporate income as consumption, that share still reached 17.8% in 2008. That is, corporate savings were almost as large as household savings. In the early 1990s, China had balanced international trade although household savings at that time were the largest contributor to the national savings. Therefore, the rise of corporate savings is as important as, if not more important, than the rise of household savings in causing China's large current account surpluses in the 2000s.

The contribution of government savings has been more or less stable. On average, government savings accounted for 6.5% of the GDP. However, this was obtained when the government was increasing transfers to the household sector.[44] Therefore, the government must have saved more and more out of its disposable income, which is indeed what

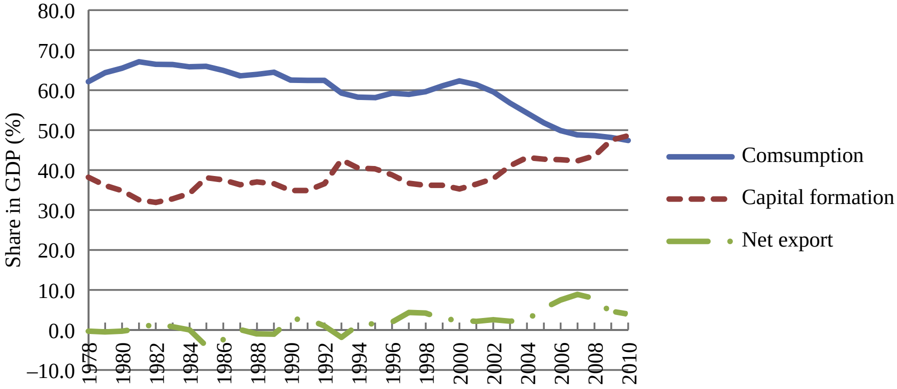


Figure 7.25 China's GDP expenditures: 1978–2010. *Source: NBS at* *www.stats.gov.cn*.

[43] Between 1995 and 2005, close to 50 million SOE workers lost their jobs. In 1998 alone, 20 million were dismissed from their SOE employers (Garnaut et al. 2005).

[44] Between 1993 and 2000, household income on average was increased by merely 4.3% as a result of government transfer. This was increased to 10.4% between 2001 and 2007.

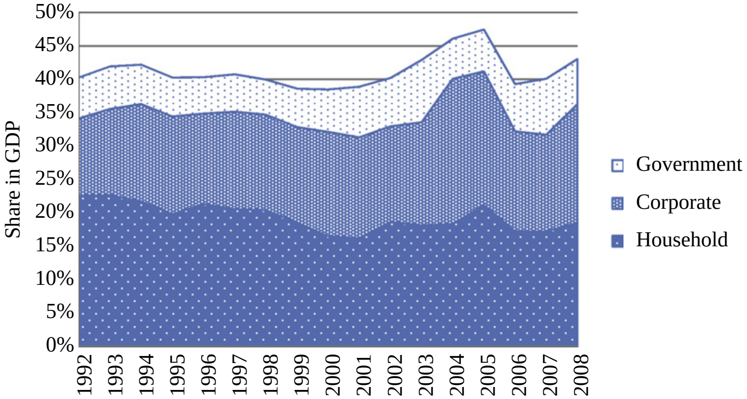


Figure 7.26 Components of China's national savings. *Source: NBS, The Flow of Funds Table, www.stats.gov.cn*.

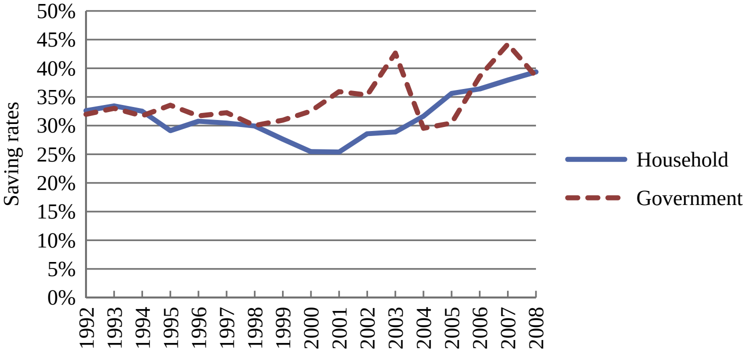


Figure 7.27 Household and government saving rates. *Notes:* Household saving rate is defined as household savings (including purchases of new homes) divided by household disposable income. Government saving rate is defined as government-conducted capital formation divided by government's disposable income (i.e. government revenue net of transfers). *Source: NBS, The Flow of Funds Table, www.stats.gov.cn*.

Figure 7.27 has shown.[45] As a matter of fact, the government saving rate has been higher than the household saving rate in most years. Most of the government savings have gone into infrastructural building. This should be envied by countries stranded by large foreign debts owned by the government; yet at the scale of China's, government savings could become a mixed blessing because government investment, especially investment in industrial parks, is likely to tilt the playing field of the economy.

[45] The sudden drop of the government saving rate in 2005 and 2006 is suspicious. It could be the result of changed definitions during the period 2004–2006.

Finally, Puzzle 3 is again shown by Figure 7.25. The share of capital formation (or the rate of investment, broadly defined) in GDP was largely in line with the national saving rate before 2002, so China's net export was minimum. The rate of investment increased together with the growth of the saving rate since then, but with a slower pace. The result was that China began to experience large current account surpluses since 2003. In 2007 and 2008, China's net export reached 8.9% and 7.9% of GDP, respectively. This is one of the areas that has raised serious concerns in the international community about China's ELG. In the literature, most studies link China's high current account surplus to its high saving rates. However, it is not guaranteed that high saving rates always lead to high current account surpluses. For example, China's national saving rate continued to increase in 2009 and 2010, but its current account surplus declined substantially to 4.3% and 4.0% of GDP, respectively.

The current account surplus is puzzling because the return to capital is high in China. Using GDP accounting data, Bai et al. (2006) find that the aggregate rate of return to capital fell from roughly 25% between 1979 and 1992 to about 20% between 1993 and 1998 and has remained in the vicinity of 20% since 1998. This is confirmed by studies using disaggregated data. For example, Feng Lu and his colleagues use industry-level data and find that the rate of return to capital fell in the range of 10–20% in the 2000s (CCER, 2007). Another indicator of the high return to capital is the discrepancy between the official interest rates and the interest rates charged in the informal financial market. While the official base lending rate has been 6–7% in recent years, the rates prevailing in the informal market of Wenzhou, a city famous for its private business development, have been more than 20% (Rosen, 2011). On the other hand, China's official foreign reserves have increased dramatically since 2004 and reached $3.18 trillion by the end of 2011. More than one third of it has been used to buy US treasury bonds, which only promise a rate of return of 2–3%. So the question is: why does capital not stay in China to reap the high returns instead of becoming the official foreign reserves and flowing to other countries to buy low-return assets such as United States treasury bonds?

In the next several subsections, we will present the explanations for the three puzzles. Section 7.6.2 presents three explanations that treat the three puzzles as by-products of China's double transition and its higher growth rates relative to other countries. Section 7.6.3 sets out to explain rising household saving rates. Sections 7.6.4 and 7.6.5 then discuss the roles of the financial sector and the government in magnifying China's imbalance problems. Lastly, Section 7.6.6 deals with the issues of the exchange rate.

7.6.2 Structural Change, Economic Growth, and the Puzzles

It is clear that Puzzle 1 is inconsistent with one of the Kaldor Facts (Kaldor, 1957), namely, the shares of national income received by labor and capital are roughly constant over long periods of time. To resolve this inconsistency, one may have to consider China's drastic structural change that has taken place since the 1980s. Specifically, two mechanisms are worth exploring.

One of them is related to the different shares of labor income in agriculture, manufacturing, and services. The manufacturing sector is more capital intensive than agriculture and services in developing countries. Before manufacturing reaches the highest point of its hump-shaped trajectory of employment, it draws labor out of agriculture so the share of labor income in the national income declines. After manufacturing climbs over the highest point of its hump-shaped trajectory, services begin to pick up and the share of labor increases. That is, the labor share in national income should exhibit a U-curve as a country develops. Li et al. (2009)'s cross-country panel study has confirmed this U-curve. It can also be generalized to the consumption share. For example, both Japan and Korea have experienced a U-curve in their consumption shares (Cai, 2011). Because Chinese manufacturing is still on the left side of the hump, we can then understand Puzzle 1 and the first part of Puzzle 2, namely, the national saving rate increases. In addition, adding two more facts about China can explain another part of Puzzle 2, i.e. the rising shares of corporate and government savings. One is that Chinese companies do not distribute profits often, and the other is that companies reinvest most of their retained earnings and the government invests a large part of its revenue.

However, the above explanation may not be able to account for everything that is happening to China. Japan's consumption share declined before 1971 and has increased since that year. In Korea, the trough happened in 1989. However, the consumption shares at the trough in both countries were higher than China's today. In Japan, it was 52%; in Korea it was 60%. In contrast, China's consumption share was 48% in 2010 and seemed to continue to decline. Therefore, it seems that the different shares of labor income in the three sectors alone cannot fully explain China's deeper trough.

The other mechanism is through suppressed wage rates. When surplus labor exists in agriculture, the industrial wage rate is constant. In reality, manufacturing wage rate grew by an average of 8% per annum in the period 1992–2008 (Conference Board, 2010). This is high by international standards. However, it is lower than the growth rate of labor productivity in the same sector, which was 16.8% per annum (Figure 7.12) between 1991 and 2009. This large discrepancy between wage and productivity growth can explain Puzzle 1 and part of Puzzle 2. To see that, let us consider a simple case in which A stands for the labor productivity (measured as per-worker value-added) in the whole economy, w stands for the wage rate, and L stands for the total number of workers in the economy. Then GDP is simply AL, labor income as a share of GDP is w/A, and profit (returns to capital) as a share of GDP is $1 - w/A$. In most cases, w should grow roughly at the same rate of A, so the shares of labor income and profit are constant over time. When the wage rate grows more slowly than labor productivity, then the share of labor income in GDP declines. In the Chinese case, more than 90% of household income comes from labor earnings. Therefore, household income as a share of GDP also declines. This explains Puzzle 1. Because companies do not distribute their profits often, but instead reinvest most of them, and the government also invests a large portion of its revenue, the national saving rate increases. So does the share of savings contributed by companies and the government. This explains part of Puzzle 2.

The growth of labor productivity comes from two potential sources, TFP growth and capital deepening—that is, the growth of per-worker capital stock. As Section 7.3.2 showed, the growth of TFP has been moderate compared with the growth of labor productivity. Capital deepening is a more substantial contributor. Then, how does capital deepening happen?

Let k stand for the capital intensity (per-worker capital stock), s_c stand for the saving rate in the corporate + government sector, and s_h stand for the saving rate in the household sector. To make it consistent with the Chinese reality, we assume $s_c > s_h$. In addition, we make the simple assumption that the household sector is only comprised of workers and household income is only comprised of labor income. Then the growth of k can be conveniently expressed by the following:

$$\dot{k} = A\left[s_c\left(1 - \frac{w}{A}\right) + s_h\frac{w}{A}\right]. \tag{7.1}$$

That is, it is the sum of per-worker savings in the corporate + government sector and the per-worker savings in the household sector. It is obvious that $\dot{k}$ is positive. More importantly, its size increases as long as the share of household income w/A declines. That is, capital intensity grows at an accelerated pace. Capital deepening happens in every economy as long as it grows. What distinguishes China from other countries, therefore, is that capital deepening happens at an accelerated rate in China.

The key here is that the wage rate grows more slowly than labor productivity. Even without surplus labor in agriculture, this can still happen. In a two-sector model with agriculture and industry, it is easy to envision that the industrial wage rate is determined by the marginal product of labor in agriculture. In such a simple framework, the gap between industrial labor productivity and its wage rate is determined by three factors: (i) the gap between the TFP growth rates in the two sectors; (ii) the gap between the rates of capital formation in the two sectors; and (iii) labor migration from agriculture to industry. The growth of TFP is not low in agriculture, mostly due to fast biotech innovations. But capital investment in agriculture has been minimal and will perhaps remain low in the future. Therefore, we have to wait for labor migration to reach a certain point to see the wage rate grow faster than labor productivity.

Structural change, however, cannot explain the rise of the household saving rate and large current account surpluses. Before we present more detailed explanations for those two phenomena, we first discuss the role of differential rates of growth to create global imbalances. In this regard, the prevailing view is set by Engel and Rogers (2006) that higher expected growth rates imply a larger future share of the country in the world output, so consumers in this country should borrow from other countries today. Obviously, China's high current account surpluses are inconsistent with this view because the share of the Chinese economy in the world is increasing due to its faster growth rate. Historical evidence also shows that most countries had current account surpluses in their high-growth periods (Xu and Yang, 2012). One of the reasons may be that Engel and Rogers do not consider how the current growth rate affects savings and investment.

In the spirit of the life-cycle hypothesis (LCH), Liu and Yao (2012) consider both the national saving rate and the investment rate for the relationship between the current growth rate and the current account. In their overlapping generation model with a single economy, agents live for two periods; they save when they are young and consume the savings when they are old. Consumption is comprised of a self-produced part and a purchased part with the share of the first part declining. The production side features decreasing returns to scale and labor-augmenting technological progress. In such an economy, the capital-output ratio is no longer a constant but declines when output grows faster. Then the national saving rate is a convex function of the growth rate while the investment rate is almost linear in it. So the current account should exhibit a U-curve when an economy grows faster. Liu and Yao's empirical study of 216 economies for the period 1960–2010, based on various specifications, confirms the existence of the U-curve. The trough happens when an economy grows by about 6% per annum.

Taking Liu and Yao (2012)'s results, we can understand both China's surpluses and the United States' deficits. China is on the right side of the U-curve so it is more likely to have surpluses when it grows faster; the United States is on the left side of the U-curve so it is more likely to have deficits when it grows faster. Furthermore, the United States has been growing faster than other advanced countries, so it is more likely to incur a deficit than those other countries. Therefore, Liu and Yao (2012) can accommodate the result of Engel and Rogers (2006).

7.6.3 Explaining Rising Household Saving Rates

China's rising household saving rates have caught wide international attention in recent years. It is therefore worthwhile to set aside a separate subsection to review the literature on this important issue. A large volume of literature has emerged in recent years. This subsection will not be able to provide a full review for all the relevant papers; instead, we will concentrate on two strands of explanations related to income growth, demographic transition, and precautionary savings.

An early attempt to explain China's high household saving rates is Modigliani and Cao (2004) in the framework of the life-cycle hypothesis. The LCH implies that income growth—not income level—and population structure are the two factors determining a country's national saving rate. To be precise, the national saving rate is proportional to the GDP growth rate under a stable capital-output ratio, and a higher working-age ratio increases the proportion. Because of the setup of the LCH, the national saving rate is equivalent to the household saving rate in Modigliani and Cao's theoretical model. So they study the national saving rate in their empirical work. They find that income growth has been the dominant factor behind the dramatic increase in China's saving rate in the period 1953–2000. Their point estimate shows that 1% point increase of the GDP growth rate leads to 2% point increase of the saving rate. Since the GDP growth rate was increased by 6–8% points in the reform period, accelerated GDP growth can account for 12–16% point increase of China's national saving rate in this period. On the other hand, demographic structure has a lesser impact. One percentage point increase of the working-age ratio would only lead to 0.0015% point increase of the saving rate. The total effect of demographic transition therefore is small although the rising working-age ratio was increased by more than 50% points.

These results are confirmed by Horioka and Wan (2007) studying provincial panel data and Ang (2009) comparing China and India. However, the above studies have all relied on a reduce-form approach to study the effects of rising working-age ratios and could underestimate the impacts of demographic transition. For example, China's economic growth was accelerated by about 2% points in the 2000s compared with the previous two decades, and its working-age ratio increased by 50% points between 2000 and 2010. Using Modigliani and Cao's results, then, accelerated growth can explain a 4% point increase in the national saving rate, and the impact of higher working-age ratio can be ignored. However, the national saving rate was increased by 14% points (Figure 7.25) and the household saving rate was increased by 15% points (Figure 7.27) in the 2000s. Apparently, the LCH can only account for a small fraction of those increases.

Curtis et al. (2011) aim at remedying the potential shortcomings of the reduce-form approach. They build an explicit overlapping generation model in which agents live for 85 years, and study the change of China's household saving rates in the period 1963–2009. What they have arrived at is a strong result: under the parameters used in their calibration, the change of China's household saving rates in this period can be almost entirely explained by demographic transition. In particular, their model result predicts 25% for the household saving rate in 2009, only 2% points short of the actual rate.[46] However, Curtis et al. (2011) do not calibrate their model by periods and thus may have underestimated the contribution of the change of the GDP growth rate in different periods.

One of Modigliani and Cao (2004)'s purposes is to show that the LCH performs much better than the Keynesian model of savings/consumption that ties the current savings/consumption to the current level of income. While the Keynesian model does not perform well at the aggregate level, as Modigliani and Cao have shown, it performs very well at the household level. This is evidently shown in Chamon and Prasad (2010) who study household saving behavior using the NBS urban household survey data for the

[46] Curtis et al. (2011) exclude purchases of new homes from household savings albeit the official statistics (such as those shown in Figure 7.27) include them.

period 1990–2005. In their summary regressions, per-capita income has very significant effects on the household saving rate. When per-capita income is doubled, the household saving rate will increase from 14.5 to 19.4% points depending on the regression specification. Wei and Zhang (2011), using provincial data for the period 1980-2007, find even large effects in the range of 20 (urban areas) to 45 (rural areas)% points. Because per-capita income in urban areas more than doubled in the period 2000–2010, income growth alone can more than explain the growth of the household saving rate in this period of time.

Extending the Keynesian model, one may also study the impact of worsening income distribution on China's aggregate household saving rates. High-income households have higher average propensities to save than low-income households. Therefore, the aggregate household saving rate increases as income is being concentrated to higher-income households. Because income distribution has been deteriorating fast in the 2000s, there is a good reason to believe that a more skewed income distribution is one of the contributors to China's growing overall household saving rates.

Many recent studies have resorted to the motivation of precautionary savings to explain China's rising household saving rates. For example, Wen (2010) calibrates a theoretical model featuring borrowing constraints and future consumption uncertainties, and finds that China's high household saving rate can be mostly explained by precautionary savings under borrowing constraints. Chamon and Prasad (2010) find that household saving rates increase in all age groups, particularly in the young and old groups. Despite their strong results on the level of income, they tend to attribute their findings to higher income uncertainties; and the lack of social security. In a related paper, Chamon et al. (2010) use the China Health and Nutrition Survey data to formally study those two factors. Adopting a precautionary saving model, their calibration shows that rising saving rates among younger households are consistent with rising income uncertainties; and higher saving rates among older households are consistent with a decline in the pension replacement ratio for those retiring after 1997. They conclude that rising income uncertainty and pension reforms can account for over half the increase in the urban household saving rate in China since the mid-1990s.

While the precautionary saving thesis has a lot to recommend, the causes behind the precautionary motives need to be scrutinized more closely. One of the frequently invoked causes is the lack of social security. However, historical data may suggest the opposite. In the 1990s, China's social security system was greatly eroded because of SOE privatization; in the meantime, household saving rates declined. Since the end of 1999, both the coverage and the benefits of social security have been indisputably improved (Shen and Yao, 2009), yet the household saving rate has increased. That is, the empirical evidence since the early 1990s does not support the precautionary saving thesis.

There are studies directly estimating the effects of social security on savings/consumption using household data. However, the effects seem to be small. For example, Ma et al. (2010) find that the new medical insurance scheme has raised farm households' food consumption by ¥81 each year, or about 2% of their total annual consumption expenditure. Bai and Wu (2011)'s study concurs with this finding; they find that the medical insurance has increased farm households' total consumption by 5%. Yao and Zhou (2012) provide a comprehensive study on the impacts of social security on household consumption using the newly released urban and rural household data by the China Family Panel Studies (CFPS).[47] Their main novelty is to estimate the effects by quantiles. Their premise is that income distribution is highly skewed in China and most people at the richer end have already got good social security coverage, so the aggregate effects of social security expansion could be small although it might have large effects on people at the poorer end. The results of their quantile regressions have confirmed their conjecture. While the impacts on poorer households are high—for example, expanding medical insurance from its current coverage to universal coverage can increase household consumption of the lowest 10% of the population by almost 30% in both the countryside and the city—the aggregate effect is small: household consumption would only increase by 0.3% points in the national GDP if all urban and rural families were covered by the current medical insurance, and would only increase by 2.6% points if urban families are also fully covered by pension and housing funds.[48]

A more plausible explanation along the line of precautionary savings is to look into the role of the rising housing prices. It is widely observed that housing prices have increased dramatically in the 2000s in most Chinese cities. It is also observed that Chinese home buyers often pay a higher down payment than required by law (Chamon and Prasad, 2010). As a result, they have to save quite a lot in order to buy a home. Chamon et al. (2010) attribute the high saving rates of young households to income uncertainty; but they can also be caused by those households' desire to buy new homes. Chen et al. (2012) concur with this view. They use the 1998 housing reform as a natural experiment and find that the reform—it privatized public housing and stopped government-provided housing—has caused families not owning a home to increase their saving rates by 2.3% points. This effect is small compared with the large increase in the household saving rate. This is probably because Chen et al. (2012) only study the average effect of the commercialization of the housing market, but not the rising housing prices per se.

[47] CFPS is modeled on the American Panel Studies of Income Dynamics (PSID). It is the first independent longitudinal survey in China. It covers more than 9000 households in rural and urban China, and the survey is conducted every 2 years. The first wave was conducted in 2010. The Institute of Social Science Survey at Peking University administers the survey.

[48] Pension coverage is currently very low and housing funds (funds that people can borrow from to buy homes) virtually do not exist in the countryside, making the estimation highly unreliable. The growth of 2.6% points of the household consumption in national GDP is not trivial compared with China's current account surplus of recent years, which was 3.5% of GDP in 2010. However, the cost to fulfill full social security coverage can be very high because the current coverage is very low. For example, the CFPS data show that 47.7% of the population did not have medical insurance and 85.7% of the qualified population did not have a pension in 2010.

In a recent paper, Wei and Zhang (2011) propose and test an interesting thesis for China's high household saving rate from the angle of high sex ratios. China's sex ratio at birth has increased over the years because of stringent family planning policy. Instead of 105–107 of the normal range, the sex ratio is 122 in China. Wei and Zhang reason that high sex ratios intensify the competition among men in the marriage market and force them to increase their values in the market. One way to do this is to increase savings intended for buying a home, car, and other status and wealth-related items. Wei and Zhang's empirical analysis finds that both cross-regional and household-level evidence supports this hypothesis; high sex ratios can potentially account for 60% of the actual increase in the household saving rate during the period 1990-2007. But this effect seems too high. One factor Wei and Zhang are unable to control is the cultural heterogeneities in different parts of China that are simultaneously correlated with saving behavior and the preference for sons. High sex ratios have contributed to China's high household saving rate, but their significance is not likely to be as large as Wei and Zhang have shown.

In summary, the growth of China's household saving rates in the 2000s is likely to be linked with rising per-capita income, higher GDP growth rates, and precautionary saving motives, particularly those associated with rising housing prices. Worsening income distribution can also be a significant factor. High sex ratios play a role, but it is not likely to be a significant factor. More studies are needed for the link between rising housing prices and higher saving rates.

7.6.4 The Financial Sector and China's External Imbalances

High national saving rates do not necessarily lead to high current account surpluses; it is not clear why China has to run large current account surpluses when the aggregate return to capital remains high. Liu and Yao (2012) can explain why a high-growing country like China is prone to run surpluses, but they cannot explain why China maintains higher returns to capital than other countries. To fully explain Puzzle 3, therefore, we need a more structured approach. In this case, the financial sector can be a focal point of such an approach because it is the intermediary between capital providers and capital users.

The recent literature has emphasized the role of finance in creating global imbalances (e.g. Caballero et al. 2008; Mendoza et al. 2009; and Ju and Wei, 2010). When the financial markets of countries differ in their capabilities to allocate capital, capital flows from countries with less efficient financial markets to countries with more efficient financial markets. That is, countries with more efficient financial markets are more likely to become debtors and countries with less efficient financial markets are more likely to become creditors in the global balance of payment. China's financial sector is one of the least reformed sectors in the country; it is much underdeveloped compared with the financial sectors in advanced countries. Viewed against the above new literature, it is then hardly surprising to find that China runs large current account surpluses. To understand how a weak financial sector has played out in China's external imbalances, it is worthwhile to first take a brief review of its deficiencies.

China has a bank-based financial system with bank credits accounting for more than 70% of total finance. The capital market is underdeveloped. There are less than 3000 companies listed in the stock market; local capital markets at the subnational levels are very thin if they exist at all. In the stock market, there are virtually no corporate bonds. Within the banking sector, state-owned banks dominate, and the number of banking institutions is small, less than 3000 even if rural credit unions are included. This can be compared with more than 18,000 in the United States, a country whose nominal GDP is 2.5 times of China's. In addition, interest rates are directly controlled by the government. The saving rate has been lower than the inflation rate since 2004 and the base lending rate is less than one third of the lending rate in Wenzhou's informal financial market (Rosen, 2011).

Among the consequences of these deficiencies, the following have direct implications for China's domestic and international imbalances. First, households' financial income is suppressed. One of the functions of a well-functioning financial market, especially the capital market, is to allow ordinary people to share the fruits of future economic growth. China's financial sector is not doing a good job in this respect; instead, it transfers wealth from ordinary depositors to banks and corporations through suppressed deposit rates and a low propensity to distribute dividends. That is, it contributes to Puzzle 1. Second, large companies, large SOEs in particular, are favored by the financial sector and the supply of capital and credits to them is abundant. In contrast, small and medium enterprises (SMEs) are consistently rationed by financial institutions. But SMEs provide 80% of urban jobs, so discrimination against them hurts the growth of employment, which then contributes to Puzzle 1 again. Third, also because of the rationing, SMEs have to raise funds on their own, most of the time relying on retained profit to augment their working capital and to take on new investment projects. As a result, corporate savings increase. Thus we have part of Puzzle 2. Fourth, the different treatments received by privileged and unprivileged firms have also the potential to create a mismatch between growth and the availability of financial resources. The growth potential of large firms is smaller than that of smaller firms, at least at the aggregate level. The abundance of capital to large firms will ultimately meet the constraint of diminishing marginal returns whereas credit rationing forces SMEs to operate at a stage where the return to capital remains still high. Diminishing returns in large firms and rationing on SMEs could even reach the point when part of the capital has to be invested outside the country, so Puzzle 3 follows.

Song et al. (2011) take up the last idea seriously and build and calibrate a general equilibrium model to show how a defective financial system can lead to a large trade surplus while the country sustains high output growth and high returns of capital. They distinguish between two kinds of firms, entrepreneurial firms and SOEs. The former are more efficient than the latter, but the latter are favored by the financial sector while the former are rationed. Growth comes from entrepreneurial firms who have to rely on retained profit to invest to generate further growth. On the other hand, the share of the SOE sector shrinks, forcing the financial sector to invest abroad.

Song et al. (2011)'s categorization of the two types of firms is disputable. After the massive privatization carried out between 1995 and 2005, the SOEs left are generally as efficient as private firms. One of the reasons for their success is that many of them operate in monopolistic sectors or receive government support. On the other hand, banks do not favor all SOEs; they discriminate against small SOEs as well as private SMEs. In addition, they favor large, private firms as well as large SOEs. The dichotomy of the availability of credits Song et al. (2011) have imposed on the two types of firm has simplified their modeling, but does not fully reflect the reality.

Mao et al. (2012) extend the literature of finance and global imbalances and study how a country's comparative advantage in finance and manufacturing affects its current account balance. Their theoretical model shows that a country with a comparative advantage in manufacturing would end up with current account surplus, and vice versa. Using a panel data of OECD countries and defining the finance-manufacturing comparative advantage of a pair of countries by the ratio between their relative labor productivity in the financial sector and manufacturing sector, their empirical study has found that countries with a comparative advantage in finance tend to have current account deficits. Tan et al. (2012) complement Mao et al. (2012) by studying how a country's financial structure affects its current account balance. Their theoretical argument is that a bank-based financial system is more likely than a market-based financial system to generate surplus. The element that makes the difference is the finance of SMEs. Because SMEs have higher risks than large firms, and banks are inherently averse to risks, the financial need of SMEs is not likely to be met in a bank-based system and SMEs have to rely more on their retained profits for finance. In a market-based system, though, it is easier for SMEs to get finance through the capital market, and because of that, it is also easier for them to get bank finance—their finance through the capital market can boost banks' confidence to lend to them. Tan et al. (2012) first study a large panel of countries and find that financial structure matters for a country's current account balance. A closer study of the OECD countries has found that only corporate savings are affected by financial structure whereas household and government savings are not. Then studying cross-country firm survey data provided by the World Bank, Tan et al. (2012) find that SMEs tend to retain more profits for investment in countries with a more bank-based financial sector than in countries with a more market-based financial sector, whereas large firms are not different when the financial structure changes.

The results of these two studies are indicative for China although they do not study China per se. When the financial sector is added to the equation, it is relatively easier for us to explain the anomaly of the coexistence of high returns to capital and current account surpluses. A weak financial sector and a relatively strong manufacturing sector give China comparative advantage in manufacturing over finance. As a result, China tends to concentrate on manufacturing and buys financial services from (i.e. export capital to) countries with advantages in finance. The returns to capital can be high in the manufacturing sector, but the deficiencies of the financial sector induce outflows of capital. On the other hand, the dominance of banks in the financial sector forces SMEs to rely on retained profits and creates a wedge of returns to capital between large firms and SMEs. Because international capital flows are determined by the return to capital among the privileged large firms, China ends up with exporting capital while the country's aggregate returns to capital remain higher than the international level.

7.6.5 The Role of the Government

The Chinese government controls a large portion of the Chinese economy. Its budgetary income (mainly taxes) accounts for a quarter of the national GDP; adding other forms of income and social security, the income directly controlled by the government can be as high as one third of GDP.[49] In addition, the state sector accounts for about 30% of the national GDP (Yao, 2011). Although many SOEs have become commercial entities, the government still maintains a strong influence on their investment and strategic plans, and above all, appoints their managers. On top of that, subnational governments at various levels borrow heavily from banks. In the last several years, infrastructural investment has accounted for more than one third of total bank lending, most of which has been undertaken by governments (Rosen, 2011). The 10 trillion yuan in debt accumulated by local governments are a result of this investment frenzy.

Because the government directly or indirectly controls more than 60% of the Chinese economy, it is not surprising that China's internal and external imbalances are linked with the behavior of the government. In particular, the government has aggravated the imbalance problem in the following three areas.

First, government revenues are spent heavily in areas related to economic growth and transfers back to citizens are limited. Overall, government spending on economic affairs accounted for 20.1% of total government spending in 2008, much higher than other countries; in contrast, government spending on health care and social security was only 7.4% and 20%, respectively, much lower than other comparable countries (Bai et al. 2010). As a result, government savings are very high, contributing to one fifth of the national savings in recent years (Figure 7.27).

Second, government investment favors more capital-intensive producers in manufacturing. Government infrastructural investment is not confined to highways and railways; a large fraction of infrastructural investment conducted by local governments is directed to the numerous industrial parks that local governments build to attract investors. One of

[49] According to a report released by the Institute of Finance and Trade Economics, Chinese Academy of Social Sciences on September 10, 2010 (IFTE, 2010), government budgetary income accounted for 25.4% of GDP in 2009, but government total income was increased to 32.3% if government funds income, extra-budgetary income, land sales revenues and social security income are added.

the hard constraints those industrial parks face is the limited supply of land. This has led to the paradoxical observation that, on the one hand, land in those parks is sold with a price much below the cost local governments have incurred in purchasing and preparing it; and on the other hand, local governments require investors to meet investment and tax quotas designated for units of land. The result is that firms entering the industrial parks are much more capital intensive than those outside. Because most local governments have concentrated local industrial development into industrial parks, firms that cannot enter the parks often find it very hard to get land. Many potential firms could be forced out of the market because of the lack of land. In addition, smaller firms outside the parks have to rent land, reducing their ability to collateralize their borrowings. The macroeconomic consequence is that the wedge between the privileged and unprivileged is widened and more imbalances are created.

More than that, a consequence of heavy government investment is the lowering of household consumption. Apparently, government investment has a direct crowding-out effect on household consumption. In addition to that, it lowers household consumption indirectly by distorting the economic structure. Chen and Yao (2011) take up this idea to study a panel of provincial data for the period 1978–2006. They find that when the share of infrastructural investment in a province's government spending increases by 1% point, the share of household consumption in GDP declines by 0.31% points in that province. Their further exploration shows that this happens by two channels. One is that government infrastructural investment promotes the secondary sector, and the other is that it increases the returns to capital owners in that sector. Both reduce the share of labor income in GDP, which further leads to low shares of household consumption.[50]

Third, the government provides large subsidies to producers through suppressed factor prices. Capital is made cheap to privileged firms; resource prices are lower than in many countries; labor standards are laxly enforced; land is sold under the cost; and the environment is grossly underpriced. Adding up the subsidies implied by those distortions, the total subsidies can be as high as 10% of GDP (Huang and Tao, forthcoming; Huang and Wang, 2010).

In summary, the Chinese government is qualified as a production government (Yao, 2011), i.e. a government that puts paramount efforts and resources into the production process and in the meantime cares much less about the improvement of citizens' welfare. The upside of such a government is that the national economy grows fast; its downside is that it facilitates an ongoing process of wealth transfer from ordinary citizens to producers and capital owners. This is yet another example of the paradoxical consequences of the Chinese government's spearheaded efforts to achieve the paramount goal of growth. Economic growth is thought to be crucial for the CCP's legitimacy, so the party has spent every effort to ensure fast economic growth in the country. However, to achieve

[50] Chen and Yao (2011) do not find that government infrastructural investment has any effect on the household saving rate; nor do they find that it has any effect on the level of per-capita consumption.

this single-minded goal, the party has unconsciously adopted policies that would turn its head on its very initial objective to gain political support from the general public. The reason for this paradox is exactly what Dahl (1991) has pointed out for guardianship: the guardians—even if they have the will to work for the people—do not have the capacity to understand the whole consequence of what they are doing. In its newly released report, *China 2030*, the World Bank has called for the Chinese government to seriously take actions to further reform the country's SOEs.[51] But this is not enough. To push things onto the right track, serious reforms, including some form of political reform, have to be taken on the government itself.

7.6.6 The Exchange Rate and China's External Imbalances

One of the contentious issues that China's extraordinary growth of trade surplus has brought about is China's exchange rate policy. Before 1994, when the official and market exchange rates were unified, the Chinese yuan had gone through a period of devaluation to correct its overvaluation in the period of command economy. Between 1994 and 2005, the yuan was effectively pegged to the dollar at 8.25 yuan to one dollar. Under the pressures from the United States and other countries, the yuan began to appreciate from June 2005 and by August 2008 its value had gained 23% against the dollar. The yuan stopped its pace of appreciation for about 2 years and then started again in June 2010. By the end of 2011, it had gained another 8.4% against the dollar. That is, the yuan appreciated against the dollar by 31.4% between June 2005 and the end of 2011. Adding the gap of inflation between China and the US, the real appreciation of the yuan against the dollar was about 36% in this time period. However, some scholars believe that this pace of appreciation is far from enough to restore a sustainable level of current accounts in China and the United Sates. This view has been repeatedly articulated by William Cline and John Williamson of the Peterson Institute of International Economics in a series of the institute's policy briefs (Cline and Williamson, 2008, 2009, 2010, 2011). For example, they claimed in their 2011 policy brief that the yuan needed an upward revaluation of 28.5% against the dollar to bring China's current account down to 3% in GDP, a level they impose on every country as sustainable in the long run.

There are many ways to estimate the so-called equilibrium exchange rate for a currency (Isard, 2007), among which the macroeconomic balance (MB) method and the purchasing power parity (PPP) method, adjusted for the Balassa-Samuelson and Penn effects, are the two most popular. The MB method had been used by the IMF until recently and is based on the following identity (Isard, 2007).[52]

Current account ≡ the equilibrium level of current account.

[51] For an executive summary and the full text of the report, see http://www.worldbank.org/en/news/2012/02/27/china-2030-executive-summary.

[52] In 2012, the IMF introduced a new method called external balance assessment method. By this method, the IMF team directly estimates a determination function for the current account (or real effective exchange rate). See http://www.imf.org/external/np/res/eba/index.htm.

The left-hand side of the identity is represented by a country's underlying current account (UCUR) position which is assumed to be a function of the exchange rate as well as other macroeconomic variables. The right-hand side of the identity is independent of the exchange rate and estimated by a country's autonomous level of capital account, a sustainable position of its net foreign assets, or its equilibrium net domestic savings (i.e. national savings minus domestic investment). Both sides of the identity can be country-specific.

Cline and Williamson use a variant of the MB method. Instead of estimating the equilibrium level of current account, they impose a uniform level of 3% of GDP that they believe is sustainable in the long run. In addition, they do not estimate a country's UCUR, but instead rely on the IMF medium-term projection to determine a country's would-be level of current account if no exchange rate adjustment happens. However, it is hard to defend why 3% of GDP is a sustainable level of the current account for all the countries. Moreover, the IMF projections are often wrong. For example, it predicted that China's current account would be 5.7% of GDP in 2011 (Cline and Williamson, 2011; Table 1), but China's actual figure was only 4%. Finally, the elasticity of the current account with respect to the exchange rate is often an issue of debate.

In a survey article, Dunaway et al. (2006) have reviewed a set of studies that use the MB method to gauge the yuan's equilibrium exchange rate. They find that those studies have reached very different estimates because they adopt different methods to estimate China's equilibrium level of current accounts and their estimations of China's UCUR arrive at very different exchange rate elasticities. Dunaway et al. (2006) also find that relatively small perturbations in the estimation method would lead the estimated undervaluation of the yuan to change by up to 23% points.

The PPP method adjusted for the Balassa-Samuelson and Penn effects focuses on longer-term equilibrium exchange rates. According to the theory of PPP, the relative values of two currencies in the long run should be equal to the inverse of the two countries' price levels. The Balassa-Samuelson effect adds the impact of the relative price of tradable and non-tradable goods inside a country. However, empirical supports to the effect have been mixed (Isard, 2007). On the other hand, the Penn effect—i.e. higher-income countries tend to have higher real exchange rates—is strongly supported by cross-sectional data, and this regularity provides a convenient way to estimate a country's equilibrium real exchange rate (Isard, 2007). However, the PPP method does not fare better than the MB method when it comes to estimating China's equilibrium exchange rate; its results are very sensitive to the inclusion and exclusion of certain variables (Dunaway et al. 2006). One of the problems is that it imposes the same coefficients for all the countries although the Balassa-Samuelson and the Penn effects may differ by a country's stage of economic development.

One of the main channels for the Balassa-Samuelson and the Penn effects to work is through rising domestic price levels. However, in a developing country with large amounts of underutilized resources, especially human resources, this channel may be substantially weakened. The reason is that the growth of the export sector brings out the underutilized resources, which in turn suppress the growth of the domestic price level. Taking this idea forward, Wang and Yao (2009) estimate an equation of exchange rates based on the PPP method adjusted for the Penn effect using a panel data set of 186 countries and regions for the period 1960–2004. The novelty of their estimation is to interact a country's relative per-capita GDP with its share of rural population. While the coefficient of the relative per-capita GDP remains significantly positive, meaning that the Penn effect holds, the coefficient of the interactive term is significantly negative for medium and low-income countries but is insignificant for high-income countries. That is, the Penn effect is significantly weaker when a country is still experiencing structural changes. Using the estimates obtained from their main regression, Wang and Yao (2009) calculate the elasticity of the Penn effect for China in the period 1994–2009 and find that it is 0.27, which is exactly the elasticity of China's real exchange rate with respect to its relative income in that period. They also find that the yuan's nominal value was undervalued by 6.5% against the dollar in June 2008, much lower than other findings.

The link between development stages and the strength of the Balassa-Samuelson and Penn effects has an implication for one to interpret the role of the fixed exchange rate regime (FERR) for China's economic growth and external imbalances. In the medium and long run, the exchange rate regime would affect a country's external balances only if the Balassa-Samuelson and Penn effects did not fully work because if they did, then the country's competitiveness would be adjusted down and its external balances would be restored. Therefore, if China's FERR has contributed to China's economic growth and external imbalances, as many believe, it would be because structural change has weakened the two effects in the country. That is, the root cause is still economic fundamentals. That may be why international evidence provides mixed results for the relationship between a country's exchange rate regime and its current account balances. For example, Chinn and Wie (2008) compile a large data set of over 170 countries for the period 1971–2005 and carefully study whether exchange rate regime flexibility affects a country's current account reversion (i.e. from a surplus to a deficit or from a deficit to a surplus). Their conclusion is that there is no strong, robust, or monotonic relationship between exchange rate regime flexibility and the rate of current account reversion.

In summary, serious internal and external imbalances have developed in the Chinese economy since the end of the 1990s. Their causes are related to China's demographic and economic fundamentals, a weak financial sector, and government distortions. There is a deficit of research on those causes. In particular, we do not have a good understanding of why the household saving rate has increased dramatically, how government behavior has exacerbated the imbalances, whether China's external imbalances would persist, and what steps China should take to rebalance its economy. China's imbalance problems will be unlikely to disappear very quickly; more research is needed if those problems are to be corrected.

7.7. INEQUALITY AND THE MIDDLE-INCOME TRAP

China used to be one of the countries with the most equal income distribution; in 1981 its Gini coefficient of per-capita income was only 0.29 (Cheng, 2007). Thirty years after, the coefficient has reached 0.48 by official data (Yin and Liu, 2011). Related to rising income inequality there is a worry that China would be falling into the middle-income trap, a situation in which a country stops its catch-up process when it reaches the level of middle income. International experiences, especially those of Latin American countries and some Southeast Asian countries, have shown that the middle-income trap often goes hand in hand with high-income inequality. The worry about China is thus warranted against the rising income inequality. This section will provide a selective review of China's income inequality and its causes. In addition to reviewing the results using the official data, we will also report results based on the data provided by the first wave of the China Family Panel Studies (CFPS), an independent national survey managed by Peking University. Studies on the middle-income trap have been scant; this section will only point out several directions in which research can be carried out in order to enrich our understanding of the issue.

7.7.1 Facts of Income Inequality

The NBS began to carry out household surveys in the early 1980s, providing data for researchers to compile a time series of China's income Gini coefficients. Figure 7.28 presents one set of such estimations for rural areas, urban areas, and the country as a whole, respectively, for the period 1981–2007. Except in the mid-1990s, income inequality increased in both the city and the countryside, as well as the whole country over the period of 26 years. The countryside has been more unequal than the city, mostly because the rural population is more diverse than the urban population in terms of occupation, stock of wealth, and human capital. By 2007, the Gini coefficient of the whole country reached 0.45. The number then was increased to 0.48 in 2009 (Yin and Liu, 2011).

The official data, however, do not reflect the income of the highest percentiles. Li and Luo (2011) have tried to remedy this deficiency. Their starting point is that the NBS data do not cover two groups of high-income people. One comprises those who in theory should be covered by the NBS survey but in reality are not because they often refuse to participate. The other is the group of extremely rich people who are not likely to be covered by any household survey at all. For people in the first group, Li and Luo use the executives of listed companies as a sample for them. The income of this sample of people can be found from the information made open by the listed companies. For people in the second group, Li and Luo first obtain the stocks of wealth of the 868 richest persons from the Forbes and Hurun 2007 lists of China's richest persons.[53] Assuming a rate of

[53] The Hurun List of China's Richest Persons has been compiled independently by Rupert Hoogewerf since 1999. Huren is his Chinese name.

return of 5%, they then estimate the annual income of those people in 2007. Finally, they estimate the country's whole income distribution by assuming that the income of the above two groups of people and the people covered by the NBS survey follow a Pareto distribution. Using the NBS definition of income, they find that China's overall Gini coefficient was 0.53 in 2007.[54]

The CFPS provides even higher Gini coefficients than Li and Luo (2011)'s adjusted figures.[55] According to the household data provided by the CFPS, in 2010 the Gini coefficient was 0.556 for the whole country, 0.488 for the countryside, and 0.513 for the city (Shen and Lei, 2011).[56] That is, by the CFPS data, China has entered the rank of the most unequal countries in the world.

The inequality is the most pronounced between urban and rural residents. Figure 7.29 shows the ratio of urban per-capita disposable income and rural per-capita net income for the period 1985–2011 using official data. In the early 1980s, the rural-urban income gap declined substantially because of the rural reform. By 1985, urban income was only barely above 1.8 times of rural income. Since then, the gap increased steadily until the mid-1990s when inflation drove up the prices of agricultural products more than those of other products. The gap began to rise again in 1997. It reached the peak of 3.3 times in the period 2007–2010. This is by far the highest in the world. The encouraging sign is

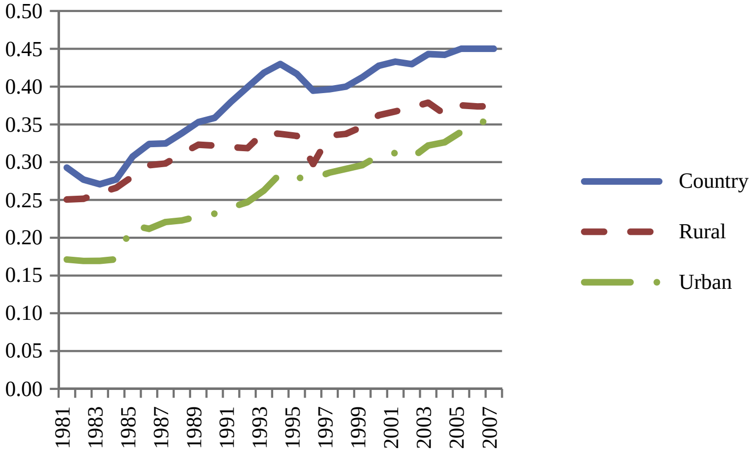


Figure 7.28 Gini Coefficients of per-capita income: 1981–2007. *Source:* Cheng (2007).

[54] Their estimation of the Gini coefficient without adjusting for the richest percentiles is 0.48, higher than that reported by Cheng (2007).

[55] CFPS is a nationally representative survey modeled on the Panel Studies of Income Dynamics (PSID). The first wave of survey, done in 2010, covered 25 provinces, 107 districts/counties, 424 villages/communities, 9500 families, and 21,760 adults.

[56] One potential problem with the CFPS is that it might have oversampled poor households because they might be more cooperative in the survey. This problem will be checked in the next round of survey scheduled for 2012.

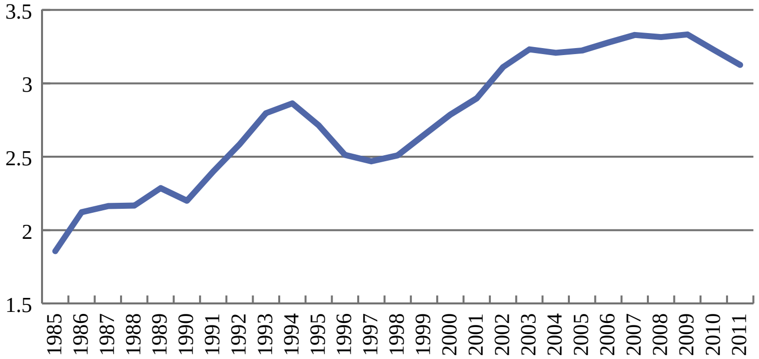


Figure 7.29 Urban-rural income gap: 1985–2011. *Source: NBS at www.stats.gov.cn. The urban-rural gap is defined as the ratio of urban per-capita disposable income over rural per-capita net income, both in current prices.*

that it declined in 2011. However, this decline was probably also due to the faster growth of agricultural prices than other prices. The CPI was 5.4%, but food prices rose by 11.8% in the year.[57]

There is a problem of how to account for the income of migrant workers when the urban-rural income gap is considered. The official statistics treat migrant workers who stay in the city for more than 180 days in a year as urban residents (but excluding household heads). This may underestimate the income brought back home by migrant workers. According to the CFPS that had a looser definition of residency status in its 2010 survey, the urban-rural income gap is 2.5 times for 2010 (Shen and Lei, 2011). The urban per-capita income in CFPS is almost the same as the figure provided by the NBS, which are ¥18,428 and ¥18,858 in 2010, respectively. However, the rural per-capita income in CFPS is much higher than that provided by the NBS. The most significant difference rests in transfer income. In CFPS, it is ¥1315; in NBS, it is only ¥398. In total, rural per-capita income is ¥6421 according to the CFPS, but ¥5153 according to the NBS (Shen and Lei, 2011).

Notwithstanding increasing income disparities, the size of the middle class has increased steadily, especially in the city. Using four waves of data provided by the China Health and Nutrition Survey, Liu et al. (2009) compare the distributions of household per-capita income in 1991, 1997, 2000, and 2006. Three significant findings emerge from their comparison. First, income distributions in the city, the countryside, and the country as a whole have all moved toward the higher end in a fashion that is consistent with the first-order stochastic dominance. That is, household income has been uniformly improved. Second, consistent with rising income inequality, the distributions

[57] The NBS Annual Statistical Report at http://www.stats.gov.cn/tjgb/ndtjgb/qgndtjgb/t20120222_402786440.htm.

of income have become more dispersed. Third, the proportion of households falling into the medium-range of income has increased. In the city, the mode of the distribution has been gradually replaced by a continuum of income. That is, the size of the middle class has been increasing. It waits for a more careful study to show how much of increasing income inequality has been caused by a larger, but more diverse middle class.

7.7.2 Causes of Rising Income Inequality

When thinking about the causes of income inequality in China, one has to realize that China is a large and highly diverse country to begin with. People have sufficiently heterogeneous income capabilities that would have resulted in unequal income distribution. Market supporters would argue that people would improve their income capabilities in the long run in response to the market. However, this would happen only when the market works seamlessly, which is utopian in any country. Government actions are often needed to ensure more equal income distribution. In this regard, there are large deficits on the part of the Chinese government.

One of them is the restriction on population movement. It is noteworthy that the urban-rural income gap was already 2.78 times in 1978, largely due to the separation of the city and the countryside set by the *hukou* system. Since 2003, restrictions on labor mobility have been abolished, but the *hukou* system is still preventing migrants from permanently settling in the city. As a result, more people are forced to stay in the countryside and a wedge is created between urban and rural income.[58]

Another deficit is the lack of social security in the countryside. While pension and health care coverage has reached 60% in the city, pensions have just started and health care is quite preliminary in the countryside. In addition, the city runs a reasonable subsistence maintenance program for low-income families whereas such programs barely exist in the countryside.

A third deficit is the lack of government actions to remove the barriers to more equal income distribution in the areas of production, redistribution, and regulation. Indeed, many of those barriers are set by the government itself. In the area of production, as our review in Section 7.6.5 has shown, the government provides substantial subsidies to producers and supports capital-intensive industries more than labor-intensive industries, causing the share of household income in the national GDP to decline. In the area of redistribution, there is no consistent government plan geared toward a more equal income distribution although the central government has increased its efforts to help

[58] The government announced a new *hukou* policy on February 24, 2012. For people living in small cities and towns, they can choose to become local residents as long as they have a stable job and housing, including rented homes; and for people living in medium-sized cities, they can do so after they have worked and lived there for three consecutive years. Large cities, however, maintain the current *hukou* policy.

rural residents and inland provinces.[59] The government's redistribution policy is seriously constrained by its desire to concentrate government revenues to invest in infrastructure, research, and projects that would generate faster current economic growth. In the area of regulation, government policies hinder the ability of ordinary citizens to take a share of economic growth in an equitable way. As we showed in Section 7.6.4, heavy regulation on the financial sector has resulted in regressive transfers of wealth from ordinary citizens to corporations and prevented ordinary citizens from benefiting from future economic growth. Government protection of the SOEs, especially those in the monopolistic sectors such as oil, telecom, and finance, has raised the income of those who have the privilege to work in those sectors. On the other hand, the government's loose implementation of the labor standards has suppressed the income of migrant workers.

These deficits have allowed the population to become more diverse in income capabilities. The fundamental cause of these deficits is related to the production nature of the Chinese government at this stage. For most officials, the primary task of the government is to make the pie bigger, not to decide how to divide it. What they have not realized is that the way the pie is divided matters for its growth. The discussion of the middle-income trap may ring a bell, though.

7.7.3 The Middle-Income Trap

The notion of the middle-income trap was made popular by a World Bank report *An East Asian Renaissance: Ideas for Economic Growth* (Gill and Kharas, 2008) and the authors' other writings (e.g. Kharas and Kohli, 2011). It refers to the situation in which a country fails to continue its catch-up process when its per-capita income has reached the middle-income level. According to Kharas and Kohli (2011), it happens when a middle-income country is not able to compete with either low-wage economies or highly skilled advanced economies. One of the examples of the trap is the Soviet Union and other former socialist countries in Eastern Europe. Their economies stopped to grow when their per-capita income reached a quarter of the American level. Latin American countries and some Southeastern countries are also believed to have experienced the trap. Table 7.7 is the income transition matrix of 112 countries between 1980 and 2009 using the World Bank categorization of income groups.[60] Among the 112 countries, 71 were qualified as middle-income countries (lower-middle and higher-middle income countries together) in 1980. Only eight of them, all of which were higher-middle income countries in

[59] In addition to its regular rural poverty alleviation programs and urban subsistence maintenance programs, the central government has abolished the taxes levied on agriculture; made mandatory education free and provided subsidies to boarding students in the countryside; and increased unconditional transfers to inland provinces.

[60] The World Bank defines income groups by absolute income in current dollars, but revises upward when time gets by. For example, in 1980, a country was qualified as a low-income country if its per-capita GDP was less than 370 dollars, whereas in 2009, the bar was raised to 995 dollars.

Table 7.7 The income transition matrix in the world: 1980–2009

	Low income	Lower-middle income	Higher-middle income	High income	Total
Low income	11	4	0	0	15
Lower-middle income	12	22	12	0	46
Higher-middle income	0	2	15	8	25
High income	0	0	0	26	26
Total	23	28	27	34	112

Notes: The income groups follow the World Bank definition.
Source: PWT 6.0.

1980, moved to the high-income group by 2009. Among the 46 originally lower-middle income countries, 12 moved to the higher-middle income group, but another 12 dropped to low-income group. There were also two originally higher-middle income countries that moved down by one group. This means that the majority of the middle-income countries of 1980 did not manage to narrow their income gaps with the high-income countries. In this sense, they are trapped.

Kharas and Kohli (2011) point out two causes for a country to fall into the middle-income trap. First, low-income countries can maintain high growth rates by focusing on job creation, but this type of cheap growth is no longer possible for middle-income countries because underutilized human resources have already been depleted. A failure for a middle-income country to expand demand and to improve its total factor productivity would then lead the country to a middle-income trap. Second, international experiences show that the middle-income trap has often appeared together with high economic and social inequality. This is no more evident in the comparison between East Asia and Latin America. Except Hong Kong, East Asian economies have maintained fairly equal income distribution while they caught up with advanced economies. In contrast, Latin American countries have stagnated for thirty years while their income inequality remained high. Both causes are very pertinent to China.

As this review has shown, China's economic growth has been mainly driven by abundant labor supply offered by its demographic transition and rural-urban migration and large quantities of capital investment offered by its high national saving rates although TFP improvement has been substantial. This model of growth may be reaching its limits for several reasons. First, the growth of labor has begun to decelerate and China's total labor force may start to decline by 2020. In the meantime, structural change centered at labor movement from the countryside to the city will also reach its steady state soon. As a result, the period of cheap growth is approaching its end. Second, capital accumulation may not sustain future economic growth for two reasons. At the aggregate level, it will inevitably face the law of diminishing marginal returns if technological progress does not keep up the pace. At the structural level, capital investment has weak effects to generate demand.[61] Third, relying on investment for future growth has the tendency to reduce the share of households in the national income, as the evidence of Section 7.6.5 strongly attests. As a result, domestic demand cannot easily catch up with the pace of domestic supply, forcing China to continue relying on external demand. However, the extraordinary high growth of export in the period 2001–2008 was more likely to be the result of the one-shot effect of trade liberalization than an inherent part of China's normal growth trajectory. China needs to enhance its domestic demand to generate future growth. Fourth, due to the distortions in finance, investment-driven growth can also worsen the income distribution in the household sector. On the one hand, investment is heavily controlled by the government and geared toward highly protected sectors such as infrastructure, telecom, and finance; on the other hand, for investment not controlled by the government, banks favor large and capital-intensive companies. Either way, people working in those privileged sectors end up enjoying higher income than people working in other sectors.[62]

The adverse effects of inequality on economic growth have been well-established in the literature. The challenge to link inequality and the middle-income trap is to show why inequality is particularly bad for a country to escape the trap. More specifically, one has to explain why inequality does not hinder a country to reach the level of middle income but does prohibit it from attaining higher levels of income. In this regard, two explanations are pertinent in the case of China.

The first explanation is related to the size of the domestic market. Inequality limits the size of domestic market, but this may not be a serious constraint for economic growth when a country is poor because it can rely on export to grow. However, domestic markets would become more important when a country reaches the level of middle income because higher labor costs reduce its competitiveness in the world market. As a result, inequality can become detrimental to further growth in the country. China is a large country; its sheer size may render it problematic to rely on the world market for sustainable growth. In this regard, rising inequality can become a serious hindrance for the country to obtain higher income.

The second explanation is related to the stock and distribution of human capital. Empirical evidence shows that the return to education is not constant as educational

[61] This might be one of the causes leading to the collapse of the Soviet Union. In the country, capital investment was concentrated in the heavy industry, especially in the military industry. But the demand of the heavy industry for more investment is limited. Without a growing consumer goods sector, capital investment would inevitably hit the wall.

[62] Zhou et al. (2012) group Chinese industries into three sectors, labor-intensive, capital-intensive and resource-based, and study the wage inequality among them. Controlling the quality of labor, their decomposition finds that between-sector differentials have become more significant in explaining the wage inequality for the period 1993–2007.

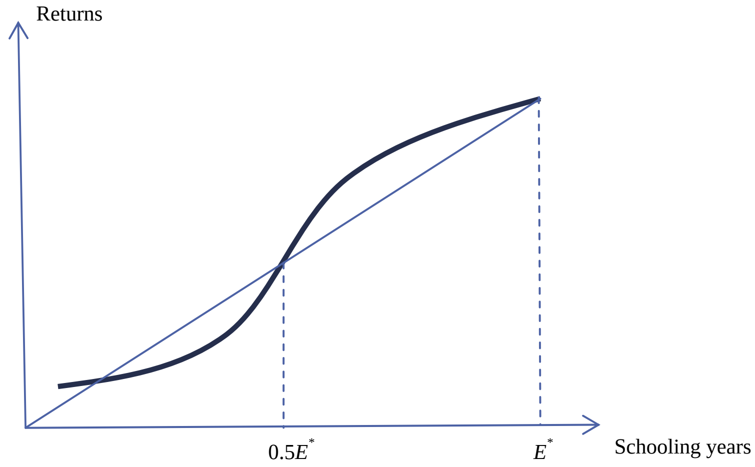


Figure 7.30 Returns to education.

attainment increases; it increases the fastest at the stage of high school and university education in China (Li et al. 2012). In general, it exhibits an S-curve as shown in Figure 7.30. Now think about the fact that the government has a fixed budget B to be allocated to the education of two persons. Suppose that this budget is enough to raise one person's education to E if it is all allocated to him. It is reasonable to believe that E increases in B. For the sake of simplicity, let us assume that E is a linear function of B. Then there exists a value of B, say B^*, such that allocating B to one person (so his educational attainment is E^* as shown in the figure and the other person gets zero education) yields higher aggregate returns than equal allocation (so each person gets education of $0.5E^*$) when B is less than B^* and the reverse is true when B is larger than B^*. It is possible to extend this argument to the whole population so that an unequal distribution of education helps economic growth when a country is relatively poor and the government does not have much resource to allocate, but a more equal distribution is more desirable when the country passes a certain level of income and the government has more resources to allocate. In reality, however, educational attainment is determined by both government support and individual decision. One important constraint for the latter is family income. In this regard, inequality can lead to slow growth of human capital because it discourages people at the lower end of the distribution from obtaining sufficient education. With more tax income after reaching the level of middle income, the government can help people from poorer families to obtain more education and boost economic growth.

The situation in China, however, is worrisome. Figure 7.31 presents the educational pyramids of urban and rural adults using data provided by the 2010 CFPS survey. It is clear from the figure that the gap between rural and urban areas persisted or became even

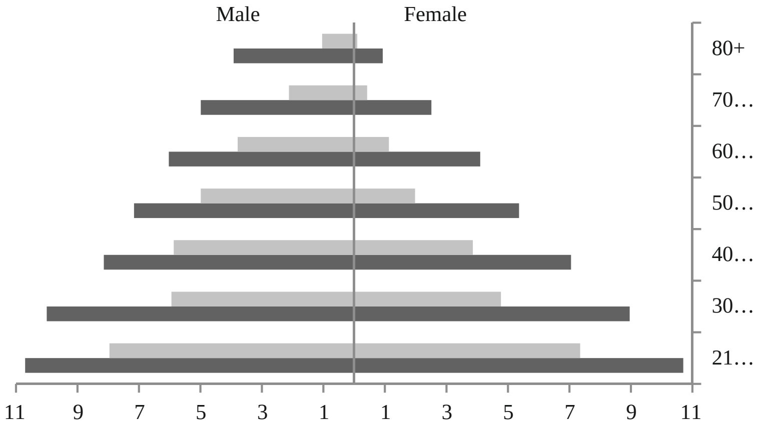


Figure 7.31 Educational pyramids in rural and urban areas (2010). *Notes:* The dark color bars are for urban areas, and the light color bars are for rural areas. Data are for 9357 adults in 25 provinces in 2010. *Source: CFPS 2010.*

larger from older age groups to younger age groups, although young people obtained more education than older people in both areas.

Table 7.8 presents the gap by gender and age group. The gap of males of above 80 years old was 2.88 years of schooling. It then declined until the group of 50–59 years old (who were born in the 1950s), but then increased in the next two age groups. In particular, the gap jumped to 4.06 years for people between 30 and 39 years old. Those people were born in the 1970s and mostly started their education in the late 1970s and early 1980s, a period when the rural reform took place. It seems that the reform took a toll on rural education. Schools were mostly financed by the budget of local governments including the village government. The dismantling of collective farming might have reduced school quality because of deteriorating village budgets. Another possible reason was that farm families wanted their kids to work earlier on their newly acquired land. The gap was narrowed to 2.74 years for people between 21 and 29.[63] But this gap was still between those of the age groups 60–69 and 70–79.

The situation for women was even worse. The urban-rural gap increased from the group of 80+ to the group of 30–39 except for the group of 40–49. This was because the educational attainment of urban women was increased much faster than rural women. As a matter of fact, in the youngest group, urban women had exactly the same amount of schooling as urban men. The urban-rural gap of the youngest group of women was still 3.35 years, higher than that of the group of 40–49 years of age, although it was smaller than that of group 30–39 years, which was the highest record of 4.19 years.

[63] Younger age groups are not included in the pyramids because they might have not completed their schooling by 2010.

Table 7.8 The urban-rural educational gap by gender

Age	Male	Female
21–29	2.74	3.35
30–39	4.06	4.19
40–49	2.27	3.19
50–59	2.17	3.38
60–69	2.24	2.97
70–79	2.87	2.09
80+	2.88	0.83

Source: Calculated from Figure 7.30.

A close look at the youngest group of people shows that the average urban youth was one year short of finishing high school whereas the average rural youth was one year short of finishing middle school. Both are inadequate for China to grow into a high-income country; rural education is more so. By the official estimate (see Section 7.3.3), China's raw college admission rate will be increased to 40% by 2020. However, if the education of the average person remains low, a bifurcation will be created among the next generation of people. In the 2020s, it is most likely that the major sources of employment will be medium-level manufacturing and services. This means that the bulk of the demand will be for workers who have medium-level technical trainings. The average rural youth will not be qualified for such jobs and the average urban youth will be barely qualified. To grow into a high-income country, China has to find ways to eliminate the gap created by the current trend of bifurcation.

In summary, the risks of China falling into the middle-income trap are likely to lie in China's current investment-driven growth model and the bifurcation of educational attainment in the population. There are high demands for China to change its growth model from both inside and outside the country; the Chinese government has also realized the need for the country to move to a more consumption-driven economy. The problem is that it may not have found the way to make the change happen. On the other hand, it is more complicated to stop the bifurcation of education. Income and wealth inequality has definitely played a role leading to the bifurcation, but it takes a long time to correct such inequality. Government commitment has a more direct effect. The *National Plan of Educational Reform and Development: 2010–2020* sets a high goal for the educational system to serve for an innovative society. As a result, formal education, university education in particular, is emphasized. More than that, elitist universities are given a priority among university education. Government programs (such as the 211 Project, the 985 Project, and the 2011 Programs) are geared toward providing large funds to a small number of elitist universities. While they have enhanced China's innovative capacities, they have also worsened the bifurcation of education in the society. To make a change, the Chinese government has to realize that its elitist approach to education is not consistent with the Chinese reality in the next 10–20 years.

7.8. CONCLUSIONS

This chapter has provided a comprehensive review of the causes and consequences of the Chinese growth miracle. Several conclusions can be drawn from the review. First, China's economic success is largely a result of following the standard recommendations of neoclassical economics including high savings and investment; technological progress; human capital accumulation; and macroeconomic stability. In addition, China had favorable initial conditions, noticeably a relatively high level of human development, a sound industrial base, and an economically and socially equal society, when its economy took off at the end of the 1970s. From this point of view, the Chinese growth miracle is not miraculous at all. Second, what is interesting about China is how it has managed high economic growth while transforming itself from a command economy to a mixed economy. The key to understanding this is the contingent institutions China has created along the way of transition and growth. Instead of transplanting institutions as they appear in advanced countries, China has adapted them to suit the political and economic constraints at the time of transplantation. As a result, the resulting contingent institutions were not pure, but achieved the most urgent goals at the time. Third, China's economic growth has been largely driven by investment and the manufacturing sector, and export has become one of the major growth drivers since China joined the WTO in 2001. In the meantime, serious internal and external imbalances have emerged. Those imbalances are likely to be the by-product of several fundamental forces moving the Chinese economy, particularly massive structural change including large rural-urban migration, abrupt demographic transition, and high growth itself. Several structural deficiencies have also played a role. Among them, underdevelopment of the financial sector and the government's pro-producer policies are the two most important. Fourth, inequality has risen fast and may have negative consequences for China to grow into a high-income country. In particular, it is contributing to the bifurcation of education in the Chinese society. To overcome the bifurcation, the Chinese government has to make a commitment to raising the educational level of the average person in the countryside as well as in the city.

The Chinese experience provides ample opportunities for economists to study economic growth, especially the political economy of economic growth. In this regard, several areas are particularly worth further exploration. The first is the incentive structure provided for government officials in a successful authoritarian regime. The literature has a relatively good understanding of the incentive structure under a successful democracy, but the study of authoritarian regimes has barely started in economics. Authoritarian regimes are more diverse than democracies and thus need more careful studies to understand their economic performance. The Chinese authoritarian regime is relatively more economically successful than other authoritarian regimes. In addition, China is a large country and the central government has to adopt delicate mechanisms to motivate local officials. It thus provides an interesting case for careful studies. The second area is the political economy of the Chinese growth model. While China's investment and export-led growth model has its roots in economic and historic fundamentals, it is not deniable that it is also reinforced by government policy. How can this model continue when it does not provide proportional gains to the public? The third area is how inequality is transformed into uneven distribution of political power and becomes a hindrance to economic growth. China started out being one of the most equal societies and has become one of the most unequal in the last 30 years. There are also signs that inequality is leading to the concentration of political power. How has this happened? Is it a rule for any authoritarian regime? The last area is the study of China's democratization process. Will China follow the prediction of the theory of social development as its per-capita income continues to rise? As this review has shown, China has succeeded economically, mainly because the country has adopted the standard growth recipe prescribed by neoclassical economics. That is, China is a normal country. There are enduring authoritarian regimes with high-income levels, but they are mostly monarchs. Will China follow other kinds of authoritarian regimes, especially those relying on bureaucratic rules, to democratize?

These areas do not comprise an exhaustive list of the interesting topics regarding China's growth story. But they are certainly the most important for China's future growth. China is still an unfolding story. It will remain an exciting source of academic research for social scientists.

ACKNOWLEDGMENTS

I thank Steven Durlauf, Shuanglin Lin, Thomas Rawski, Jinghai Zheng for their helpful comments and suggestions. Qinggang Liu, Guangjun Shen, Luanying Wu, and Jing Zhou have provided competent research assistance.

REFERENCES

Acemoglu, Daron, 2003. Why not a political coase theorem? Social conflict, commitment, and politics. Journal of Comparative Economics 31 (4), 620–652.

Acemoglu, Daron, Johnson, Simon, Robinson, James, 2006. Institutions as the fundamental cause of long-run growth. In: Aghion, Philippe, Durlauf, Steven (Eds.), Handbook of Economic Growth. North Holland, Amsterdam.

Ang, James, 2009. Household saving behaviour in an extended life cycle model: a comparative study of China and India. Journal Of Development Studies 45 (8), 1344–1359.

Bai, Chong-En, Chang-Tai, Hsieh, Yingyi, Qian, 2006. The return to capital in China. Brookings Papers on Economic Activity (2), 61–88.

Bai, Chong-En, Qian, Zhenjie, 2009. Who is squeezing out household income—an analysis of the national income distribution in China (Shui zai jizhan jumin de shouru—zhongguo guomin shouru fenpei geju fenxi). Social Sciences in China (zhongguo shehui kexue) (5), 99–115.

Bai, Chong-En, Wang, Dehua, Qian, Zhenjie, 2010. Several issues for public finance to speed up structural change (gonggong caizheng cujin jiegou zhuanbai de ruogan wenti). Comparative Studies (bijiao) (3).

Bai, Chong-En, Wu, Binzhen, 2011. Social insurance and household consumption in China. Paper presented in the 16th World Congress of the International Economic Association, August 2011, Beijing.

Bell, Daniel, 2010. China's New Confucianism: Politics and Everyday Life in a Changing Society. Princeton University Press, Princeton (April 2010).

Besley, Timothy, Kudamatsu, Masayuki, 2008. Make Autocracy Work. In: Helpman, Elhanan (Ed.), Institutions and Economic Performance. Harvard University Press, Cambridge (November 2008).

Bishop, John, Liu, Haiyong, 2008. Liberalization and rent-seeking in China's labor market. Public Choice 135 (3–4), 151–164.

Blanchard, Olivier, Kremer, Michael, 1997. Disorganization. Quarterly Journal of Economics 112 (4), 1091–1126.

Blanchard, Olivier, Shleifer, Andrei, 2001. Federalism with and without Political Centralization: China versus Russia. IMF Staff Papers 48, 171–79.

Blecker, Robert, Razmi, Arslan, 2010. Export-led growth, real exchange rates and the fallacy of composition. In: Setterfield, March (Ed.), Handbook of Alternative Theories of Economic Growth. Elgar, Northampton, Mass and Cheltenham, U.K., pp. 379–396, May 2010.

Bloom, David, Finlay, Jocelyn, 2009. Demographic change and economic growth in Asia. Asian Economic Policy Review 4 (1), 45–64.

Brandt, Loren, Ma, Debin, Rawski, Thomas, 2011. From divergence to convergence: re-evaluating the history behind China's economic boom. Manuscript.

Bremmer, Ian, 2011. The End of the Free Market: Who Wins the War between States and Corporations. Portfolio Trade, New York (September 2011).

Boycko, Maxim, Shleifer, Andrei, Vishny, Robert, 1997. Privatizing Russia. MIT Press, Cambridge, Mass (January 1997).

Brandt, Loren, Hsieh, Chang-tai, Zhu, Xiaodong, 2008. Growth and structural change in China. In: Brandt, Loren, Rawski, Thomas (Eds.), China's Great Transformation. Cambridge University Press, New York (April 2008).

Bromley, Daniel. 2009. Sufficient Reason: Volitional Pragmatism and the Meaning of Economic Institutions. Princeton University Press, Princeton (July 2009).

Bulte, Erwin, Damania, Richard, 2008. Resources for Sale: Corruption, democracy and the natural resource curse. B.E. Journal of Economic Analysis and Policy: Contributions To Economic Analysis And Policy 8 (1).

Caballero, Ricardo, Farhi, Emmanuel, Gourinchas, Pierre-Olivier, 2008. An equilibrium model of "global imbalances" and low interest rates. American Economic Review 98 (1), 358–393.

Cai, Fang, 2010. Demographic transition, demographic dividend, and Lewis turning point in China. China Economic Journal 3 (2), 107–120.

Cai, Fang, 2011. On the international experiences of structural adjustment (guanyu jiegou tiaozhen de guoji jingyan). Comparative Studies (bijiao) 51, 31–35.

Cai, Hongbin, Treisman, Daniel, 2007. Did government decentralization cause China's economic miracle? World Politics 58 (4), 505–535.

CCER, 2007. An estimation of the rate of return to capital in China (woguo ziben huibaolv guce). China Economic Quarterly (jingjixue jikan) 6 (3), 723–758.

Chamon, Marcos, D., Prasad, Eswar, S., 2010. Why are saving rates of urban households in china rising? American Economic Journal: Macroeconomics 2 (1), 93–130.

Chamon, Marcos, Liu, Kai, Prasad, Eswar, 2010. Income Uncertainty and Household Savings in China. IMF Working Papers: 10/289.

Chen, Binkai, Yao, Yang, 2011. The cursed virtue: government infrastructural investment and household consumption in Chinese provinces. Oxford Bulletin of Economics and Statistics 73 (6), 856–877.

Chen, Binkai, Yang, Rudai, Zhong, Ninghua, 2012. Housing reform and saving rate of China. Paper presented in Restructuring China's Economy, the 2012 ASSA Annual Meeting, January 5–8, 2012, Chicago.

Cheng, Yonghong, 2007. China's overall Gini coefficient and its decomposition by rural and urban areas since reform and opening (gaige yilai quanguo zongti jini xishu de yanbian jiqi chengxiang fenjie). Social Sciences in China (zhongguo shehui kexue) (4), 45–60.

Cheten, Ahya, et al., 2006. India and China: New Tigers of Asia. Mogen Stanley Research Report, April 2006.

Chinn, Menzie, Wei, Shang-Jin, 2008. A Faith-based Initiative: Does a Flexible Exchange Rate Regime Really Facilitate Current Account Adjustment? NBER Working Papers: 14420.

Cline, William, Williamson, John, 2008. New Estimates of Fundamental Equilibrium Exchange Rates. Peterson Institute of International Economics Policy Brief, Number PB 08–7.

Cline, William, Williamson, John, 2009. 2009 Estimates of Fundamental Equilibrium Exchange Rates. Peterson Institute of International Economics Policy Brief, Number PB 09–10.

Cline, William, Williamson, John, 2010. Currency Wars? Peterson Institute of International Economics Policy Brief, Number PB 10–26.

Cline, William, and Williamson, John, 2011. Estimates of Fundamental Equilibrium Exchange Rates, May 2011. Peterson Institute of International Economics Policy Brief, Number PB 11–5.

Curtis, Chadwick C., Lugauer, Steven, Mark, Nelson C., 2011. Demographic Patterns and Household Saving in China. NBER Working Papers: 16828.

Dahl, Robert, 1991. Democracy and Its Critics. Yale University Press, New Heaven (July 1991).

Devadason, Evelyn, 2011. Reorganization of intra-ASEAN 5 trade flows: the "China factor". Asian Economic Journal 25 (2), 129–149.

Dunaway, Steven, Leigh, Lamin, Li, Xiangming, 2006. How Robust are Estimates of Equilibrium Real Exchange Rates: The Case of China. IMF Working Paper WP/06/220.

Engel, Charles, Rogers, John, 2006. The U.S. current account deficit and the expected share of world output. Journal of Monetary Economics 53 (5), 1063–1093.

Enikolopov, Ruben, Zhuravskaya, Ekaterina, 2007. Decentralization and political institutions. Journal of Public Economics 91 (11–12), 2261–2290.

Feenstra, Robert, 2011. How Big Is China? Yanfu Memorial Lecture in Economics, China Center for Economic Research, Peking University (June 2011).

Fung, K.C., Lau, Lawrence J., Xiong, Yanyan, 2006. Adjusted estimates of United States-China bilateral trade balances: an update. Pacific Economic Review 11 (3), 299–314.

Gandhi, Jennifer, 2008. Political Institutions under Dictatorship. New York, Cambridge University Press.

Garnaut, Ross, 2010. Macro-economic implications of the turning point. China Economic Journal, 3 (2), 181–190.

Garnaut, Ross, Song, Ligang, Tenev, Stoyan, Yao, Yang, 2005. Ownership Transformation in China. The World Bank, Washington, DC (July 2005).

Gehlbach, Scott, Keefer, Philip, 2008. Investment without Democracy: Ruling-party Institutionalization and Credible Commitment in Autocracies. Manuscript. Department of Political Science, the University of Wisconsin-Madison and the World Bank, May 2008.

Gerschenkron, Alexander, 1962. Economic Backwardness in Historical Perspective, A Book of Essays. Belknap Press of Harvard University Press, Cambridge, Massachusetts (January 1962).

Gill, Indermit, Kharas, Homi, 2008. An East Asia renaissance: Ideas for economic growth. World Bank, Washington DC.

Glaeser, Edward, La Porta, Rafael, Lopez-de-Silanes, Florencio, Shleifer, Andrei, 2004. Do institutions cause growth? Journal of Economic Growth 9 (3), 271–303.

Han, Li, 2007. Marketing Politics? Economic Reforms and the Selection of Political Elites in China. Manuscript. Division of Social Sciences, Hong Kong University of Science and Technology,

He, Daxing, Yao, Yang, 2011. Social equality, the disinterested government and economic growth in China (shehui pingdeng, zhongxing zhengfu yu zhongguo jingji zengzhang). Economic Research Journal (jingji yanjiu) (1), 4–17.

Horioka, Charles, Wan, Junmin, 2007. The determinants of household saving in China: a dynamic panel analysis of provincial data. Journal of Money, Credit, and Banking 39 (8), 2077–2096.

Hu, Yunzhi, Yao, Yang, 2012. Rents, Elitism, and the Returns to the CCP Membership. Manuscript, China Center for Economic Research, Peking University.

Huang, Yiping, Tao, Kunyu, 2010. Factor market distortion and the current account surplus in China. Asian Economic Papers 9 (3), 1–36.

Huang, Yiping, Wang, Bijun, 2010. Cost distortions and structural imbalances in China. China & World Economy 18 (4), 1–17.

IFTE (Institute of Finance and Trade Economics), 2010. China Fiscal Policy Report 2009/2010, <http://cmsold.cass.cn/showNews.asp?id=30783>. September 10, 2010.

Isard, Peter, 2007. Equilibrium Exchange Rates: Assessment Methodologies. IMF Working Paper WP/07/296.

Jin, Hehui, Qian, Yingyi, Weingast, Barry R., 2005. Regional decentralization and fiscal incentives: federalism, Chinese style. Journal of Public Economics 89 (9–10), 1719–1742.

Ju, Jiandong, Wei, Shang-Jin, 2010. Domestic institutions and the bypass effect of financial globalization. American Economic Journal: Economic Policy 2 (4), 173–204.

Kaldor, Nicholas, 1957. A model of economic growth. The Economic Journal 67 (268), 591–624.

Kharas, Homi, Kohli, Harinder, 2011. What is the middle income trap, why do countries fall into it, and how can it be avoided? Global Journal of Emerging Market Economies 3 (3), 281–289.

Kim, Jong-Il, Lau, Lawrence, (1996). The sources of Asian Pacific economic growth. The Canadian Journal of Economics 29 (special issue, part 2), S448–S454.

Knight, John, Deng, Quheng, Li, Shi, 2011. The puzzle of migrant labour shortage and the rural labour surplus in China. China Economic Review 22 (4), 585–600.

Koopman, Robert, Wang, Zhi, Wei, Shang-Jin, 2012. Estimating domestic content in exports when processing trade is pervasive. Journal of Development Economics 99 (1), 178–189.

Kuijs, Louis, 2006. How Will China's Saving-investment Balance Evolve? World Bank Policy Research Working Paper #3958.

Lardy, Nicholas, 1998. China's Unfinished Economic Revolution. Brookings Institution Press, Washington DC (July 1998).

Lau, Lawrence, Qian, Yingyi, Roland, Gerard, 2000. Reform without losers: an interpretation of China's dual-track approach to transition. Journal of Political Economy 108 (1), 120–143.

Lau, Lawrence, et al., 2007. Non-competitive input-output model and its application: an examination of the China-U.S. trade surplus. Social Sciences in China 2007 (5), 91–103.

Lee, Keun, 2008. The BeST Consensus. Paper presented in the Shanghai Forum, May 2008.

Lewis, W. Arthur, 1954. Economic Development with Unlimited Supplies of Labor. Manchester School of Economic and Social Studies, vol. 22, pp. 139–191.

Li, Bobai, Andrew, Walder, 2001. Career advancement as party patronage: sponsored mobility into the Chinese administrative elite, 1949–1996. American Journal of Sociology 106 (5), 1371–1408.

Li, Daokui, Lin, Linlin, Wang, Hongling, 2009. The U-shape law of the labour share in GDP (GDP zhong laodong fen'e yanbian de U xing guilu). Economic Research Journal (jingji yanjiu) (1), 70–82.

Li, Hongbin, Zhou, Li-An, 2005. Political turnover and economic performance: the incentive role of personnel control in China. Journal of Public Economics 89 (9–10), 1743–1762.

Li, Hongbin, Liu, Pak Wai, Ma, Ning, Zhang, Junsen, 2007. Economic returns to communist party membership: evidence from Chinese twins. Economic Journal 117 (523), 1504–1520.

Li, Hongbin, Liu, Pak Wai, Zhang, Junsen, 2012. Estimating returns to education using twins in urban China. Journal of Development Economics 97 (2), 494–504.

Li, Shi, Luo, Chuliang. 2011. How unequal is China? Economic Research Journal (jingji yanjiu) (4), 68–79.

Li, Yining, 2012. Economic Reform and Development in China. Cambridge University Press, Cambridge (January 2012).

Lin, Justin, Liu, Zhiqiang, 2000. Fiscal decentralization and economic growth in China. Economic Development and Cultural Change 49 (1), 1–21.

Lin, Justin, Yao, Yang, 2001. Chinese rural industrialization in the context of the east Asian miracle. In: Stiglitz, Joseph, Yusuf, Shahid (Eds.), Rethinking the East Asian Miracle. The World Bank and Oxford University Press, Washington DC (June 2001).

Lin, Justin, Cai, Fang, Li, Zhou, 2003. The China Miracle. The Chinese University Press, Hong Kong (July 2003).

Lin, Justin, 2009. Economic Development and Transition: Thought, Strategy, and Viability. Cambridge University Press, Cambridge (March 2009).

Liu, Jing, Zhang, Juwei, Mao, Xuefeng, 2009. The dynamics of China's income distribution between 1991 and 2006 (zhongguo 1991–2006 nian shouru fenbu de dongtai bianhua). Journal of World Economy (shijie jingji) 2009 (4), 3–13.

Liu, Qinggang, Yao, Yang, 2012. Differential Growth Rates and Global Imbalances. Manuscript. CCER, Peking University.

Lu, Feng, Liu, Liu, 2007. Productivity growth in manufacturing and services and an international comparison (1978–2005) (woguo liang bumen laodong shengchanlu zengzhang ji guoji bijiao 1978–2005). China Economic Quarterly (jingjixue jikan) 6 (2), 357–380.

Ma, Shuang, Zang, Wenbin, Gan, Li, 2010. The new rural cooperative medical insurance and household food consumption (xinxing nongcun hezuo yiliao baoxian dui nongcun jumin shiwu xiaofei de yingxiang fenxi). China Economic Quarterly 10 (1), 249–270.

Maddison, Angus, 2001. The World Economy: A Millennial Perspective. OECD, Brussels (June 2001).

Mao, Rui, Yao, Yang, 2012. Structural change in an open economy. Pacific Economic Review 17 (1), 29–56.

Mao, Rui, Jianwei, Xu, Yao, Yang, 2012. Finance-Manufacturing Comparative Advantage and Global Imbalances. Manuscript. China Center for Economic Research, Peking University.

Mendoza, Enrique G., Quadrini, Vincenzo, Rios-Rull, Jose-Victor, 2009. Financial integration, financial development, and global imbalances. Journal of Political Economy 117 (3), 371–416.

Modigliani, Franco, Cao, Shi, 2004. The Chinese saving puzzle and the life-cycle hypothesis. Journal of Economic Literature 42 (1), 145–170.

Morris, Ian, 2011. Why the West Rules—for Now. Farrar, Straus and Giroux, New York (October 2011).

Murphy, Kevin, Shleifer, Andrei, Vishny, Robert, 1992. The transition to a market economy: pitfalls of partial reform. Quarterly Journal of Economics 107 (3), 889–906.

Nee, Victor, 1989. A theory of market transition: from redistribution to markets in state socialism. American Sociological Review 54 (5), 663–681.

North, Douglass, Thomas, Robert, 1973. The Rise of the Western World: A New Economic History. Cambridge University Press, Cambridge.

Opper, Sonja, Brehm, Stefan, 2007. Networks versus Performance: Political Leadership Promotion in China. Department of Economics, Lund University (July 2007).

Park, Donghyun, Shin, Kwanho, 2009. The People's Republic of China as an Engine of Growth for Developing Asia?: Evidence from Vector Autoregression Models. ADB Economics Working Paper Series No. 175.

Qian, Yingyi, Weingast, Barry, 1997. Federalism as a commitment to preserving market incentives. Journal of Economic Perspectives 11 (4), 83–92.

Qian, Yingyi, Rolan, Gerard, Chenggang, Xu, 2006a. Coordination and experimentation in M-form. Journal of Political Economy 114 (2), 366–402.

Qian, Yingyi, Roland, Gerard, Chenggang, Xu, 2006b. Coordination and experimentation in M-form and U-form organizations. Journal of Political Economy 114 (2), 366–402.

Ramo, Joshua, 2004. The Beijing Consensus: Notes on the New Physics of Chinese Power. Foreign Policy Centre. <http://www.fpc.org.uk/fsblob/244.pdf>, May 2004.

Rao, M. Govinda, Singh, Nirvikar, 2004. The Political Economy of India's Federal System and its Reform. Santa Cruz Center for International Economics, Department of Economics, University of Carlifornia - Santa Cruz, April 2004. <http://escholarship.org/uc/item/4gc7c4px>.

Rawski, Thomas, 2001. What is happening to China's GDP statistics? China Economic Review 12 (4), 347–354.

Rawski, Thomas, 2011. Human resources and China's long economic boom (renli ziyuan yu zhongguo changqi jingji zengzhang). China Economic Quarterly (jingjixue jikan) 10 (4), 1153–1186.

Rodriguez-Pose, Andres, Ezcurra, Roberto, 2011. Is fiscal decentralization harmful for economic growth? Evidence from the OECD countries. Journal of Economic Geography 11 (4), 619–643.

Rodrik, Dani, 2006. What's so special about China's exports? China And World Economy 14 (5), 1–19.

Rosen, Daniel, 2011. The role of the state in China's economy. Brookings Institute, March 1, 2011.

Sachs, Jeffrey, Woo, Wing Thye, 1994. Structural factors in the economic reforms of China, eastern Europe, and the former soviet union. Economic Policy 9 (18), 101–145.

Sen, Amartya, 1966. Peasants and dualism with and without surplus labor. Journal of Political Economy 74 (5), 425–450.

Schott, Peter K., 2008. The relative sophistication of Chinese exports. Economic Policy 53, 5–49 (January).

Shen, Yan, Yao, Yang, 2009. CSR and Competitiveness in China. Foreign Languages Press, Beijing, August 2009.

Shen, Yan, Lei, Xiaoyan, 2011. Report of Data Cleaning for CFPS, 2010. Manuscript, China Center for Economic Research. Peking University, November 2011.

Song, Zheng, Storesletten, Kjetil, Zilibotti, Fabrizio, 2011. Growing like China. American Economic Review 101 (1), 196–233.

Sugihara, Kaoru, 2003. The east Asian path of economic development: a long-term perspective. In: Arrighi, Giovanni, Hanashita, Takeshi, Selden, Mark (Eds.), The Resurgence of East Asia: 500, 150 and 50 Year Perspectives. Routledge, London, July 2003.

Tan, Zhibo, Wei, Shang-Jin, Yao, Yang, Zhao, Yue, 2012. Financial Structure, Corporate Savings and Global Imbalances. Manuscript, China Center for Economic Research, Peking University.

Valli, Vittorio, Saccone, Donatella, 2008. Structural change and economic development in China and India. The European Journal of Comparative Economics 6 (1), 101–129

Wang, Shaoguang, Hu, Angang, 2001. The Chinese Economy in Crisis: State Capacity and Tax Reform. M E Sharpe Inc., New York (April 2001).

Wang, Xianbing, Xu, Xianxiang, 2008. Sources, attrition, and tenure of local leaders and economic growth: evidence from Chinese provincial governors and party secretaries (difan guanyuan laiyuan, quxiang, renqi he jingji zengzhang: laizi zhongguo shengzhang shengwei shuji de zhengju). Management World (guanli shijie) (4), 16–26.

Wang, Zetian, Yao, Yang, 2009. Structural change and the Balassa-Samuelson effect (jiegou bianhua he Balassa-Samuelson xiaoying). Shijie jingji (The Journal of World Economy) 2009 (4), 38–49.

Wei, Shang-Jin, Xiaobo, Zhang, 2011. The competitive saving motive: evidence from rising sex ratios and savings rates in China. Journal of Political Economy 119 (3), 511–564.

Wen, Yi, 2010. Saving and Growth under Borrowing Constraints Explaining the "High Saving Rate" Puzzle. Federal Reserve Bank of St. Louis Working Paper No. 2009–045C.

Williamson, John, 1990. What Washington means by policy reform? In: Williamson, John (Ed.), Latin American Adjustment: How Much Has Happened? Institute for International Economics, Washington, March 1990.

Woo-Cumings, Meredith, 1997. The political economy of growth in east Asia: a perspective on the state, market, and ideology. In: Aoki, Masahiko, Kim, Hyung-Ki, Okuno-Fujiwara, Masahiro (Eds.), The Role of Government in East Asian Economic Development: Comparative Institutional Analysis. Clarendon Press, Oxford (May 1997).

Wu, Jinglian, 2005. Understanding and Interpreting Chinese Economic Reform. Texere, London (December 2005).

Wu, Ho-mou, 2012. China's local government debt: issues and concerns. Paper presented in China's Economy in 2012: Views from Chinese Economists, China Center for Economic Research, Peking University and the National Committee on United States-China Relations, the New York Stock Exchange, January 9, 2012.

Wu, Li, 2001. A study of the size of the "price scissors" in China in 1949–1978 (1949–1978 nian zhongguo jiandaocha cha'e bianzheng). Researches in Chinese Economic History (zhongguo jingjishi yanjiu) (4), 3–12.

Wu, Yaowu, Zhao, Quan, 2010. Higher education expansion and employment of university graduates (gaoxiao kuozhao yu daxuesheng jiuye). Economic Research Journal (jingji yanjiu) (9), 93–108.

Xing, Yuqing, Detert, Neal, 2010. How iPhone Widens the US Trade Deficits with PRC. GRIPS Policy Research Center Discussion Paper: 10–21.

Xu, Jianwei, Yang, Panpan, 2012. A hundred year history of global imbalances. Manuscript.

Xu, Chenggang, 2011. The fundamental institutions of China's reforms and development. Journal of Economic Literature 49 (4), 1076–1151.

Yang, Dali, 1998. Calamity and Reform in China. Stanford University Press, Stanford (August 1998).

Yao, Yang, 2004. Political process and efficient institutional change. Journal of Institutional and Theoretical Economics 160 (3), 439–453.

Yao, Yang, Zheng, Dongya, 2007. Externalities and the development of heavy industry. Frontier of Economics in China 2 (4), 467–489.

Yao, Yang, Zheng, Dongya, 2008. Heavy industry and economic development: the Chinese planning economy revisited (zhonggongye he jingji fazhan: jihua jingji zai kaocha). Jingji Yanjiu (Economic Research Journal) 43 (4), 26–40.

Yao, Yang, Zhang, Ye, 2008. Measuring the domestic technological contents of China's exports: evidence from Jiangsu and Guangdong provinces and China as a whole (zhongguo chukoupin guonei jishu hanliang shengji de dongtai yanjiu: laizi quanguo ji jiangsusheng guangdongsheng de zhengju). Zhongguo Shehui Kexue (Social Sciences in China) (2), 67–82.

Yao, Yang, 2009. Economic reform and institutional innovation. In: Shi, Zhengfu (Ed.), 30 Years of China's Reform Studies Series. Gale Asia/Cengage Learning, Singapore, October 2009.

Yao, Yang, 2011. The Role of the Government in China. Background paper prepared for China: 2020, The Asian Development Bank.

Yao, Yang, Zhang, Muyang, 2011. Sub-national Leaders and Economic Growth: Evidence from Chinese Cities. China Center for Economic Research Working Paper E2011010, November 2011.

Yao, Yang, Zhou, Jing, 2012. Social Security and Consumption in China. Paper presented in Restructuring China's Economy, the 2012 ASSA Annual Meeting, January 5–8, 2012, Chicago.

Yin, Hongpan, Liu, Xulin, 2011. Trends of China's overall Gini coefficient (zhongguo zongti jinni xishu de bianhua qushi). China Population Science (zhongguo renkou kexue) (6), 11–20.

Young, Alwyn, 2003. Gold into base metals: productivity growth in the People's Republic of China during the reform period. Journal of Political Economy 111 (6), 1220–1261.

Zhang, Tao, Zou, Henfu, 1998. Fiscal decentralization, public spending, and economic growth in China. Journal of Public Economics 67 (2), 221–240.

Zheng, Jinghai, Bigsten, Arne, Angang, Hu, 2009. Can China's growth be sustained? A productivity perspective. World Development 37 (4), 874–888.

Zhou, Shen, Li, Ke'ai, Gong, Xuejiao, Zhang, Liang, 2012. Wage differentials among Chinese industrial sectors (zhongguo gongye hangye jian tongzhi laodong shouru chaju wenti). Yunnan Shehui Kexue (Social Sciences in Yunan) 2012 (1), 99–103.

Zhu, Rong, 2011. Individual heterogeneity in returns to education in urban China during 1995–2002. Economics Letters 113 (1), 84–87.

CHAPTER EIGHT

Growth from Globalization? A View from the Very Long Run

Christopher M. Meissner
Department of Economics, University of California, Davis and NBER, Davis, CA 95616, USA

Abstract

What is the connection between different forms of globalization, economic growth, and welfare? International trade, cross-border capital flows, and labor movements are three areas in which economic historians have focused their research. I critically summarize various measures of international integration in each of these spheres. I then move on to discuss and evaluate the ongoing and active debate about whether globalization is significantly associated with growth in the past. I pay particular attention to the role of globalization in the Great Divergence, the tariff-trade-growth debate, and the globalization of capital markets in the 19th century.

Keywords

Great divergence, De-industrialization, Capital flows, First wave of globalization, Integration

JEL Classification Codes

F63, F15, F34, F43, N70, N20

8.1. INTRODUCTION

What is the connection between globalization and economic growth? Free international trade is traditionally seen as welfare enhancing and Pareto optimal. Since Adam Smith formulated his dictum that the extent of the market determined the division of labor, economists have both theoretically and empirically confirmed the gains from trade. Skepticism about the benefits of international market integration has always been on the scene however. Observe, just to name a few, the delusions of misinformed mercantilists; the protectionist policies promoted by figures as diverse as Alexander Hamilton or Friedrich List; the Prebisch and Singer thesis that commodity exporting nations would fare poorly in the open international markets; the price-theoretic analysis by Newberry and Stiglitz (1984) showing trade to be inefficient in the presence of certain types of uncertainty; all the way to the loud squelches of protest from anti-globalizing activists.

Beyond these stark and extreme views, a voluminous theoretical and empirical body of scholarly research exists analyzing the subtle details of the connections between globalization and economic growth. Much of the literature continues to agree with Smith's

Handbook of Economic Growth, Volume 2B
ISSN 1574-0684, http://dx.doi.org/10.1016/B978-0-444-53540-5.00008-2

bottom line that there are significant gains from trade for all parties involved. Then again, a healthy dose of well-informed skepticism exists, arguing that globalization is not unambiguously beneficial. This view has emerged from careful analysis of the long-run record and greater thought about the interaction between market failures and globalization. This chapter surveys a select amount of the large literature mainly written or influenced by economic historians in order to provide one view about what the long-run record has to say about globalization and growth.

The explosion of empirical and theoretical work on the connections between globalization and growth that occurred over the last several decades has greatly improved our understanding of this process. It has broadened the scope of analysis to include the impact of integration not only in goods markets but also in the markets for labor, capital, and even ideas relevant to the economic processes also known as production technologies. The findings of this literature, as they pertain to economic growth, are largely, but not uniformly, supportive of the idea that globalization has been positively associated with growth. Those who are less supportive often suggest that the relationship is conditional and certain other factors might influence the gains from the process.

Economic historians have long argued that the Industrial Revolution could not have occurred without international trade. Recent research continues to support this notion albeit with some new views on the mechanisms behind this relationship. An earlier literature also looked at whether European economies gained dominance because of exploitative international relationships with colonies and traditional societies. There turns out to be little evidence these relations were decisive, as we will see. Another strand of the literature argues that colonization and the slave trade damaged the prospects for growth in non-European economies. While these are old ideas, new data and new methodologies continue to support this notion.

Finally, a new and exciting strand of the literature which is less supportive of the positive association between trade and growth argues that globalization can help explain the large gap in incomes between Europe and its selected offshoots and the rest of the world that opened up ca.1800 and which persists today. This gap has come to be known as the Great Divergence.[1] This is not to pin the entire blame on globalization. Other factors such as factor endowments, institutional quality, and political factors seem to interact with globalization to enhance or limit the gains. Economists and historians are only beginning to understand the origins and persistence of these institutions and their complex interaction with global forces, but recent research leaves little doubt about their importance.

The new literature on growth and globalization gives many specific reasons for how the positive relationship might break down. These reasons often center around the patterns of

[1] Some evidence suggests that between 1970 and 2006 this divergence of outcomes has begun to be eliminated as global inequality has fallen (Pinkovskiy and Sala-i-Martin, 2009). Yet, many countries remain much poorer than the richest nations, and as an historical matter, the Great Divergence is certainly an important phenomenon.

specialization induced by international trade. Commodity price volatility has been one major problem (Williamson, 2011). The historical record suggests that for small open economies specialized in resources or agriculture, globalization enhances commodity price volatility. There is also economic volatility directly related to financial crises which have their roots in the globalization of capital flows. Certain types of countries specializing in natural resources have also faced political and economic challenges broadly labeled the "resource curse." Globalization may have negative side effects in certain circumstances. The systematic study of the conditions that determined the historical relationship between globalization and growth is in its early stages. Still, because of these observations from the long run, economists are not yet able to say that globalization is unambiguously associated with higher growth.

Before arriving at this conclusion, the chapter offers an introduction to how economists and economic historians measure and track globalization, or more precisely, global integration. Section 8.3 provides a limited survey of some relevant insights from the literature linking economic growth and market integration. Section 8.4, the first of three sections on the historical connections between economic growth and globalization, gives a review of the period 1500 to 1800. The next section illustrates how recent research views the connection between the British Industrial Revolution and globalization. Sections 8.6 and 8.7 supply critical reviews of what we know about the diffusion of the industrial revolution to other parts of the world and the role of globalization. The focus is on the cross-country comparative literature. Here, I also take a look at the role of globalization in explaining the Great Divergence. In light of this evidence, the conclusion discusses the rationale behind the assertion that globalization may not always have a positive association with economic growth.

8.2. MEASURING "GLOBALIZATION"

8.2.1 Commodity Markets

Globalization is defined broadly as the economic and social connections between the world's nations. Economic historians have recently debated the question of when globalization began, and come to no conclusion. This is largely because it is a semantic question (O'Rourke and Williamson, 2002 versus Flynn and Giraldez, 2004). De Vries (2010) differentiates hard globalization or market-based connections familiar to modern economists, from soft globalization which concerns other more qualitative connections between regions. International economists and many economic historians often prefer to study and measure integration between different markets. Roughly speaking, integration is the degree of connection between any two markets, regions or nations. Numerous ways of looking at integration exist of course, and these vary depending on whether one is investigating commodity markets, labor markets, capital markets, or the market for ideas or production technologies.

In commodity markets, economic historians have a long tradition of investigating the price gap between markets of single homogeneous goods. The logic of the law of one price demands that arbitrage eliminate price differentials until no further profit opportunities exist. Any price gap that exists must be less than or equal to the transaction or trade costs of eliminating the price gap via arbitrage operations. Price differentials in this framework persist due to physical or political barriers to trade such as tariffs or transport costs. Figure 8.1 illustrates this logic. The barriers to trade are defined as the length of line segment *tt* in the simple supply and demand framework of Figure 8.1. When *tt* shrinks to *tt′* in panel A, integration is said to have risen. We might also see trade volumes rise as the supply curve S shifts down to S' in Panel B of Figure 8.1, but there is no reduction in trade barriers between two countries in this case. Looking at trade volumes alone to say something about integration can be misleading according to O'Rourke and Williamson (1994). The level of trade has risen, but in this case it could be due to productivity advance or other favorable supply shocks with no reduction in the cost of trade. Trade has obviously grown, but integration, which is related to the sum total of all barriers to international trade, has not changed.

Further conceptual refinement on this topic by Jacks (2006) looks not only at the time-varying price gaps between markets for a standardized commodity but also at the dynamics of the price differential itself. The approach models the price gap between two markets as a threshold auto-regressive process.[2] To understand this, begin by noting that in a competitive market with forces of arbitrage, the difference in price P between market o (origin) and market d (destination) must be within a narrow band, as follows:

$$-\tau_t^{do} \leq P^d - P^o \leq \tau_t^{od}.$$

Here, the variable τ_t^{od} measures the total cost of sending one unit of the good from market o to market d. When the price differential holds there is a band within which it can fluctuate equal to $\tau_t^{od} + \tau_t^{do}$. The width of this band is determined by the cost of arbitraging price differences. When shocks hit either market causing price differentials to escape the band, one of these inequalities is violated. At this point, price differentials are eliminated, but not instantaneously. The forces of arbitrage take time to eliminate these profit opportunities as information travels slowly and shipping takes time. Therefore, the price differential is assumed to follow a random walk within the bands but it follows an autoregressive process if the price difference jumps outside of the bands. The width of the bands can therefore be estimated to give an indication of all of the economic barriers impeding arbitrage otherwise referred to as trade costs. Note that these trade costs can include shipping costs, tariffs, and other trade policies, as well as the financing costs of arbitrage or even the impact of uncertainty on the market.

[2] Jacks (2006) analyzes an asymmetric model where route-specific trade costs can matter. The exposition here simplifies and requires that trade costs be the same in both directions between trading nodes. The possibility of storage is also ignored here.

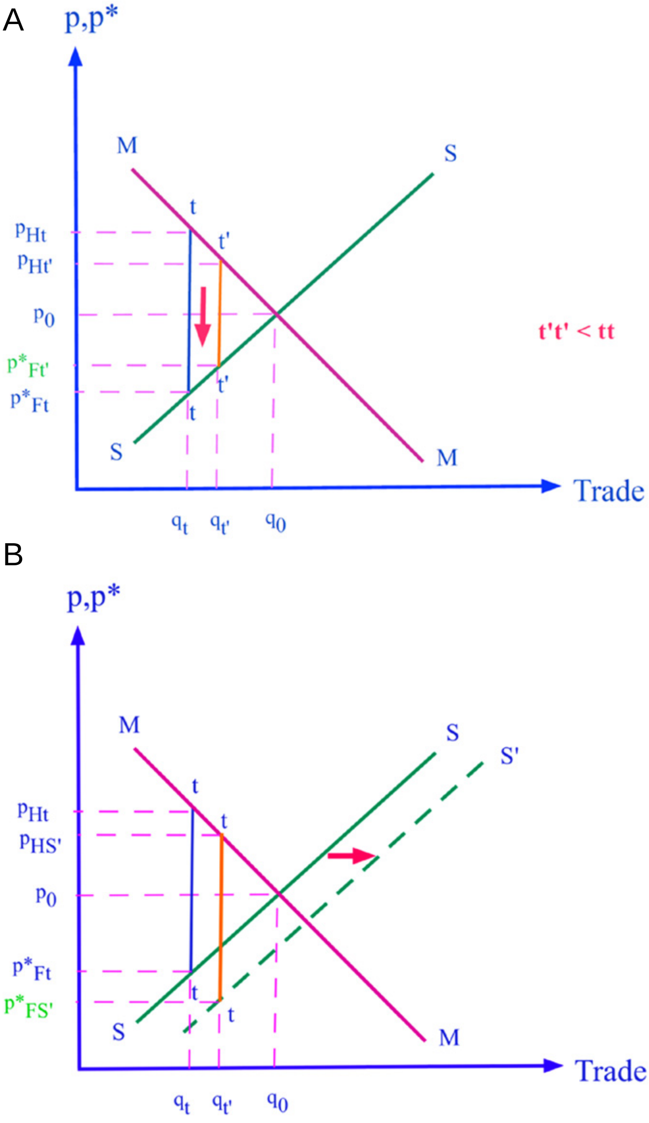


Figure 8.1 (A) A rise in integration and trade between two markets as a consequence of lower trade costs. (B) A rise in trade between two markets with no rise in integration.

Price-based measures have their pitfalls as do all measures of integration. In the case of price differentials, there is no micro-founded theory to the price differentials in question. The voluminous literature on pricing-to-market for goods in the industrial sector where market power is often evident suggests that price differentials can be the outcome of preferences manifested as market-specific elasticities of substitution—as well as technologies of arbitrage. Coleman (2007) makes the subtle, but crucial, point which is often disregarded in many historical studies of this sort, that a necessary condition to infer trade costs between two markets o and d is that trade between o and d be strictly positive. In other words, there are significant errors in many of the studies that have used price differentials between two unconnected markets to infer something about trade costs. Finally, in two markets where factor endowments, preferences, and technologies are the same, and shocks are perfectly correlated, prices will be equal with or without barriers to trade. Price equalization is not guaranteed to be perfectly correlated with integration.

As an alternative to the study of price gaps, gravity models of international trade have been used. Empirical gravity models are a powerful, yet parsimonious tool to measure international integration. The gravity approach is inspired by Newtonian physics and has been pondered in economics, as Anderson (2011) recently emphasized, since at least the 19th century (e.g. Ravenstein, 1889). Gravity says international trade is positively related to the size of two markets and negatively related to distance. Size in economics is related to total expenditure or incomes, and distance is a proxy for the barriers to trade or trade costs. Micro-founded partial or general equilibrium models of international trade give rise to gravity as discussed in Anderson (1979), Bergstrand (1985), and Head and Ries (2001) among many others. Anderson and van Wincoop (2004) go so far as to argue that gravity is consistent with *any* underlying structure of production. In other words, both Ricardian *and* factor endowment-based models of trade with positive trade costs give equivalent gravity models. This makes it hard to distinguish which forces are "causing" trade, but in some cases, where integration is the object of analysis this turns out to be irrelevant. A particularly intuitive expression of the gravity model is given as:

$$x_{od}x_{do} = \frac{x_{oo}x_{dd}}{(\tau_{od})^{(\sigma-1)}}, \tag{8.1}$$

where $x_{od}x_{do}$ is the product of trade flows or expenditure on foreign goods between country o and d, $x_{oo}x_{dd}$ is the product of the two countries' expenditure on tradable domestic goods, τ_{od} is the product of the ratio of international trade costs to domestic trade costs, and $\sigma > 1$ is the (constant) elasticity of substitution between any two goods, domestic or foreign.[3] Higher trade costs reduce expenditure on foreign goods, pushing demand toward domestic goods. This equation can be estimated by using proxies for domestic trade (e.g. total GDP minus exports) and trade costs (e.g. tariffs, distance, shipping costs, common monetary standards, and languages). Coefficients on the trade costs proxies then provide the partial effect of such frictions on international trade, and under certain assumptions, allow for inference on the elasticity of substitution. The gravity model also

[3] In this specification, trade costs are modeled as a trade cost factor equal to one plus the tariff equivalent of trade costs. Normalize the domestic trade cost (i.e. the cost of getting something from the factory or farm gate to the consumer) to one. Then, the total import price in country d for a good shipped from o is $p_{od} = \tau_{od}p_o$.

leads immediately to a very useful measure of integration. Solving (8.1) for the trade barriers allows one to infer trade costs as the scaled ratio of foreign trade to domestic expenditure. If an estimate of the elasticity of substitution is at hand, or a value is assumed, a particular value for these trade costs in tariff equivalent terms is readily available.

To measure the degree of international integration over time and across countries, Jacks et al. (2011) solve Equation (8.1) for the unobservable trade cost term. This value is the difference, or the wedge, between total expenditure on domestically produced tradables and total expenditure on foreign produced tradables. This ratio is directly related to trade costs by inspection of Equation (8.1). With an elasticity of substitution of 11, the (unweighted) average bilateral trade costs for the US between 1870 and 1913 were the equivalent of a 70% tariff on foreign goods while for the UK they equaled about 50%. With a lower elasticity of substitution one finds higher trade costs. The average tariff equivalent for a large sample of bilateral pairs is 140% in 1910, 158% in 1933, and 124% in 2000. During the period before World War I, the average annual growth rate of bilateral trade costs was on the order of −0.8% prior. Between the world wars, these grew at roughly +0.4%, and after World War II they fell at a rate of −0.5%. A U-shaped pattern of long-run integration emerges clearly from such data. Integration was high in the 19th century, fell in the interwar period, and then rose to new heights by the end of the 20th century.

One can also think of this trade cost measure as a residual in the spirit of growth accounting exercises. In this case, the residual is the gap between actual international trade and that predicted by the size of the two nations alone. Jacks et al. (2011) show how the gravity equation allows for an accounting exercise similar in spirit to growth accounting. Here, any growth not accounted for by expansion of the domestic trade terms is attributed to changes in trade costs. This is just as in growth accounting where changes not attributed to changes in inputs would be attributed to changes in total factor productivity. The gravity model exhibited above is consistent with a wide range of demand-side and supply-side frameworks, as one can easily show. More general translog expenditure functions, however, yield different expressions for bilateral trade, as do nested CES functions that allow different elasticities of substitution across varieties of goods (Novy, 2013; Feenstra et al. 2011). The assumption that the elasticity of substitution is not homogeneous across goods or countries does not seem to give highly misleading results as discussed in Jacks et al. (2011), but is certainly one of the potential pitfalls for using this structural approach.

8.2.2 Integration in Capital Markets

Global financial flows are also governed by arbitrage and gravity-like relationships (Obstfeld and Taylor, 2004; Clemens and Williamson, 2004a). In terms of the gravity approach, flows are larger when the share of the receiving economy in the global economy is larger. Simple portfolio theory in a frictionless world would dictate that the portfolio share of a nation's assets would correspond to the share of total world output. Of course, informational frictions loom large in global capital markets as they do in local markets.[4] Another consideration for the direction of capital flows is the correlation structure of the returns on various assets, based on the logic of the international capital asset pricing model (ICAPM). Nations that have lower correlations (i.e. lower betas) with the market portfolio would be in higher demand under normal circumstances.

Obstfeld and Taylor (2004) study interest rate differentials from the late 19th century until the present period for several different kinds of assets including long-term sovereign bond yields and short-run money market funds. They find strong evidence for high levels of integration based on the small observed deviations from exchange-risk-free interest parity. In the interwar period, integration is found to be much lower by this same measure, while from the 1970s, the data demonstrate tight integration once again. For long-term sovereign bond yields, there is evidence of significantly lower coefficients of variation on bond yields in 1910 than in 1870. Mauro et al. (2006) compare the 19th century sovereign bond markets to those of the late 20th century. They undertake a series of event history analyses to measure the reaction of bond prices to news. Their finding is that co-movement is much higher today than in the past when bond prices reacted much more to local news than global shocks.

On the quantity side, the portfolio positions of investors are extremely hard to track historically. What we do possess is data on gross capital flows from the major capital exporting economies of the 19th century. These can be used to track foreign assets relative to global or the investor country's GDP. Schularick (2006) estimates that the ratio of gross world assets divided by global GDP was about 20% in 1913 while today that ratio stands at roughly 75%. Gross inflows (which are widely assumed in the literature to equal net flows in the 19th century) into the less developed world were much larger in terms of receiving country GDP in the first era of globalization compared to today as Obstfeld and Taylor (2004) and Schularick (2006) discuss.

8.2.3 Integration in Labor Markets

Price- and quantity-based measures of integration are also available in labor markets. The relevant price in the labor market is the wage, and often the wage of unskilled labor is used to minimize problems in comparability across occupations. Analogous to the goods market, with free movement of labor, workers gravitate to localities with higher wages subject to several economic constraints. The standard metaphors in the economic history literature for these constraints are push and pull forces. These connote factors in the sending and receiving country, respectively. Flows would be larger when wage gaps are higher, where the sending country has a large percentage of the population that is male

[4] Bordo (2003) provides an excellent survey to the information problems of global capital markets in the 19th century.

and of prime working age, and where previous migration has been high (Hatton and Williamson, 1994).

During the 19th century, the world witnessed some of the biggest waves of migration in the history of the global economy. Wages, and then wage gaps, of unskilled workers for the 19th century have been meticulously constructed by Williamson (1995) and subsequently by O'Rourke and Williamson (1994). A recent large-scale project on comparative real wages goes back much further in time. Bob Allen and his collaborators have also contributed unique data on wages of building craftsmen and laborers in dozens of cities in Europe and Asia beginning in the 14th century. Allen (2001) calculates wages in 20 European cities in terms of silver, a common monetary standard over the long run, and in terms of a real wage (nominal divided by a price index) and a subsistence wage. The latter is in terms of a fixed bundle of common consumables such as bread or grain-based victuals, alcoholic beverages, fuel, clothing, and lodging. While wage gaps closed and convergence was the rule in the 19th century Atlantic economy according to O'Rourke and Williamson, large wage gaps opened up within Europe between 1300 and 1800 in Allen's data. This divergence is consistent with the notion that labor market integration within Europe was low prior to the 19th century or that offsetting forces inhibited convergence. Again, the caveats of using price-based data apply.

8.2.4 Ideas and Technology

Flows of ideas and technologies are central to the growth and globalization literature. Nevertheless, because of measurement difficulties, the empirical historical research on integration in this domain is minimal. Because of the heterogeneous nature of ideas and technology, no observable, well-organized market in ideas and technology truly exists. Price-based measures are not systematically available as they are for commodities like wheat, coal, or iron. Equal challenges exist for quantity-based measures. Madsen (2007) argues that foreign knowledge is embodied in current and past imports of high technology goods. For the period since 1950, he uses trade within several industries (chemicals, machinery, and scientific instruments) as the key proxy for flows of ideas, while before 1950 overall imports are used.

The economic history literature has also focused on qualitative information regarding technology transfer, miscellaneous prices on factory equipment and the pre-fabricated factories for shipment to foreign countries in the 19th century. Fragmentary evidence based on patent citations has been used, but the quality of information from these is low due to the variability in patent regulations prior to the 20th century and the lack in many cases of domestic rules for citation of foreign patentees.

Clark and Feenstra (2003) and Lucas (2009) use a structural approach and match the data on aggregate labor productivity to say something about technological catch-up in the nineteenth and twentieth centuries. Assuming a Cobb-Douglas production function, Clark and Feenstra report massive gaps in total factor productivity between countries that can only be explained by a failure to adopt best practice technologies in the less developed world. Clark and Feenstra note that the telegraph and shipping technologies connected the far reaches of the globe and allowed for rapid transmission of ideas and information when necessary. Politically, European empires often provided transfer of institutional technologies including strong property rights and other necessary cultural, social, and legal conditions. British and American firms also began to specialize in the production of startup packages for hopeful entrants to the textile industry in the 19th century. These packages often included capital goods as well as human capital in the form of consulting on engineering and managerial issues. Still, despite all of this, many countries lagged behind. Clearly then, ideas and technologies in important industries have had the potential to be widely shared across space and increasingly so, since the 18th century, however, systematic measurement of this process remains highly qualitative, and when it is quantitative, there are only a very limited amount of studies to date.

8.3. CHANNELS: THE THEORETICAL LINKS BETWEEN GLOBALIZATION AND GROWTH

8.3.1 Static Models and the Gains from Trade

There are many different views on the channels through which globalization, or integration might affect economic growth. It must be recognized that most theoretical results which provide inspiration for numerous investigations of these links were derived in static environments. The history of the global economy is dynamic by definition, and many of the standard arguments are in fact not well suited to explaining long-run growth or helping us understand the dynamic interaction between trade and economic growth. The intuition for the static gains from trade may not carry forth to a long-run environment where intertemporal factors affect current investment decisions. Investments in physical and human capital and of course in research and development are the key drivers of long-run growth. It took until the 1980s and 1990s for a significant literature to develop a coherent view of the connections between trade, investment, innovation, and growth, and still, many of these ideas have yet to filter into the analysis of the long-run of economic history despite their importance. Here we review a limited set of views on the connections between trade and growth which have been oriented to understanding problems in economic history.

The textbook starting point has always been the recognition that limited international integration impedes the efficient allocation of resources. It is easy to show in a static model with almost any micro-structure that under free trade, in a small open economy, a representative consumer has higher welfare than under autarky. Consumer gains in standard models arise from improvements in the terms of trade. Increases in the terms of trade are associated with higher welfare and higher incomes. Still, the gains from eliminating inefficiencies and the barriers to international trade, are not all that large when all resources are fully employed or nations have large domestic markets.

Recent research has also emphasized the gains from variety. For consumers, increasing variety in the consumer basket due to trade brings welfare gains from what are essentially new goods in the consumption basket. Feenstra (1994) shows how to measure the drop in the consumer price index from such changes. This allows one to show another set of gains in real incomes from international trade. Romer (1994) suggests similar gains for producers from an increased variety of intermediates allowed for by international integration. Desmet and Parente (2009) argue that international trade brings forth higher price elasticities. In this case, profitability rises as output expands to serve foreign (or large domestic) markets. This endogenously raises the rate of growth of technological change since more profitable producers can afford the fixed cost of technological change. Little work has been done on estimating the magnitude of these effects in the past.

To get a handle on the size of the gains of trade, a particularly intuitive expression has recently been derived in research by Arkolakis et al. (2012). They investigate the gains from trade in several leading models of international trade including perfect competition, monopolistic competition, and trade with intermediate goods. Under fairly standard conditions these gains are given by the formula:

$$(\lambda)^{\frac{1}{\varepsilon}} - 1,$$

where λ is the share of total expenditure devoted to domestic production or 1 minus the ratio of imports to total income, and ε is the elasticity of imports with respect to a change in trade costs. The modern literature's estimates of this elasticity are in the range -5 to -10. The gains from trade are interpreted as the percentage change in real income needed to compensate a consumer for a move to complete autarky. For a nation with an import share of 15% and with a fairly low elasticity of -5, the gains from trade are roughly equal to 3%. Higher elasticities would give smaller gains. To find the rise in income attributable to a rise in trade, Arkolakis et. al. present another calculation. Consider a move to free trade, say for the United States in 1890. This would be going from the historical average *ad valorem* tariff equivalent of 40%, to no tariffs. The relevant calculation for the rise in income is calculated by Arkolakis et. al. as:

$$1 - \left(\frac{\lambda}{\lambda'}\right)^{\frac{1}{\varepsilon}},$$

where λ' is the share of domestic expenditure after tariffs are lower and λ is the share before tariffs are lowered. In the case where the trade cost elasticity is -5, imports rise threefold. In the late 19th century American case, an actual import to income ratio of roughly 6% might have become 18%. The gain from this move to free trade is then calculated as 2.7% of income. The elasticity of income with respect to trade is then a very small 0.0135. Recent empirical estimates from Feyrer (2009) produce a much larger elasticity of 0.5.

Calculations similar to those above, but for a small open economy instead of the United States, can also be done. But the bottom line from such calculations is that the gains from trade in commodities alone cannot easily account for the massive rise in living standards witnessed over the last 200 years. If one wants to pursue the issue and find a significant link between trade and growth, then another tack must be taken. One possibility is that the static view of trade and income needs to be supplemented with a dynamic view for us to understand whether there can be any meaningful association between integration and growth.

8.3.2 The Dynamic Gains from Globalization

One simple way the literature has thought about dynamics is to study the one-off long-run impact on income of a change in integration in general equilibrium. Computable general equilibrium models yield predictions on how a change in globalization of trade, labor, and capital markets can lead to a change in incomes and so forth. When trade barriers fall, nations specialize in goods in which they have a comparative advantage. Subject to several important assumptions, the Stolper-Samuelson factor price equalization theorem concludes that this will lead to wage convergence.

O'Rourke and Williamson (1999) summarize a large literature, which they mostly pioneered, and argue that both trade and migration were a force for convergence in the 19th century in the Atlantic economy. Trade forced wages up in low-wage, labor-abundant countries toward the level of labor-scarce nations. Capital flows offset these convergence forces when they flowed from labor and capital abundant regions (i.e. Britain) to labor-scarce but natural resource-abundant regions (e.g. Canada, the US, etc.). Labor flowing toward economies with high wages from low-wage regions acted as a force for convergence as predicted by such models. The bottom line of this research program is that globalization is likely to lead to convergence (O'Rourke et al. 1996). Convergence, however, is a disequilibrium phenomenon. According to standard models of trade, there is no reason for the growth rate of productivity to be higher in the long run in a more globalized world. What was witnessed in the 19th century was essentially the comparative statics result outlined in a one-shot general equilibrium model of the international economy.

A conceptual revolution in understanding growth emanated from new growth theory which promised something more in terms of the benefits from trade (Rivera-Batiz and Romer, 1991). The general view from new growth theory is that larger or more integrated markets enable entrepreneurs and inventors to more easily cover the fixed cost related to the development of a new idea.[5] Open international markets also promote the sharing

[5] Dynamic issues were also considered in the earlier literature based on learning-by-doing and "infant" industry protection. Many exemplary case studies exist but the literature has not yet shown systematic evidence for such forces in history. David (1970) and Head (1994) find evidence of learning-by-doing in 19th century US cotton textiles and 19th century US steel rails. Head argues for welfare losses to users from protection. Irwin (2000) argues there were welfare losses from tariffs in the case of 19th century US tinplate.

of income enhancing ideas raising incomes and providing further stimulus for new ideas. As explained by Jones and Romer (2009), the growth rate of ideas rises as integration rises, or more generally, the incentives to innovate improve. The theoretical literature thus suggests that the growth rate is a positive function of the size of the market. Romer (1996) argues that American economic development in the 19th century was founded on economies of scale, and that America's size also helped increase the rate of advance of total factor productivity.

Another interesting avenue for dynamic gains is the possible interaction between institutions which facilitate innovation and productivity advance and the size of the market. Acemoglu et al. (2005) suggest that trade interacted with the political economy of European regions between 1500 and 1800. The urban merchant class, with an interest in strong property rights and low sovereign taxation, saw their fortunes and political influence strengthen as the Atlantic economy burgeoned between the 15th and the 18th century. In regions where absolutist monarchs ruled, like Spain, this did not occur. Here, exposure to the trade opportunities in the Americas and the broader Atlantic basin failed to foment institutions supporting commerce, trade, and urbanization.

Oppositely, outside of Europe in the 19th century, where societies came under the colonial domination of Europeans, weak institutional legacies often led to reduced incomes (Acemoglu et al. 2001). More specifically, in places where European settler mortality was high—due to endemic tropical diseases—Europeans looted and extracted resources but failed to invest in the establishment of strong property rights. These forces persist today long after de-colonization. Their evidence shows that places where settler mortality was higher have lower protection of property rights and hence, relatively poor economic performance in the last half century.

Galor (2004) and Mountford and Galor (2008) give further theoretical insight into the conditions under which globalization may fail to lead to modern economic growth and instead keep some nations locked into a Malthusian regime. The Malthusian regime in this work is characterized as a situation where long-run living standards grow only very slowly. The Malthusian regime dominates until sufficiently high labor productivity is reached which can take a long time. In the most basic framework (cf. Galor and Weil, 2000), larger populations lead eventually to sufficiently high income per capita to spark a demographic transition. This allows for lower fertility and higher standards of living with a high rate of productivity growth. Families eventually opt for greater quality of offspring rather than higher quantity when incomes reach a certain threshold since the rate of return from investing in such human capital is high and the opportunity costs of raising children rise with incomes.

In such models, international trade does not improve prospects for long-run growth in all regions. This is because some areas will not have a comparative advantage in skill intensive industry if they are resource-abundant or labor-abundant. If productivity growth depends on skill intensity in the previous period, then regions forced to specialize in low-skilled activity may remain mired in a Malthusian equilibrium. They persist in producing unskilled intensive or non-industrial goods due to their trade with higher income regions. In such regions, population growth eliminates any gains in per capita incomes due to productivity growth, these regions stay relatively poor and modern economic growth never appears. Trade does not stunt growth in all models of trade and growth, of course. A simple exploration by Eaton and Kortum (2001) of a Ricardian model of trade shows that productivity advance is invariant to barriers to trade. Larger markets incentivize innovation, but trade makes it more difficult to come up with an idea to compete with foreign technologies. Which effect dominates, if any, determines the long-run rate of growth of an economy.

Williamson (2011) asserts that trade led to de-industrialization in many regions from the 19th century. This often occurred where regions did not have the appropriate comparative advantage to specialize in industry. He highlights four reasons why a failure to industrialize might harm growth. First, industry often gives rise to urban agglomeration effects. Dense urban factor markets also bring efficiency gains. The demand for high-skill technical staff and services that facilitate industry brings productivity gains too. Finally, knowledge transfer is facilitated in urban industry. Williamson also notes that places that specialize in non-industrial pursuits have often fallen victim to the Dutch Disease due to an overvalued real exchange rate. Commodity specialization also brings high export price volatility and hence lower investment. In a similar vein, Ross (2005) and Bulte et al. (2011) note that resources are often associated with political instability, low investment, and low growth. Resources create rents and enhance the ability of a country to borrow on international markets. In situations where authoritarian regimes claim property rights over all resources, borrowing or the ability to export commodities on world markets for quick cash can lead to "hit-and-run" or looting strategies. The impact is often low investment in the wider economy, political instability, and low growth. Ross (2005) examines conflict where local insurgents battle incumbents for the chance to control natural resource rents borrowing on the collateral of resource rents via "booty futures" to fund such activity. This type of conflict provides a drag on economic growth.

A proper historical treatment of the idea that trade limits economic growth would also model both supply and demand forces shaping human capital accumulation and account for the institutional and market forces allowing for movement into high-skilled products. Many nations specialized in non-industrial goods such as Canada, New Zealand, and Australia and managed to maintain high incomes and high growth rates. Today, countries in East Asia and elsewhere are promoting labor-intensive manufacturing and experiencing rising living standards although this process took a long time to appear.

International capital flows should also allow capital scarce countries to raise their standards of living and converge. Many observers believed that the historically unprecedented outflows of European capital during the 19th century were often associated with better infrastructure and allowed for capital accumulation in the private sector. Gourinchas and Jeanne (2006) calibrate a neo-classical growth model and find that growth rises slightly in the short run from such infusions. Large impacts on living standards and growth can only arise in such a model when capital flows are associated with deeper institutional and social changes. These forces allow for a higher long-run level of income per capita and hence add to the potential for longer transition dynamics. Also, it is worth noting that in the neo-classical model of growth, a permanent rise in the rate of capital inflow would be akin to a rise in the saving rate. This would lead to temporarily higher growth rates and higher incomes in the long run but no permanent effect on growth rates.

Further work by Rancière et al. (2008) is suggestive that countries that proceed apace with financial liberalization grow more quickly as entrepreneurs leverage an expansion in the capital stock. This generates a higher probability of a financial crisis, but, overall, stronger growth dominates in the long-run compared to nations that do not liberalize. In such a case, a country will have large negative skewness of credit growth and be more susceptible to systemic crises. This would not necessarily result in a more variable growth path for incomes, but *would* be associated with higher average growth rates. We now turn to a discussion of the historical record on the relationship between globalization and growth.

8.4. GLOBALIZATION AND HISTORY: FROM ANTIQUITY TO THE 18TH CENTURY

The time from antiquity to the 18th century encompasses a period when all regions of the world were constrained under a Malthusian growth regime. This implies increases in living standards occurred only sporadically when small technological innovations arose. Higher living standards (i.e. incomes per capita) could not endure if population growth responded in the long-run. Within societies, feudalistic and other anti-competitive institutional arrangements gave ruling classes opulent lifestyles, but, by and large, economic growth was limited. Inter-regional trade was historically always an important force for sharing ideas and reducing price differences in commodities. But since transportation technologies remained limited, institutional protections for long-distance trade did not exist, and trade was not fully competitive; overall trade, specialization and income growth due to such exchange was limited.

Findlay and O'Rourke (2007) lay out the interaction between geography, military power, and technology across these centuries. In a pulsating analysis of the Arab conquerors, Viking incursions, the Pax Mongolica, Venetian dominance, the Chinese empire, and European discoveries, these authors emphasize that cross-regional trade has long affected local economies. Indeed, the search for scarce commodities and trade opportunities drove many of the major geopolitical convulsions of the past.

Arab traders from the 8th century defined a trade network encompassing the Iberian peninsula in the west to the Indus and the Oxus in the east. Strong trading connections would eventually reach as far as South Asia, China in the East, and western sub-Saharan Africa. Not surprisingly, the various Arab caliphates produced great, if not highly concentrated wealth, while intermediating trade between the East and Western

Europe. Such a network enabled the northern European economy to obtain Eastern textiles and spices. This trade was generally cooperative on both the purchasing and selling end, but underlying forces within the local economies were decidedly feudalistic leading to the (probably unanswerable) question of whether there were net welfare gains from such trade.

With the arrival in Europe of Genghis Khan and the Mongol conquerors from the East, a vast overland network that sheltered Eurasian trade unfolded. The Pax Mongolica strongly stimulated the transfer of ideas, techniques, and goods.[6] As it turned out, the Mongolian invasions led directly to the spread of the Bubonic plague which afflicted Europe beginning in the mid-14th century. The Black Death revolutionized the price and wage structure in western Europe by instantaneously raising the real wages of survivors. As a result, in western Europe, bargaining power shifted against the feudal elite toward laborers. Labor gained and labor-saving technological change was induced. Higher wages also promoted urban merchant power by raising trade imports to Europe from Asia of luxury goods.

The European voyages of discovery, intent on breaking the Venetian, Genoan and Muslim intermediation of the Far Eastern trade, began in the 15th century. The arrival of Columbus in the western hemisphere and the circumnavigation of the Cape of Good Hope initiated a new epoch in the international economy. By the late 16th century, massive silver flows to Spain led to the onset of Dutch Disease in Spain (Drelichman, 2005). Demand for luxury goods increased, while the price of non-tradables rose. Spain sourced its luxury imports from the Low Countries in return for American silver. This gave rise to a new division of international labor. Indigenous labor in the Americas extracted silver, northern Europeans produced fine cloth and other consumer goods while trade with the East in spices and textiles intensified. New goods such as coffee, tea, tobacco, sugar, and cocoa enlarged the choice set for consumers and hence provided higher welfare. O'Rourke and Williamson (2009) investigate price convergence between Asia and Europe on spices after 1500 and find evidence of price convergence attributable to the Portuguese trade. The introduction of trade routes not only facilitated productivity gains due to improved division of labor, but it also increased consumer welfare. Voth and Hersh (2009) estimate the value of access to these new goods as 10% of a common English laborer's wages. Such large gains, due to the spread of Empire and the enhanced competition on trade routes, certainly helped habituate European consumers to a higher standard of living. The availability of a broader range of goods associated with this Commercial Revolution gave an incentive to work harder.

[6] Diamond (1999) emphasizes that over the long-run, trade of ideas and techniques on the Eurasian landmass was at the root of the eventual dominance in the global economy by its inhabitants. Diamond placed emphasis on the transfer of agricultural techniques along areas of similar climate, movement of plant and pathogenic organisms, and the domestication of native animals for husbandry.

De Vries (2008) identifies a positive impact on living standards of the rise in integration from the mid-17th century. The rise in the intensity of work which coincided with these changes was called an "industrious revolution" by De Vries (1994). The argument is straightforward: as the barriers to international trade fell, and the range of consumption goods improved, the incentive to work intensely increased. Rapid changes in the consumption patterns identified in Holland and England from the 16th and 17th century could only be had if productivity and hourly wages had risen—which is unlikely—or if workers increased total hours.

Dutch command of shipping routes gave Amsterdam its pre-eminent entrepôt status for tropical goods. On the back of these changes, urbanization rates increased in the Low Countries, and financial innovations allowed for Dutch pre-eminence in financial developments in the 17th and 18th centuries including sovereign lending and trade finance. Such advantages were soon to be eliminated by the English however. The roots of the industrial revolution in the early 18th century stem not only from the domestic institutional foundations that restrained sovereign profligacy (North and Weingast, 1989), but from the English policy priority of dominating maritime trade. Findlay and O'Rourke (2007) call attention to the French envy of the dual English focus on military domination and development of commerce and trade. Contrary to other continental powers (excluding the Dutch of course) which limited economic activity to older feudal patterns and privilege, the English "combined since the time of Elizabeth to promote trade" (Crouzet, 1981 p. 65). International trade therefore is seen by many authors as a critical component of the British Industrial Revolution.

Property rights may have also been shaped by exposure to international trade. Acemoglu et al. (2005) argue that the exposure of European economies to a global economy shaped their subsequent economic growth. Regions that were heavily involved in international trade *and* which had greater checks on their rulers urbanized and grew more quickly than other regions in Europe governed by absolutist monarchs. Trade with Africa, the Americas, and Asia enhanced the bargaining power of local merchants and allowed for greater security of property rights.

8.5. GLOBALIZATION AND THE BRITISH INDUSTRIAL REVOLUTION

The British Industrial Revolution is of course a complex phenomenon and its causes include multiple factors and their interactions. This section briefly surveys the literature's views on how trade mattered for this process. It is now widely recognized that the British Industrial Revolution was a gradual process. The initial stages were isolated in a few industries such as textiles and iron-making. The standard view is that the cotton jenny, the steam engine, and improvements in iron making were some of the prime technological breakthroughs allowing greater productivity in many industrial activities.

The latter two might be viewed as General Purpose Technologies that eventually had large spillovers for the modern sector.

Allen (2009) argues natural resource abundance, high labor costs, and extensive markets mattered for the Industrial Revolution. For instance, abundant coal, located next to rich iron deposits, along with high wages, made it rational for British entrepreneurs to substitute coal-using machines for relatively expensive labor. Eventually the cotton textiles industry was mechanized. Machines including the steam engine and the railroad allowed massive productivity gains in several sectors. All of this begs the question of whether global economic forces might have given rise to such invention and innovation. For Findlay and O'Rourke (2007, p. 348), one key pillar was "the role of parliament in promoting and fostering all forms of trade and economic activity." Exports as a share of income doubled between 1700 and 1800 from 8% to roughly 16%. Was this a symptom or a cause of British industrial success?

The clear supremacy of British cotton textiles on global markets dates from the early 19th century. Earlier, the colonial trade with India had introduced fine calico cloth to the British market. Competing domestic woolen textile manufacturers lobbied for higher tariffs and protection from such superior quality cloth. These tariffs spurred growth in English cotton manufacturing as well. Findlay (1982) classified this as an early example of (successful) import substitution. Later this gave way to export diversification. By the first decade of the 19th century, British cotton textiles claimed the largest share of British exports due to high growth rates of productivity. The creation of foreign markets within the Empire, treaties with other nations assuring low tariff levels, and protection from high-seas piracy from the Royal Navy, all helped as well. Many authors have long viewed the inexpensive access to raw materials, particularly cotton, from the western hemisphere as crucial in lowering input costs.

Indeed, an early view promoted by Eric Williams (1944) and later Inikori (1987) and Darity (1992), among others, suggested that England was able to accumulate capital and profits on the back of the slave trade and exploitation of the colonial economies via coercive labor markets. Latter day Marxists lamented the colonial plunder which left England richer but the colonies destitute. O'Brien (1982) investigated the general "contributions" of the periphery to the first industrial takeoff to conclude that the profits arising from exploitation of the periphery were unlikely to be decisive in determining British fortunes. Since trade and the profits emanating from international trade were small, a maximum of 15% of gross investment could have been due to such interactions. Still, the opening of the Atlantic world and the broader global economy allowed for larger markets for British goods, lower cost raw materials, and finally some funds for re-investment. Clark et al. (2008) conclude that trade with the entire world, but not simply the North American colonies, *was* crucial for the British economy. Their simulations show that without foreign markets, British industry would have shrunk by 35% and TFP growth would have slowed by 6%.

Eltis and Engerman (2000) survey the literature on slavery and the sugar economies of the Caribbean. Since the Industrial Revolution was founded on astonishing productivity advances in cotton textiles, and slave-produced raw materials added only a fraction to the final costs of making such goods, they find no convincing evidence of a role of the Atlantic slave trade in British economic success. Of course, all of this does not negate that the slave trade, colonization, and competition from British exports mattered for the economic growth of the colonized economies.

The areas most affected by the Atlantic slave trade in western Africa seem to be significantly poorer today than comparable regions in Africa that were less exposed to the slave trade (Nunn, 2008). One possibility for this is that the slave trade encouraged slave raiding, kidnapping, and lawlessness. The legacy is poor development of property rights and other crucial institutional foundations. Nunn and Wantchekon (2009) show convincing evidence that these places also exhibit low levels of trust and social capital which further impedes exchange and economic development.

Theoretical explorations of the causal relations between trade and the industrial revolution remain scarce notwithstanding the large literature. Standard Smithian explanations for increasing returns were ruled out early on by Findlay due to the small scale of British enterprise in the 18th century. McCloskey (1970) goes onto argue that such a small fraction of total expenditure and income relied on trade that it did not seem plausible *a priori* to attribute any peculiar role to foreign trade as opposed to domestic trade. This is a point Findlay, echoing Mantoux (1961), vociferously disputes. The relevant metaphor is that only a small amount of yeast is necessary to ferment and chemically alter an enormous mass. Clearly, the non-linearities and the relationship between the micro-level activity and the macro outcomes are not well understood—even today. What seems likely, however, is that trade was the "child of industry" rather than the other way around.

Desmet and Parente (2009) take the view that market size was decisive in a theoretical contribution to the debate. In this model, larger markets spur innovation and productivity advance. The key link between market size and growth is that larger markets have larger demand elasticities. Firms that produce in a world of large demand elasticities see revenue and profits rise with expanded production. This greatly incentivizes innovation—in this case modeled as a sunk cost which can only be profitable with a sufficiently high elasticity. The elasticity in this model is a function of the size of the market. The calibrated model seems to roughly fit the stylized facts. The model predicts rising urbanization, higher TFP growth, and expansion of the domestic and foreign markets in the late 18th and early 19th century. Allen (2009) provides additional narrative support to the high elasticity theory. British cotton textiles benefited both from local engineering and technology as well as the fact that global markets were competitive with high elasticities. Contrast this with France which produced high-end lace and knitwear and could not sell into global markets. Consequently, the incentive to innovate and adapt new technologies was lower there since the demand for productivity enhancing inputs was lower. Theoretically, the

Desmet and Parente model diverges from much of the standard trade literature which has focused on a constant elasticity case as per the Dixit-Stiglitz-Norman operationalization of the love of variety. Whether or not this particular view of British industrialization based on participation in a globally competitive industry will hold up in other data sets is an interesting question. Certainly this approach opens up many new avenues for further historical research.

Joel Mokyr credits the Enlightenment with the advent of the Industrial Revolution in northwestern Europe (Mokyr, 2010). This process of scientific awakening within Europe was unique, and it coincides with the flourishing of new scientific theories and applications to practical problems. The Enlightenment in this view is a rise in the integration of the market for ideas. As Mokyr observes, during these years communities of scientific minds were frequently brought together in various scholarly societies in Great Britain and in northwestern Europe. Examples include the Royal Society and the Académie Royale both established in the 1660s which helped filter and "sanction" the intellectual leaders of the time. Eventually, the findings of those involved would help contribute not only to new general purpose technologies but innovative ways to enhance productivity and efficiency in a broad range of industrial pursuits. The Industrial Revolution in England, and its early diffusion throughout northern France, Belgium, the Low Countries, and some of the Germanic territories was a result of these idea flows, however imprecisely measured.

8.6. GLOBALIZATION AND THE INTERNATIONAL DIFFUSION OF THE INDUSTRIAL REVOLUTION: 1820–1913

From the early 1820s, international commodity markets became rapidly more integrated, while at the same time, economies outside of Britain began to experience the process of modern economic growth. A large literature sees these two processes as intimately connected. Not only did trade flows rise as transportation costs fell, tariffs dropped, and communications improved, but migration also surged, capital from Britain, Germany, and France flowed into areas of recent settlement and less developed areas, and foreign direct investment as well as technology transfer accelerated. Growth takeoffs, demographic transitions, increased urbanization, and sustained improvements in well-being significantly transformed the way of life of the average 19th century inhabitant of Europe and North America. Not every region shared equally in this increased prosperity, but most regions participated, and in most places higher incomes were associated with greater integration. Most regions were able to secure the gains predicted by static trade theory. It is an open question as to whether globalization limited attempts to achieve modern economic growth in places which specialized in non-industrial activity. Williamson (2011) suggests it did. His evidence, summarized and discussed below, notes that the first period of globalization set off a process of de-industrialization in many places outside of Europe which ultimately stunted long-run economic growth.

The British Industrial Revolution was founded on new technologies including the steam engine which also promoted market integration. Steam engines eventually powered the railroad engines that fused national and international markets. Iron-hulled ships and the steam engine made for higher quality maritime shipping (Allen, 2009). Lower tariffs reigned in England from the repeal of the Corn Laws in 1846. In 1860, the Cobden Chevalier treaty was signed. The explosion of most-favored-nation clauses afterwards promoted trade. The telegraph from the 1850s, monetary stability arising from the classical gold standard and construction of global European empires, also enabled strong integration from the middle of the 19th century.[7] Commodity price gaps closed dramatically during this period, and world merchandise exports relative to world GDP rose eightfold between 1820 and 1913 from 1% to 8% (Findlay and O'Rourke, 2003). Concurrent to these advances, Germany from the 1850s, Japan and the US from the 1860s, and many other regions began the irreversible process of modern economic growth and/or industrialization. Per capita incomes in these places more than doubled between 1870 and 1913. In many cases, and in several ways, trade and globalization catalyzed this process.

The voluminous research of O'Rourke and Williamson summarized in O'Rourke and Williamson (1999) leaves little doubt that globalization led to wage and price convergence between the areas of recent (European) settlement and Europe. Wage gaps were pushed down by the large net emigration from Europe to the Americas. O'Rourke and Williamson show that while GDP grew at 0.7% in Ireland, GDP per capita grew at almost double the pace or 1.3% due to heavy emigration. Emigration mattered in many other places too. Eastern and Southern Europeans migrated en masse to North and South America keeping wages down in the West and bringing them up in the East. O'Rourke et al. (1994) estimate US real wages would have been about 9% higher in the absence of immigration during the 19th century.

O'Rourke and Williamson (1999) summarize the literature by noting that trade and migration were substitutes. In other words, these forces worked in the same direction to promote convergence. Indeed, more than all of the large decline in real wage dispersion within the Atlantic economies is "explained" or accounted for by analysis in Taylor and Williamson (1997). Offsetting forces such as capital flows and trade responses worked to offset some of this convergence. The data from the 19th century for the now-advanced economies within Northwestern Europe and bordering the eastern Atlantic Ocean are strikingly consistent with the predictions of the Stolper-Samuelson theorem.

[7] See Lampe (2009) on the positive trade impact of the MFN clause and Accominotti and Flandreau (2008) for the opposite view. Lew and Cater (2006) argue the telegraph promoted international trade but almost always came along with new railroad lines. López-Córdova and Meissner (2003), and Estevadeordal et al. (2003) argue that the gold standard promoted international trade between 1870 and 1913. López-Córdova and Meissner (2003) and Flandreau and Maurel (2005) show evidence that monetary unions enhanced international trade. Mitchener and Weidenmier (2005) find that empire was associated with higher foreign trade. Jacks (2006) notes that commodity price integration was higher due to institutional factors like the gold standard and empire. A large literature on tariffs, income, and growth exists. We comment below.

Factor prices, especially wages, converged in the globalized Atlantic economy of the late 19th century.

The force driving this of course was integration which promoted specialization in products using their most abundant factors of production. In the Americas, this meant that growth in the resource intensive and agricultural sectors acted to put downward pressure on wages. Wright (1990) finds evidence that the USA was a net exporter of resource-intensive manufactures in the late nineteenth and early 20th century. In Europe, increased specialization in labor-intensive industrial output served to raise wages. Consequently in places like Belgium, and in Great Britain, labor interests allied with industry to advocate free trade during the 19th century.[8] Cheaper grain imports and higher demand for their specialized industrial products abroad worked to raise incomes (Huberman, 2008).

What about integration in capital markets? To be sure, capital flows increased substantially and capital market integration rose from the mid-19th century. New and competitive financial intermediaries based in the City of London, the telegraph, the gold standard, and institutional arrangements such as the Council on Foreign Bondholders and the British Empire promoted the supply of capital and deepened integration.[9] Net inflows to the receiving countries were significant. On average, the current account deficit/GDP in countries such as Australia, Canada, New Zealand, and the USA (prior to 1860 in the latter), was on the order of 3% and much higher in many years. Foreign investment often accounted for about 20% of total investment in many net capital importers of the time and up to 50% in Australia, Canada, Argentina, and Brazil (Fishlow, 1985; Williamson, 1964 on the USA). Clemens and Williamson (2004a) reveal that the Lucas Paradox, the lack of capital flows to less developed countries, was somewhat less marked in the 19th century than in the late 20th century, but that richer countries still received a disproportionate amount of the world's capital inflows.

Clemens and Williamson (2004a) also look at the demand side and the barriers to integration in 19th century capital markets. They show that capital chased migrants and natural resources. In other words it was drawn to destinations where the marginal product of capital would likely be highest. This also leads to the observation that, *ceteris paribus*, lower capital flows would lead to lower marginal products for other factors of production and hence lower factor incomes. O'Rourke and Williamson (1997) report one initial analysis for a limited set of countries. They suggest that capital-labor ratios were higher in several countries during the 1870–1913 period due to capital inflows, but many countries

[8] There was, however, a backlash to free trade emanating from Lancashire textile industrialists and workers in the late 19th century, in the UK. This appears to be due to increased competition with other industrial nations and penetration of Eastern textiles. The latter benefited from the continuous depreciation of silver against gold and hence eroded market share, jobs, and profits (Wilson, 2001).

[9] See Esteves (2007) for a recent analysis of the Council on Foreign Bondholders. Bordo and Rockoff (1996) and Obstfeld and Taylor (2003) study the gold standard and capital flows. Ferguson and Schularick (2006) argue that the British Empire lowered bond spreads. Mitchener and Weidenmier (2005) show that colonial ties with the USA lowered bond spreads in the Americas.

were hardly affected by the global capital boom. For instance in Italy, Portugal, Spain, and Ireland, capital-labor ratios and real wages seem not to have been affected by capital flows during the period since they received such small amounts. In Denmark, Norway, and Sweden, their estimates show an average increase in the capital-labor ratio of 16%, 17%, and 50%, respectively, and significant rises in the real wage were had as a consequence.

Rather than look at the capital stock directly, for which the data are somewhat limited, a second approach has looked at the financial flows of the time and correlated them with incomes in the spirit of the cross-country empirical growth literature. Bordo and Meissner (2011) and Schularick and Steger (2010) study the short-and long-run associations between capital flows and incomes between 1870 and 1913. Both studies find that foreign capital flows are associated with higher incomes. There is no evidence yet, however, that such flows raise growth rates over the long-run. Schularick and Steger provide evidence that capital flows in the 19th century raised investment rates allowing for higher incomes. Bordo and Meissner agree but also focus on the economic risks associated with financial inflows. Inflows in the 19th century appear to be highly correlated with the probability of a banking, currency, or debt crisis and these bring income down in the short run by up to 3% on average. The negative impact on incomes of the small number of debt crises studied is large and longlasting as well.

Financial globalization's direct effect in the 19th century was to allow for rising living standards, but the indirect effect was negative via financial crises. Heterogeneity in other underlying determinants of financial crises such as reserve accumulation, trade openness, exchange rate policy, and overall financial development made it so that experience in handling capital inflows differed. Some countries like those in Scandinavia, along with Canada, Australia, and the USA seemed to benefit from capital inflows. In nations with underdeveloped financial systems, non-credible commitments to fixed exchange rates and where the executive branch of government was relatively unconstrained, capital flows presented a threat to income stability due to the higher likelihood of a financial crisis.

8.7. CROSS-COUNTRY COMPARATIVE EVIDENCE FROM THE LATE 19TH CENTURY

A large literature studies the empirical connections between trade exposure and economic growth in the post-World War II period. Many of the same research designs have been implemented in the 19th century setting. The workhorse econometric model, underlying many of the studies surveyed below, is typically a regression of the growth or level of GDP per capita on measures of trade exposure and other independent variables. Not all authors agree that free trade promoted long-run economic growth in the 19th century. In parallel to the debates that still rage regarding the connection between trade and growth in the post-World War II era, economic historians continue to debate the relationship between trade and growth in the past.

The starting point for many of these studies was the historical observation that despite the rise in the global trade share during the 19th century reported earlier, many countries imposed higher tariffs after 1870. O'Rourke and Williamson (1999) present unequivocal evidence that price gaps widened wherever tariffs were raised. American tariff policy generated the equivalent of a uniform tariff of close to 40% between 1870 and 1913 (Irwin, 2010) The southern cone of Latin America kept high tariffs as well (Clemens and Williamson, 2004b). In Europe, Germany raised tariffs in 1879. France imposed the Méline Tariff in 1884. These continental giants famously increased protection for their agricultural sectors because of the so-called grain invasion. Cheaper transportation and supply-side expansion led to a large rise in grain imports from the Americas and Eastern Europe. In Germany, the political bargain involved extra protection for industrial interests. Contrary to these moves, we see that in Asia, Japan, India, China, Siam, and Indonesia signed treaties that limited rises in import tariffs from the mid-19th century.

Bairoch (1972) undertakes a study of the impact of protection on growth prior to 1914 in Germany, France, Italy, and Great Britain. The finding is that nations grew faster under higher tariffs. Since other forces are not considered in the study, this finding stands simply as an unconditional correlation that could have been due to omitted factors.

O'Rourke (2000) enlarges the sample and conditions on several other variables in a regression framework, also finding that growth in per capita income was slowest in those countries that had the *lowest* tariffs. While Britain maintained low tariffs, the US, Canada, and Argentina boasted high average tariff rates. Recent work by Lehman and O'Rourke (2011) supports this, and goes further, suggesting that *what* countries protected mattered. Tariffs on manufacturing industries were associated with higher growth, but they were not associated with high growth in the primary sector. The explanation is compatible with a story where tariffs raise the rate of return on activity that generates externalities such as research and development, improved product quality, and an expanded variety of locally produced products. Implicit in the argument is that domestic markets are better than international markets at providing these incentives. This is at odds with much of the trade and growth literature which equates these outcomes with the overall size of the market. More research on this possibility must be a priority to fully understand the mechanisms since direct evidence has not yet accumulated. One other possibility is that tariffs are beneficial for growth but that this result is dependent on the external environment as discussed in Clemens and Williamson (2004b).

Subsequent to the early findings that tariffs coincided with high growth, a series of papers argued strenuously that the opposite was true. Irwin and Terviö (2002) estimate an instrumental variables regression where GDP per capita is the dependent variable and total trade relative to total output is the key independent variable. Geographic determinants of trade are used to predict bilateral trade and the predicted shares are aggregated across all partners to build up predicted trade shares. The latter are used as an instrumental variable for actual trade following the lead of Frankel and Romer (1999). The data for 1913 show a positive and significant relationship between trade and output per person. Irwin (2002) also disputes the notion that higher tariffs caused higher growth. Canada and Argentina, for example, relied on capital imports to create export-led, commodity-based growth. When the sample is increased to include Russia, Portugal, and Brazil, we see that they also implemented high tariffs but faced low growth. Schularick and Solomou (2011) estimate no relationship between tariffs and income using GMM techniques.

Jacks (2006) looks at a slightly different sample than O'Rourke (2000) by lengthening the time dimension and adding countries. He finds evidence consistent with both strands of the literature: openness is positively related to growth but so are higher tariffs. Following the argument of Clemens and Williamson (2004b), tariffs appear to have been associated with higher net exports and the effect seems to be dependent upon the level of foreign tariffs. In yet further work on a broader sample that includes many countries in the periphery, Blattman et al. (2002) claim that growth and tariffs were only positively associated with growth in the European core and the English speaking offshoots Canada and the USA for instance. In southern Europe and in Latin America, tariffs were high but did not correspond with growth. Their explanation is that a "country has to have a big domestic market, and has to be ready for industrialization, accumulation, and human capital deepening if the long-run, tariff-induced, dynamic effects are to offset the short-run gains from trade given up."

Another strand of the literature has taken a longer time horizon into consideration. Vamvakidis (2002) studies the 1870–1910, 1920–1940, 1950–1970, and 1970–1990 sub-periods. A positive relationship between growth and trade openness only becomes apparent after 1970. In the 1920–1940 period there is a negative relationship. Similarly, Clemens and Williamson (2004b) identify a tariff-growth paradox noting that high tariffs are associated with high growth before World War II but not after. Their explanation is that the global environment matters. In the post-World War II environment of low tariffs, nations may lower welfare by raising tariffs as penalties are imposed abroad. In a world where tariffs are high in a few large countries (i.e. prior to 1914), high tariffs might not be as damaging and may be associated with better economic performance.

In light of this great debate on openness and incomes, one might reach the conclusion that in the past there was no strong relationship between these two variables. However, a recent series of papers provides evidence that there is in fact a strong positive relationship. These papers deploy estimating equations motivated by trade theory from the last two decades. The underlying hypothesis to be tested is that lower trade costs and hence greater market access lead to higher incomes. Donaldson (2008) finds convincing evidence that in India in the late 19th century and the early 20th century establishing a railroad connection with other regions significantly raised agricultural productivity and real incomes. Rosés (2003) shows that falling trade costs in Spain in the mid-19th century led to industrial concentration and presumably to higher incomes as predicted by new trade theory. Liu and Meissner (2012) find that greater market access had a positive and significant relationship with income per capita in a sample of 25 countries in 1900. Moreover, Liu and Meissner simulate the general equilibrium effect on welfare of the elimination of international borders which seem to stifle trade as a uniform tariff of roughly 50% on all foreign goods would. The simulation suggests a rise in real incomes of 10% for large and wealthy nations like France and Germany. For smaller countries like Belgium, the Netherlands, and Switzerland the rise in real incomes is on the order of 30%. The conclusion from these country-case studies and the cross-country evidence is that a decrease in trade costs significantly raises real incomes.

Still, while it might be true that tariffs boost economic activity in protected sectors, for this to lead to a long-run welfare gain for consumers it is necessary that non-convexities exist. In other words, in a dynamic setting, industrialization, even if artificially induced by trade barriers, would have to lead to significant learning-by-doing or other productivity gains. The evidence on the latter is limited, but a new strand of the literature on the Great Divergence is consistent with this argument.

8.8. GLOBALIZATION AND THE GREAT DIVERGENCE: THE PERIPHERY FALLS BEHIND

The case that trade is universally correlated with high growth and rising living standards has also been challenged based on the historical record of the periphery. How was it that during the 19th century, a period of deep integration, many nations fell behind and failed to industrialize? During the 19th century global boom, Galor and Mountford (2008) observe that (unexplained) early advantages in factor endowments were decisive for the now richest countries. Specifically, these nations were relatively technologically advanced being abundant in semi-skilled and high-skilled workers and hence less land- and labor-abundant by the early 19th century when the global trade boom erupted. In these nations, trade augmented the incentive to invest in human capital while in those nations endowed with natural resources or abundant in low-skilled labor, the incentive to invest in human capital was low. These nations became increasingly specialized in low-skill intensive industries. Per capita incomes did not rise as fast in the periphery as in Europe and North America, there was a delayed demographic transition, and the gap between the richest and poorest countries increased. This is not to say that the periphery did not see rising incomes in the 19th century. The best available evidence suggests that many nations did grow. However, they did not grow as fast as the core industrial countries and they failed to maintain these growth rates over the long run. The well-known reversal of fortune in Argentina—once one of the highest income countries in the world—is an extreme case here.

For this story to hold, the demand for human capital cannot be high even despite an apparently high rate of return in such countries. Galor and Mountford (2008) rationalize this by arguing these rates of return "reflect a suboptimal investment in human capital in an environment characterized by credit market imperfections and limited access to schooling." Clearly, a better understanding of the market for education and the institutional foundations of the supply for education is merited here. These are likely to be very important constraints. Lucas (2009) also investigated the diffusion of the industrial revolution to poor countries and posited a wedge that inhibited accumulation of technology or know how. The traditional sector is large in poor countries and less receptive to frontier industrial technologies. The conclusion that backward countries would never converge does not stand in simple versions of this model. Eventually, the incentive to adopt new industrial technologies would be large enough to induce a change. In the short- to medium-run, however, gaps between rich and poor open up as there is also a strong incompatibility between industrial know-how and local traditional production techniques. Clark and Feenstra (2001) take one approach to studying these gaps in the 19th century and find that the incentive to adapt leading technologies was ostensibly too small to induce change in many non-European economies. Since the 1950s, many once-poor nations have joined in the process of convergence, particularly those in East Asia, Chile in Latin America, and several nations from Eastern Europe. The important role of human capital in this process in East Asia has been highlighted by Crafts (1999) among others. It may be the case that the theoretical models discussed above can explain this late industrialization. To do so they would need to explain the timing, and this should be related to the flow of useful ideas from the more developed part of the world or from domestic sources and the incentives to exploit them.

Williamson (2011) explores the historical dimensions of de-industrialization in what has now come to be known as the periphery or the less developed countries. These are places where manufacturing once flourished but where such activity witnessed a steady decline over the 19th century and in the early 20th century. For example, India and China produced more manufactures as a percentage of world output prior to the 19th century than Great Britain and the rest of the Western economies, but this obviously was not the case throughout the 19th and 20th centuries. Such nations were not highly specialized in industrial activity over the last 200 years. Williamson (2011) equates industrial activity with higher long-run growth and hence de-industrialization creates divergence by definition. In Williamson's view, urban industry creates high demand for skilled workers, it benefits from agglomeration, and it allows factor markets to be thicker and hence more efficient. The corollary is that specialization in non-industrial activity, induced by trade with places with a comparative advantage in industrial goods, should lower the capacity for long-run growth. Resource-based economies fall victim to the resource curse as rents accrue to a wealthy elite and Dutch Disease lowers investment and productivity advances in the industrial tradable sectors. Resources also give rise to conflict as actors compete to gain access to the rents created by these endowments. Terms of trade volatility, documented to be much higher in primary producing areas, also seem to have been correlated with lower investment and lower growth. Williamson builds the case that places like Japan, which was labor-abundant and resource scarce, were able to industrialize and avoid a resource curse precisely due to this set of factor endowments. In the late 20th century, countries that have actively promoted industry and especially labor-intensive exports have also fared better as in East Asia.

The striking divergence of different sets of resource-based economies from the 19th century raises a challenge to Williamson's thesis, however. After all, Canada, Australia, and to a large degree the United States built successful, high growth economies from their resource-based comparative advantages (Keay, 2007; McLean, 2004; Wright, 1990). The staple theory of economic growth, as applied to Canadian development, proposes that an economy can build on forward and backward linkages (Watkins, 1963). Norway's oil-finds of the late 20th century did not lead it down a conflict-ridden path followed by many of the west African states "blessed" by oil reserves, nor has it fallen victim to a Dutch Disease. Chile may also be heralded as a success case. This nation has steadily managed to elevate its economic status in the last three decades by becoming a net agricultural and resource exporter. These exceptions illustrate that resources and non-industrial pursuits are not always a curse. Systematic evidence from the recent past is examined by Robinson et al. (2006) and Mehlum et al. (2006) who argue that where property rights, institutions, and political arrangements promote stability and efficiency resources do not bring a curse. What we know from the United States case is that government involvement in the US geological survey, promotion of technical ability and research in geology, agriculture, and metallurgy and active entrepreneurs seeking to capitalize on a resource-oriented industrialization enabled a high income/high growth outcome in the United States. In Australia, McLean (2004) notes a parallel development of public support for agricultural research. The history lesson is that even in agriculture, technical advance is not doomed to be slow. Olmstead and Rhode (2008) survey US agricultural development over the 19th century and provide ample evidence that innovation, based on careful experimentation and adaptation to climate and pest environments, allowed for rapid land and labor productivity advances in crops and livestock. What remains to be investigated systematically and using the experience of many countries over time is the determinants of demand for skill in the primary and agricultural sector and the political and economic determinants of the supply of institutions necessary to satisfy such demands.

The case of Mexico also provides interesting evidence in this debate. During the late-19th century, Mexico is often seen as an early example of export-led growth based on primary products. Catão (1998) finds however that the Mexican export sector was isolated from the modern sector. Silver, lead, copper, and petroleum made up 80% or more of total exports generating substantial surpluses for the Mexican economy. Mining processes used in Mexico were apparently highly capital intensive leading to no significant increases in the demand for labor as a complement in the production process. Forward linkages between the mining and industrial sector also never materialized despite their theoretical plausibility. As an illustration, it is estimated that only about 5% of the total mineral production of Mexico was refined or smelted domestically. Lead was shipped to New York to be refined rather than in Mexico itself. While imports failed to keep pace with exports, little of the surplus appears to have been absorbed domestically. Henequen production displayed evidence of the Dutch Disease emphasized by Williamson. Foreign expertise and capital were ubiquitous—machinery and equipment for an entire henequen processing facility were reportedly shipped to Mexico from abroad in the late ninetieth century (Beatty, p. 400). It does not appear that barriers to trade in technology were a serious constraint. Instead, social and political problems persisted. Local land-owners preferred to over consume foreign luxury goods rather than save the proceeds of their surpluses which might have financed further growth.

Nevertheless, much of this story has been re-interpreted. Beatty (2000) documents a mineral-led export boom, but notes Mexico's exports were only less diversified than Argentina and Peru out of all Latin American producers. Mexico's government played an active role in encouraging mining and primary production by sponsoring infrastructure investment which lowered the cost of such exploitation. By the early 20th century, Mexico's government attempted an early version of import substitution by raising tariffs on selected industrial imports and crafting a tax policy that subsidized various industries. The data reveal that many new industries had grown up to supplant imports of consumer goods during this period:

> *The list of such industries…includes both manufactured goods destined for consumers as final products (such as cigarettes, cotton textiles, beer, soap, footwear, candles, paper, ink, and food products) as well as goods destined for use in various kinds of extractive and manufacturing processes (such as diverse iron and steel products explosives, window and bottle glass, cement, bricks, paints, leather, chemicals, and processed minerals)…(Beatty, 2000 p. 419)*

Cortes Conde (1992) summarizes similar evidence from Argentina and Brazil prior to World War I. He suggests that in these nations there was a burgeoning local industry like in Mexico. Clearly, something more than de-industrialization was going on in these places.

8.9. CONCLUSIONS

Coming forward into the 20th century, the connections between growth and globalization have exhibited an equally complicated relationship as during the golden age of globalization prior to World War I. The two decades of instability between the World Wars witnessed an up and then a down in global integration. The 1920s gave back some of the global linkages established prior to 1914 that were eroded by the war, and growth was relatively strong as nations stabilized. The 1930s witnessed varying outcomes, but, by and large, growth resumed as trade recovered.

Great Britain and its trade partners in the Imperial Preferences system forged a path to recovery by forming a trade bloc. British devaluation also helped fuel recovery via exports as in many other nations. Germany, with the aid of its satellites in the Reichsmark bloc and significantly more autarkic policies, re-armed for total war and witnessed economic recovery in the late 1930s. Meanwhile, the United States (re-) negotiated low tariffs with its trade partners from 1934 under the Reciprocal Trade Acts thus reversing the inability of periphery nations to export their way to solvency. Latin American withdrawal from world markets has its roots in the experiences of the interwar period. As a consequence of all of this, while trade re-emerged, along with recovery from the Depression, trade grew much more slowly than world output after 1933 (Madsen, 2001).

After World War II, the best evidence is that tariffs and growth were negatively related. Latin American nations imposed prohibitive tariffs intended to spur local industry and reduce reliance on foreign manufactures. Taylor (1999) documents significantly lower investment and hence lower incomes for these nations. Estevadeordal and Taylor (2008) provide recent evidence that nations that raised the price of foreign capital goods and machinery via trade policy experienced lower growth.

From the 1970s onwards, cross-country economic evidence purported to show that distortions induced by trade policy and exchange controls were associated with lower economic performance. This contributed to the building of a Washington Consensus that liberal policies were best for growth. Subsequent analysis of the East Asian Miracle by Rodrik (1995) argued that rather than engage in a laissez-faire model of growth, Taiwan and South Korea got "certain interventions right." In South Korea, the government promoted heavy investment and accumulation of foreign equipment prior to the export boom. The early investment push led to a subsequent export boom. Rather than export-led growth, this successful program involved a number of distortions that subsidized and encouraged capital accumulation. Additional factors in the Korean case must include good initial conditions of high human capital, favorable demographics, and low levels of wealth inequality.

The lesson from the literature then is not necessarily that closing up is beneficial when a nation has fallen behind. The most obvious example is the dismal economic failure of North Korea, but many other nations have experienced equally poor outcomes by closing themselves off. Other conditions must obtain to achieve long-run economic growth in closed economies. And again, certain economies have grown *despite* high tariffs. Trade itself may not yield higher growth rates as some evidence suggests for the economies that de-industrialized in the 19th century. In terms of other forms of globalization, we have learned that international labor mobility acted as a force for convergence in the 19th century. Since the interwar period, global migration became a pale shadow of its former self, ruling out the possibility that the international movement of people has played as decisive of a role for understanding growth and convergence in the global economy over the last 70 years.

As for global capital markets, the historical record shows that the large cross-border flows of the 19th century were a force for growth in the receiving countries but also led to divergence. Capital flowed to regions with abundant natural resources and working age males. Foreign capital helped raise incomes in institutionally advanced commodity-based exporters and industrializers. Elsewhere, these flows exposed many nations to financial crises. These events dealt large shocks to many receiving and sending countries across the 19th and 20th century. The Great Depression represented the end of the slavish adherence to the gold standard at historically defined parities and the severe constraints on achieving domestic macroeconomic balance this institution engendered. The rise of representative democracy and a lack of international cooperation made the gold standard hard to maintain in the face of capital flows in the advanced core from the 1930s (Eichengreen, 1992). Nations also opted to forego significant international capital flows for many decades after World War II due to their fear of the destabilizing speculation they witnessed in the 1930s. The strong resurgence of those capital flows beginning in the late 20th century has been associated with the 1980s debt crisis, the Asian financial crisis of 1997–1998, and even the global crisis that began in 2007–2008. The research for the modern period prescribes proper sequencing of financial liberalization and other potential pitfalls to wholesale liberalization. Domestic financial conditions, policies, and institutions need to be "adequate" and "sound" before opening up to international markets can yield positive growth benefits (Klein and Olivei, 2008). This means that is has been hard to find convincing evidence that liberalized capital accounts have always, over the long run, been associated with significantly stronger growth in the long run in all cases.

Adam Smith and David Ricardo's logic that free trade brings benefits to both parties of the exchange is impeccable. Nations have generally experienced aggregate income gains from specialization whether they were distributed equally or not. But free trade may not yield the efficient outcome when more complex environments are considered and the program to be solved becomes dynamic. Using the long run of history to investigate the relationship between growth and globalization reveals that the relationship between these two outcomes is rather complex and nuanced. The historical record largely confirms that there is no one-way positive relationship between the growth and globalization at all times and for all countries.

It now appears that when globalization has demanded specialization in natural resources, Dutch Disease, low investment and physical conflict over "rents" can occur. Conflict over the right to control such endowments is clearly inefficient and it is very likely to be catalyzed by the incentive to export these goods to world markets and the ability to finance their exploitation with foreign capital under imperfect governance structures (Ross, 2005; Bulte et al. 2011). This suggests one way in which institutional pre-requisites must be satisfied to fully enjoy the static *and* dynamic gains from trade.

Dynamic models that predict specialization in primary production by some countries may lead to welfare losses for future generations in places where globalization delivers low rates of human capital accumulation and lower TFP growth. Further research should sort out whether the historical record is at odds or is consistent with this view. In particular, it remains to be seen whether the specialization in primary products with low-productivity growth is a function of other underlying factors or not. If it is, then trade and specialization may not be at fault per se.

A second inefficiency associated with globalization arises when international capital markets are open. Information asymmetries; the unenforceability of repayment of debts; strong lender and borrower moral hazard; and other market imperfections generate inefficient levels of international lending. These forces can create significant welfare losses, crises, and economic volatility in the growth rate of liberalized economies. These market imperfections are surely more important in international markets than in domestic markets where sovereignty is not an issue and where regulations are typically better enforced. Measured consideration of the benefits and costs of cross-border flows is necessary after a careful look at the long-run record.

The historical record suggests therefore that liberalizing international markets is not necessarily a policy that will raise economic growth. Other pre-conditions and other policies to promote growth seem to be just as important, if not more so, in many cases. All of this does *not* argue that adding further distortions to the policy mix by limiting the process of globalization in some narrow sense via tariffs or closed capital accounts is necessarily advantageous. Instead, the long-run historical record merely reminds us that economists be significantly more cautious when making the claim that globalization or free trade is unambiguously efficient.

ACKNOWLEDGMENTS

The generous support of the Center for the Evolution of the Global Economy (CEGE) at the University of California, Davis; and Alan M. Taylor is gratefully acknowledged. I acknowledge Rivka Shenhav for excellent research assistance and early comments. Errors are solely attributable to the author.

REFERENCES

Accominotti, O., Flandreau, M., 2008. Bilateral treaties and the most-favored-nation clause: the myth of trade liberalization in the nineteenth century. World Politics 60 (2), 147–188.

Acemoglu, D., Johnson, S., Robinson, J.A., 2001. The colonial origins of comparative development: an empirica. Investigation. American Economic Review 91 (5), 1369–1401.

Acemoglu, D., Johnson, S., Robinson, J.A., 2005. The rise of europe: Atlantic trade, institutional change, and economic growth. American Economic Review 95 (3), 546–579.

Allen, R.C., 2001. The great divergence in European wages and prices from the middle ages to the First World War. Explorations in Economic History 38, 411–447.

Allen, R.C., 2009. The British Industrial Revolution in Global Perspective. Cambridge University Press, Cambridge.

Anderson, J.E., 1979. A theoretical foundation for the gravity equation. The American Economic Review 69, 106–116.

Anderson, J.E., 2011. The gravity model. Annual Review of Economics 3, pp. 133–160.

Anderson, J.E., van Wincoop, E., 2004. Trade costs. Journal of Economic Literature 42, 691–751.

Arkolakis, C., Costinot, A., Rodríguez-Clare, A., 2012. New trade models, same old gains? American Economic Review 102 (1), 94–130.

Bairoch, P., 1972. Free trade and European economic development in the 19th century. European Economic Review 3 (3), 211–45.

Beatty, E.N., 2000. The impact of foreign trade on the Mexican economy: terms of trade and the rise of industry, 1880–1923. Journal of Latin American Studies 32 (2), 399–433.

Bergstrand, J.H., 1985. The gravity equation in international trade: some microeconomic foundations and empirical evidence. The Review of Economics and Statistics 67 (3), 474–481.

Blattman, C., Clemens, M., Williamson, J.G., 2002. Who Protected and Why? Tariffs the World Around 1870–1938 Mimeo. Department of Economics Harvard University.

Bordo, M.D., 2003. The globalization of international financial markets: what can history teach us? In: Auernheimer, L. (Ed.), International Financial Markets: The Challenge of Globalization, University of Chicago Press, Chicago, pp. 29–78.

Bordo, M.D., Meissner, C.M., 2011. Foreign capital, financial crises and incomes in the first era of globalization, European Review of Economic History 15 (1), 61–91.

Bordo, M.D., Rockoff, H., 1996. The gold standard as a good housekeeping seal of approval. Journal of Economic History 56, 389–428.

Bulte, E., Meissner, C.M., Sarr, M., Swanson, T., 2011. On the looting of nations. Public Choice 148 (3–4), 353–380.

Catão, L.A., 1998. Mexico and export-led growth: the porfirian period revisited. Cambridge Journal of Economics 22 (1), 59–78.

Clark, G., Feenstra, R.C., 2003. Technology in the great divergence. In: Bordo, M.D., Taylor, A.M., Williamson J.G. (Eds.), Globalization in Historical Perspective. University of Chicago Press, Chicago, pp. 277–322.

Clark, G., O'Rourke, K.H., Taylor, A.M., 2008. Made in America? the new world, the old, and the industrial revolution. American Economic Review 98 (2), 523–528.

Clemens, Michael A., Williamson, J.G., 2004a. Wealth bias in the first global capital market boom, 1870–1913. Economic Journal 114, 304–337.

Clemens, Michael A., Williamson, J.G., 2004b. Why did the tariff-growth correlation change after 1950? Journal of Economic Growth 9 (1), 5–46.

Coleman, A., 2007. The pitfalls of estimating transactions costs from price data: evidence from trans-Atlantic gold-point arbitrage, 1886–1905. Explorations in Economic History 44 (3), 387–410.

Cortes Conde, R., 1992. Export-led growth in Latin America: 1870–1930. Journal of Latin American Studies 24, 163–179.

Crafts, N.F.R., 1999. East asian growth before and after the crisis. IMF Staff Papers 46 (2), 139–166.

Crouzet, F., 1981. The sources of England's wealth: some French views in the eighteenth century. In: Cottrell, P.L., Aldcroft, D.H., (Eds.), Shipping, trade and commerce: essays in memory of Ralph Davis. Leicester University Press, London, pp. 61–79.

Darity, W.A., 1992. British industry and the West Indies plantations. In: Inikori, Joseph E., Engerman, Stanley L. (Eds.), Trade, The Atlantic Slave. Duke University Press, Durham NC, pp. 247–279.

David, P., 1970. Learning by doing and tariff protection: a reconsideration of the case of the Ante-Bellum United States cotton textile industry. Journal of Economic History 30 (3), 521–601.

Desmet, K., Parente, S.L., 2009. The evolution of markets and the revolution of industry: a quantitative model of England's development, 1300–2000 mimeo. University of Illinois Department of Economics.

DeVries, J., 1994. The industrial revolution and the industrious revolution. The Journal of Economic History 54 (2), 249–270.

De Vries, J., 2008. The Industrious Revolution. Cambridge University Press, New York.

DeVries, J., 2010. The limits of Globalisation in the early modern world. Economic History Review 63 (3), 710–733.

Diamond, J.M., 1999. Guns, Germs, and Steel: The Fates of Human Societies. W.W. Norton & Company, New York.

Donaldson, D., 2008. Railroads of the Raj: estimating the impact of transportation infrastructure. American Economic Review.

Drelichman, M., 2005. The curse of Moctezuma: American silver and the Dutch disease, 1501–1650. Explorations in Economic History 42 (3), 349–380.

Eaton, J., Kortum, S., 2001. Technology, trade, and growth: a unified framework. European Economic Review 45 (4–6), 742–755.

Eichengreen, B., 1992. Golden Fetters, The Gold Standard and the Great Depression, 1919–1939. Oxford University Press, New York.

Eltis, D., Engerman, S.L., 2000. The importance of slavery and the slave trade to industrializing Britain. Journal of Economic History 60, 123–144.

Estevadeordal, A., Taylor, A.M., 2008. Is the Washington Consensus Dead? Growth, Openness, and the Great Liberalization, 1970s–2000s. NBER Working Paper 14264.

Estevadeordal, A., Frantz, B., Taylor, A.M., 2003. The rise and fall of world trade, 1870–1939. The Quarterly Journal of Economics 118 (2), 359–407.

Esteves, R.P., 2007. Qui custodiet quem? Sovereign Debt and Bondholders' Protection before 1914. Mimeo Department of Economics, Oxford University.

Feenstra, R.C., 1994. New product varieties and the measurement of international prices. American Economic Review 84 (1), 157–77.

Feenstra, R., Obstfeld, M., Russ, K.N., 2011. In search of the armington elasticity. Mimeo UC Davis Department of Economics.

Ferguson, N., Schularick, M., 2006. The empire effect: the determinants of country risk in the first age of Globalization, 1880–1913. Journal of Economic History 66 (2), 283–212.

Feyrer, J., 2009. Trade and Income—Exploiting Time Series in Geography. NBER Working Paper 14910.

Findlay, R., 1982. International distributive justice: a trade theoretic approach. Journal of International Economics 13 (1–2), 1–14.

Findlay, R., O'Rourke, K.H., 2003. Commodity market integration, 1500–2000. In: Bordo, M.D., Taylor, A.M., Williamson, J.G. (Eds.), Globalization in Historical Perspective. University of Chicago Press, Chicago, pp. 13–64.

Findlay, R., O'Rourke, K.H., 2007. Power and Plenty: Trade, War, and the World Economy in the Second Millenium. Princeton University Press, Princeton, NJ.

Fishlow, A., 1985. Lessons from the past: capital markets during the 19th century and the interwar period. International Organization 39 (3), 383–439.

Flandreau, M., Maurel, M., 2005. Monetary union, trade integration, and business cycles in 19th century Europe. Open Economies Review 16 (2), 135–152.

Flynn, D.O., Giraldez, A., 2004. Path dependence, time lags and the birth of globalization: a critique of O'Rourke and Williamson. European Review of Economic History 8, 81–108.

Frankel, J.A., Romer, D., 1999. Does trade cause growth? The American Economic Review 89 (3), 379–399.

Galor, O., 2004. From stagnation to growth: unified growth theory. In: Durlauf, P.A. (Ed.), Handbook of Economic Growth. North-Holland.

Galor, Oded, Mountford, A., 2008. Trading population for productivity: theory and evidence. The Review of Economic Studies 75 (4), 1143–1179.

Galor, O., Weil, D.N., 2000. Population, technology, and growth: from Malthusian stagnation to the demographic transition and beyond. American Economic Review 90 (4), 806–828.

Gourinchas, P.O., Jeanne, O., 2006. The elusive gains from international financial integration. Review of Economic Studies 73 (3), 715–741.

Hatton, T.J., Williamson, J.G., 1994. What drove the mass migrations from Europe in the late nineteenth century? Population and Development Review 20 (3), 533–559.

Head, K., 1994. Infant industry protection in the steel rail industry. Journal of International Economics 37 (3), 141–165.

Head, K., Ries, J., 2001. Increasing returns versus national product differentiation as an explanation for the pattern of US-Canada trade. American Economic Review 91 (4), 858–876.

Huberman, M., 2008. Ticket to trade: Belgian labour and globalization before. The Economic History Review 61 (2), 326–359.

Inikori, J., 1987. Slavery and the development of industrial capitalism in England. In: Solow, Barbara, Engerman, Stanley (Eds.), British Capitalism and Caribbean Slavery: The Legacy of Eric Williams. Cambridge University Press, Cambridge, 79–101.

Irwin, D.A., 2000. Did late-nineteenth-century U.S. tariffs promote infant industries? evidence from the tinplate industry. Journal of Economic History 60 (2), 335–360.

Irwin, D.A., 2002. Interpreting the tariff-growth correlation of the late nineteenth century. American Economic Review 92 (2), 165–169.

Irwin, D.A., 2010. Trade restrictiveness and deadweight losses from US tariffs. American Economic Journal: Economic Policy 2 (3), 111–133.

Irwin, D.A., Tervio, M., 2002. Does trade raise income?: Evidence from the twentieth century. Journal of International Economics 58 (1), 1–18.

Jacks, D.S., 2006. What drove 19th century commodity market integration? Explorations in Economic History 43, 383–412.

Jacks, D.S., Meissner, C.M., Novy, D., 2011. Trade booms, trade busts and trade costs. Journal of International Economics 83 (2), 185–201.

Jones, C., Romer, P., 2009. The New Kaldor Facts: Ideas, Institutions, Population, and Human Capital. NBER Working Paper 15049.

Keay, I., 2007. The engine or the caboose? resource industries and twentieth century Canadian economic Performance. Journal of Economic History 67 (1), 1–32.

Klein, M., Olivei, G.P., 2008. Capital account liberalization, financial depth, and economic growth. Journal of International Money and Finance 27 (6), 861–875.

Lampe, M., 2009. Effects of Bilateralism and the MFN clause on international trade—evidence for the Cobden-Chevalier network, (1860–1875). The Journal of Economic History 69 (4), 1012–1040.

Lehman, S., O'Rourke, K.H., 2011. The structure of protection and growth in the late nineteenth century. Review of Economics and Statistics 93 (2), 606–616.

Lew, B., Cater, B., 2006. The telegraph, co-ordination of tramp shipping, and growth in world trade, 1870–1910. European Review of Economic History 10 (2), 147–173.

Liu, D., Meissner, C.M., 2012. Market Potential and the Rise of U.S. Productivity Leadership. NBER Working Paper 18819.

López-Córdova, J.E., Meissner, C.M., 2003. Exchange rate regimes and international trade: evidence from the classical gold standard era, 1870–1913. American Economic Review 93 (1), 344–353.

Lucas, R.E., 2009. Trade and the diffusion of the industrial revolution. American Economic Journal: Macroeconomics 1 (1), 1–25.

Madsen, J.B., 2001. Trade barriers and the collapse of world trade during the Great depression. Southern Economic Journal 67, 848–68.

Madsen, J.B., 2007. Technology spillover through trade and TFP convergence: 135 years of evidence for the OECD countries. Journal of International Economics 72 (2), 464–480.

Mantoux, P., 1961. The Industrial Revolution in the Eighteen Century, revised ed. Haper Torchbooks, New York.

Mauro, P., Sussman, N., Yafeh, Y., 2006. Emerging Markets and Financial Globalization: Sovereign Bond Spreads in 1870–1913 and Today. Oxford University Press, Oxford.

McCloskey, D.N., 1970. Britain's loss from foreign industrialization: a provisional estimate. Explorations in Economic History 8 (2), 141–152.

McLean, I., 2004. Australian economic growth in historical perspective. The Economic Record 80 (250), 330–345.

Mehlum, H., Moene, K., Torvik, R., 2006. Institutions and the resource curse. Economic Journal 116, 1–20.

Mitchener, K.J., Weidenmier, M., 2005. Empire, public goods, and the Roosevelt corollary. Journal of Economic History 65, 658–692.

Mokyr, J., 2010. The Enlightened Economy. Yale University Press, New Haven.

Newberry, D., Stiglitz, J., 1984. Pareto inferior trade. Review of Economic Studies 51 (1), 1–12.

North, D.C., Weingast, B.R., 1989. Constitutions and commitment: the evolution of institutional governing public choice in seventeenth-century England. The Journal of Economic History 49 (4), 803–832.

Novy, D., 2013. International trade without ces: estimating translog gravity. Journal of International Economics. 89 (2), 271–282.

Nunn, N., 2008. The long-term effects of Africa's slave trades. The Quarterly Journal of Economics 123 (1), 139–176.

Nunn, N., Wantchekon, L., 2009. The Slave Trade and the Origins of Mistrust in Africa. NBER Working Paper 14783.

O'Brien, P., 1982. European economic development: the contribution of the periphery. The Economic History Review 35, 1–18.

Obstfeld, M., Taylor, A.M., 2003. Sovereign risk, credibility and the gold standard: 1870–1913 versus 1925–31. Economic Journal 113 (487), 241–275.

Obstfeld, M., Taylor, A.M., 2004. Global Capital Markets: Integration, Crisis, and Growth. Cambridge University Press, New York.

Olmstead, A.L., Rhode, P.W., 2008. Creating Abundance. Cambridge University Press, Cambridge, New York.

O'Rourke, K.H., 2000. Tariffs and growth in the late 19th century. Economic Journal 110 (463), 456–483.

O'Rourke, K.H., Williamson, J.G., 1994. Late 19th century Anglo-American factor price convergence: were Heckscher and Ohlin right? Journal of Economic History 54 (4), 892–916.

O'Rourke, K.H., Williamson, J.G., 1997. Around the European periphery 1870 1913: Globalization, schooling and growth. European Review of Economic History 1 (2), 153–190.

O'Rourke, K.H., Williamson, J.G., 1999. Globalization and History. MIT Press, Cambridge, Mass.

O'Rourke, K.H., Williamson, J.G., 2002. When did globalisation begin? European Review of Economic History 6 (1), 23–50.

O'Rourke, K.H., Williamson, J.G., 2009. Did Vasco da Gama matter for European markets? testing Frederick Lane's hypotheses fifty years later. Economic History Review 62 (3), 655–84.

O'Rourke, K.H., Taylor, A.M., Williamson, J.G., 1996. Factor price convergence in the late nineteenth century. International Economic Review 37 (3), 499–530.

Pinkovskiy, M., Sala-i-Martin, X., 2009. Parametric Estimations of the World Distribution of Income. NBER Working Paper 15433.

Rancière, R., Tornell, A., Westermann, F., 2008. Systemic crises and growth quarterly. Journal of Economics 123 (1), 359–406.

Ravenstein, E.G., 1889. The laws of migration. Journal of the Royal Statistical Society 52 (2), 241–305.

Rivera-Batiz, L.A., Romer, P.M., 1991. Economic integration and endogenous growth. Quarterly Journal of Economics 106 (2), 531–555.

Robinson, J.A., Torvik, R., Verdier, T., 2006. Political foundations of the resource curse. Journal of Development Economics 79 (2), 447–468.

Rodrik, D., 1995. Getting interventions right: how South Korea and Taiwan grew rich. Economic Policy 20, 55–107.

Romer, P.M., 1994. New goods, old theory, and the welfare costs of trade restrictions. Journal of Development Economics 43 (1), 5–38.

Romer, P.M., 1996. Why, indeed, in America? theory, history, and the origins of modern economic growth. The American Economic Review 86 (2), 202–206.

Rosés, Joan R., 2003. Why isn't the whole of Spain industrialized? new economic geography and early industrialization, 1797–1910. Journal of Economic History 63 (4), 995–1022.

Ross, M., 2005. Booty Futures Mimeo. Department of Political Science, University of California, Los Angeles.

Schularick, M., 2006. A tale of two globalizations: capital flows from rich to poor in two eras of global finance. International Journal of Finance & Economics 11 (4), 339–354.

Schularick, M., Solomou, S., 2011. Trade and Growth: Historical Evidence. Journal of Economic Growth 16 (1), 33–70.

Schularick, M., Steger, T., 2010. Financial integration, investment, and economic growth: evidence from two eras of financial globalization. The Review of Economics and Statistics 92 (4), 756–768.

Taylor, A.M., 1999. On the costs of inward-looking development: price distortions, growth, and divergence in Latin America. Journal of Economic History 58 (1), 1–28.

Taylor, A.M., Williamson, J.G., 1997. Convergence in the age of mass migration. European Review of Economic History 1 (1997), 27–63.

Vamvakidis, A., 2002. How robust is the growth-openness connection? historical evidence. Journal of Economic Growth 7 (1), 57–80.

Voth, J., Hersh, J., 2009. Sweet Diversity: Colonial Goods and the Rise of European Living Standards After 1492. C.E.P.R. Discussion Paper 7386.

Watkins, M., 1963. The staple theory of economic growth. Canadian Journal of Economics and Political Science 29 (2), 141–158.

Williams, E., 1944. Capitalism and Slavery. University of North Carolina Press, Chapel Hill, NC.

Williamson, J., 1964. American Growth and the Balance of Payments. University of North Carolina Press, Chapel Hill, NC.

Williamson, J.G., 1995. The evolution of global labor markets since 1830: background, evidence and hypotheses. Explorations in Economic History 32, 141–196.

Williamson, J.G., 2011. Trade and Poverty: When the Third World Fell Behind. MIT Press, Cambridge, Mass.

Wilson, Ted, 2001. Battles for the Standard: Bimetallism and the Spread of the Gold Standard in the Nineteenth Century. Ashgate, Aldershot, UK.

Wright, Gavin, 1990. The origins of American industrial success, 1879–1940. American Economic Review 80 (4), 651–68.