Economic Growth

New Directions in Theory and Policy

Edited by **Philip Arestis, Michelle Baddeley** and **John S.L. McCombie**

Economic Growth

New Directions in Theory and Policy

Edited by

Philip Arestis

University Director of Research, Cambridge Centre for Economic and Public Policy, Department of Land Economy, University of Cambridge and Fellow of Wolfson College, UK

Michelle Baddeley

Fellow and Director of Studies in Economics, Gonville and Caius College, Cambridge and member, Faculty of Economics, University of Cambridge, UK

John S.L. McCombie

Director, Cambridge Centre for Economic and Public Policy, Department of Land Economy, University of Cambridge and Fellow of Downing College, UK

Edward Elgar

Cheltenham, UK • Northampton, MA, USA

2007

Contents

List of figures vii
List of tables viii
List of contributors x

1 Introduction 1
Philip Arestis, Michelle Baddeley and John S.L. McCombie
2 Is growth theory a real subject? 8
Franklin M. Fisher
3 What is endogenous growth theory? 14
Mark Roberts and Mark Setterfield
4 Is the natural rate of growth exogenous? 32
Miguel Leon-Ledesma and A.P. Thirlwall
5 The representative firm and increasing returns: then and now 44
Stephanie Blankenberg and G.C. Harcourt
6 A dynamic framework for Keynesian theories of the business cycle and growth 65
Pedro Leão
7 A Keynesian model of unemployment and growth: theory 83
John Cornwall
8 A Keynesian model of unemployment and growth: an empirical test 101
Wendy Cornwall
9 The relevance of the Cambridge–Cambridge controversies in capital theory for econometric practice 117
G.C. Harcourt
10 Foreign direct investment and productivity spillovers: a sceptical analysis of some OECD economies 136
Carlos Rodríguez, Carmen Gomez and Jesus Ferreiro
11 Increasing returns and the distribution of manufacturing productivity in the EU regions 151
Bernard Fingleton and Enrique López-Bazo
12 The role of wage-setting in a growth strategy for Europe 178
Andrew Watt

13 Economic growth and beta-convergence in the East European Transition Economies 200
Nigel F.B. Allington and John S.L. McCombie

14 Knowledge externalities and growth in peripheral regions 223
Fabiana Santos, Marco Crocco and Frederico Jayme Jr

15 Knowledge, human capital and foreign direct investment in developing countries: recent trends from an endogenous growth theory perspective 238
Diana V. Barrowclough

16 Is growth alone sufficient to reduce poverty? In search of the trickle down effect in rural India 259
Santonu Basu and Sushanta Mallick

17 Strategy for economic growth in Brazil: a Post Keynesian approach 279
José L. Oreiro and Luiz Fernando de Paula

Index 305

Contributors

Nigel F.B. Allington, Downing College, Cambridge and Cardiff University, UK

Philip Arestis, University of Cambridge, UK

Michelle Baddeley, University of Cambridge, UK

Diana V. Barrowclough, UNCTAD, Geneva, Switzerland

Santonu Basu, Queen Mary College, University of London, UK

Stephanie Blankenberg, School of Oriental and Asian Studies, London, UK

John Cornwall, Dalhousie University, Canada

Wendy Cornwall, Mount Saint Vincent University, Canada

Marco Crocco, Universidade Federal de Minas Gerais, Brazil

Jesus Ferreiro, University of the Basque Country, Spain

Bernard Fingleton, University of Cambridge, UK

Franklin M. Fisher, Massachusetts Institute of Technology, USA

Carmen Gomez, University of the Basque Country, Spain

Frederico Jayme Jr, Universidade Federal de Minas Gerais, Brazil

G.C. Harcourt, Jesus College, Cambridge, UK

Pedro Leão, Universidade Técnica de Lisboa, Portugal

Miguel Leon-Ledesma, University of Kent, UK

Enrique López-Bazo, Universitat de Barcelona, Spain

Sushanta Mallick, Queen Mary College, University of London, UK

John S.L. McCombie, University of Cambridge, UK

José L. Oreiro, Federal University of Paraná, Brazil

Luiz Fernando de Paula, University of the State of Rio de Janeiro, Brazil

Mark Roberts, University of Cambridge, UK

Carlos Rodríguez, University of the Basque Country, Spain

Fabiana Santos, Universidade Federal de Minas Gerais, Brazil

Mark Setterfield, Trinity College, Hartford, USA

A.P. Thirlwall, University of Kent, UK

Andrew Watt, European Trade Union Institute, Brussels

1. Introduction

Philip Arestis, Michelle Baddeley and John S.L. McCombie

In September 2005, the *Cambridge Centre for Economic and Public Policy* – based in the Land Economy department of the University of Cambridge, UK – hosted its second official conference. The theme selected for this conference focused on the nature, causes and features of economic growth across a range of countries and regions. This volume is a collection of some of the key papers presented at this conference, and they address a broad range of growth-related topics – from theoretical analyses of economic growth in general to empirical analyses of growth in the OECD, transition economies and developing economics.

In Chapter 2, this volume begins with 'Is growth theory a real subject?', in which Franklin Fisher presents the paper given at the conference as an after-dinner talk. It has been left in that form rather than making it more formal because the anecdotes are interesting as well as amusing. But the paper's informality should not conceal that fact that, together with the sugar, it is administering very strong and bitter medicine to growth theory and to macroeconomics generally. Franklin Fisher questions the widespread use of aggregate production functions in growth theory and also raises the issue of what is meant by the words 'capital', 'investment', 'labor', 'productivity' and 'output'. These concepts cannot be freely used as though they are related by a production function. Macroeconomic theories must question these links if they are to be more than a spurious application of microeconomic theory.

In Chapter 3, 'What is endogenous growth theory?', Mark Roberts and Mark Setterfield provide a critical survey of the literature on endogenous growth theory. Two definitions of endogenous growth theory are initially identified, on the basis of which the substantive content of various different models of endogenous growth is then explored. A particularly important point that emerges from this analysis is that endogenous growth can be either 'Keynesian' or 'neoclassical' in nature. A third definition of endogenous growth theory is then introduced that permits identification and exploration of a special subset of endogenous growth models in which the growth process is path dependent.

In Chapter 4, ‘Is the natural rate of growth exogenous’, Miguel Leon-Ledesma and A.P. Thirlwall examine the question of whether the natural rate of growth is exogenous or endogenous to demand, and whether it is input growth that causes output growth or vice versa. This question lies at the heart of the debate between neoclassical growth economists on the one hand, who treat the rate of growth of the labour force and labour productivity as exogenous to the actual rate of growth, and economists in the Keynesian/post-Keynesian tradition on the other, who maintain that growth is primarily demand-driven because labour force growth and productivity growth respond to demand growth. The latter view does not imply that demand growth determines supply growth without limit; rather, aggregate demand determines aggregate supply over a range of full employment growth rates. In most countries, demand constraints tend to bite long before supply constraints are ever reached. This paper shows that there is an easy way to discriminate between these competing hypotheses by using a simple technique to estimate the natural growth rate, and then estimating whether the natural growth rate is significantly higher in periods when the actual growth rate exceeds the natural rate, and lower when the actual growth rate is below the natural rate. The model is tested for 15 OECD countries, and the results support the view that the natural rate of growth is not independent of the actual rate of growth – just as the ‘natural’ rate of unemployment is not independent of the actual rate of unemployment! The theoretical message is that to understand growth-rate differences between countries, even over substantial periods of time, there needs to be as much focus on demand as well as supply (the focus of ‘new’ growth theory). The policy message for slow-growing countries is that they need to identify constraints on demand (such as the balance of payments, and an obsession with low inflation, for example, within the EU at the present time), as well as investing in the capacity to supply.

In Chapter 5, ‘The representative firm and increasing returns: then and now’, Stephanie Blankenburg and G.C. Harcourt return to the debates from the 1920s about the concept of increasing returns and the role of the representative firm, which culminated in the 1930 symposium in the *Economic Journal*. Underlying them were confusion about whether theory could or could not be applied to ‘real world’ happenings, clarity about which was often obfuscated by ‘Marshall’s desire to be read by businessmen’ (*sic*). Many of the issues that were raised and the confusions that occurred in these debates surfaced again in the developments in the last 20 years of endogenous growth theory. The objects of the paper are to try to clarify the exchanges between the protagonists in the 1920s and then to relate the findings to the re-emergence of similar issues and confusions in the last 20 years.

In Chapter 6, 'A dynamic framework for Keynesian theories of the business cycle and growth', Pedro Leão argues that the Keynesian multiplier-accelerator model of the cycle fails to account for the self-sustained nature of real-world booms and recessions. He recasts multiplier-accelerator models in a dynamic framework, inspired by Harrod's theory of economic growth. The results are twofold. First, the resulting model provides a satisfactory explanation for the observed self-sustained nature of booms and recessions. Second, the dynamic framework suggests that a change in investment has a greater effect on aggregate demand than on aggregate supply. This is what lies at the root of booms and recessions.

In Chapter 7, 'A Keynesian model of unemployment and growth: theory', John Cornwall presents a theory of long-run unemployment, output and productivity as a two-stage recursive process generated by the interaction of aggregate demand and aggregate supply. Institutions and the distribution of power determine the strength of aggregate demand policy, and through it the level of aggregate demand and unemployment in any period. The demand-side variables as well as supply-side variables determine the productivity growth rate. Aggregate demand change, via its impact on unemployment and investment behaviour, always induces change in aggregate supply in the same direction. The dominant role of aggregate demand defines the Keynesian character of the model, which emphasizes the direct effect of policy on aggregate demand and unemployment, and its indirect impact on growth. Also, the included structural features are determinants of performance in the short run but are changed by the system's performance in the longer run, a path-dependent process that generates transformational growth rather than the steady state growth of the neoclassical model.

In Chapter 8, 'A Keynesian model of unemployment and growth: an empirical test', Wendy Cornwall presents an empirical companion to John Cornwall's chapter. She empirically assesses John Cornwall's model, using standard econometric techniques to test the model's ability to explain unemployment and growth in a group of developed OECD economies during the second half of the twentieth century. She focuses on the level and growth of aggregate demand as the outcome of policy choices endogenous to this model – with growth of aggregate demand inducing endogeneity in aggregate supply growth too. The econometric results strongly support the importance of institutions and the distribution of power on aggregate demand outcomes as well as the link between aggregate demand and supply as emphasised in John Cornwall's chapter.

In Chapter 9, 'The relevance of the Cambridge–Cambridge controversies in capital theory for econometric practice', G.C. Harcourt returns with an assessment of the modern relevance of capital theory. By assuming that the short period and long period may be viewed as if they had collapsed into one, neoclassical economists have been able to specify econometric models and use actual data in order to estimate values of parameters that are not directly observable. This is the basis, for example, of the procedures adopted by Arrow et al. (1962) in their work on CES production functions. The capital theory results show that this procedure is not acceptable, even if all neoclassical assumptions except the presence of vintages are accepted. The same sort of conceptual doubts are relevant for the specification procedures implied in co-integration.

In Chapter 10, 'Foreign direct investment and productivity spillovers: a sceptical analysis of some OECD economies', Carlos Rodríguez, Carmen Gomez and Jesus Ferreiro argue that one of the channels through which inward FDI can promote economic growth in host economies is the existence and absorption of productivity spillovers. This chapter is an attempt to evaluate the existence, size and direction of these externalities. Although there exists in the literature a high number of papers related to this issue, this chapter incorporates an OECD database not used in previous papers: 'Measuring Globalisation: the Role of Multinationals in OECD Economies'. This database is used to calculate the productivity of foreign and local firms in the economies/industries for which data are available. They estimate whether or not a relationship exists between the evolution of the productivity gap between foreign and local firms (a proxy for the existence and direction of productivity spillovers) and the presence (and its change) of foreign firms in the local economy/industry. Although they do not detect a generalized outcome about the existence and sign of the spillovers effect, there exist more probabilities of negative spillovers leading to a wider productivity gap between foreign and local firms (dependent on the industry under analysis).

In Chapter 11, 'Increasing returns and the distribution of manufacturing productivity in the EU regions', Bernie Fingleton and Enrique López-Bazo estimate an empirical model motivated by recent theoretical developments in urban and geographical economics. This model allows for the effects of technological externalities such as knowledge spillovers and congestion. The model emphasizes the diverse causes of regionally differentiated manufacturing productivity growth rates but provides empirical support for increasing returns to scale, which lies at the heart of contemporary theory. The effects of increasing returns are illustrated by simulations, the density function and stochastic kernels, which show how equilibrium productivity level distributions alter across EU regions assuming different degrees of returns to scale. They conclude by making some speculative suggestions about possible causes of changes in increasing returns.

In Chapter 12, 'The role of wage setting in a growth strategy for Europe', Andrew Watt argues that growth performance, particularly in the euro area, has been extremely disappointing. This chapter expounds the view that, contrary to the conventional wisdom, failures in the macroeconomic policymaking regime are largely responsible. In particular a lack of coordination between wage-setting and monetary policy is identified as a weakness that leads to tighter-than-necessary macroeconomic policy – and thus slower growth – to contain inflation. Given the extreme difficulty of achieving fundamental changes in the policy architecture, Andrew Watt examines the scope for behavioural changes in wage setting, and monetary (and to a lesser extent fiscal) policy, and for developing the coordination mechanisms between these two policy areas necessary to achieve faster non-inflationary growth in the euro area. Key elements are the coordination of wage policy around a productivity norm and the expansion and extension of an existing coordination instrument: the 'Macroeconomic Dialogue'. This chapter shows, using a simple model, that such changes, which do not require changes to the European Treaty, would be effective in terms of growth and employment. This raises numerous implementation issues. Problems of actor incentives, commitment and inter-country adjustment are discussed, focusing on the scope for trade unions to add a European dimension to their wage-bargaining strategies.

In Chapter 13, 'Economic growth and beta-convergence in the East European Transition Economies', Nigel Allington and John McCombie examine the question of whether the transition economies have exhibited any recent evidence of catching up with the EU15 countries in terms of productivity over the period 1994 to 2002. This is accomplished by estimating a number of specifications of the neoclassical beta-convergence growth model. An alternative measure of convergence, sigma-convergence, which we do not report here, measures whether or not the cross-country variation of group per capita income shrinks over time. Finding β-convergence is a necessary, although not a sufficient, condition for σ-convergence to occur.

In Chapter 14, 'Knowledge externalities and growth in peripheral regions', Fabiana Santos, Marco Crocco and Frederico Jayme Jr argue that in some models of the so-called endogenous growth theory, externalities play an important role because they are the main rationalization for the emergence of increasing returns to scale. Usually, endogenous growth models neglect the spatial dimension of these externalities, assuming that externalities and spillovers are perfectly mobile within national boundaries. This hypothesis has been a matter of debate among geographers and economists, since the former have pinpointed the role of institutional factors in constraining externalities. The aim of this chapter is to contribute to this debate by shedding light on the fact that these institutional aspects should include the centre–periphery dimension. Having this theoretical approach in mind, the paper analyses stimuli and constraints to the emergence and absorption of externalities in a peripheral environment. The authors' claim is that peripheral conditions – rather than underdevelopment conditions – constrain the generation and absorption of externalities and impose limits on their contributions to economic growth.

In Chapter 15, 'Knowledge, human capital and foreign direct investment in developing countries: recent trends from an endogenous growth theory perspective', Diana Barrowclough describes new trends in foreign direct investment (FDI) into developing countries, with a particular focus on investment in the highest value-added forms of human capital, such as knowledge, experience and technical expertise. These forms of human capital are important because they are key elements in the process of innovation and technological change, feeding into productivity, competitiveness and, ultimately, human and social development. Empirical evidence from economic activities in research and development (R&D) in developing countries is presented, against the counterpoint of trends of investment in tourism. The different experiences in each group of activities reflects the differing sources of comparative advantage, which for the most part stem from developing countries' endowments of expertise, knowledge and experience. In the R&D sector, developing countries are essentially acting as sellers of their endowments of human capital and expertise. In tourism, they are rather buyers. There is also a new and surprising extension of the trend, which sees developing countries beginning to invest outwards, into other countries. Hosts include both developed and developing countries, and the determinants of investment vary in each case, reflecting the home and host countries' respective endowments of human capital.

In Chapter 16, 'Is growth alone sufficient to reduce poverty? In search of the trickle down effect in rural India', Santonu Basu and Sushanta Mallick present a theoretical analysis of growth and poverty in rural India. They employ several econometric tests to examine whether the trickle-down effect took place in rural India over a long time-period. They find little evidence to suggest that the trickle-down effect did occur, suggesting that the emergence of capital–labour substitution was primarily responsible for preventing growth from reducing poverty. The decline in poverty and a higher growth rate that took place during the late 1970s and 1980s were largely a result of government anti-poverty measures and the more equitable distribution of credit and inputs to smaller and marginal farmers.

In Chapter 17, 'Strategy for economic growth in Brazil: a Post Keynesian approach', José Luís Oreiro and Luiz Fernando de Paula present a Keynesian strategy for public policies aiming at higher, stable and sustained economic growth in Brazil. They hypothesize that the current poor growth performance of the Brazilian economy is due to macroeconomic and structural constraints rather than to the lack of microeconomic reforms. They recommend a strategy in which the basic features are to adopt: firstly, a crawling-peg exchange rate regime in which devaluations in domestic currency are set by the Central Bank at a rate equal to the difference between the target inflation and the average inflation rate of Brazil's most important trade partners; secondly, market-based capital controls in order to increase the autonomy of the Central Bank to set nominal interest rates according to domestic objectives (mainly to promote a robust growth); thirdly, reduction of nominal interest rates to a level compatible with a real interest rate of 6.0% per year; and fourthly, reduction of the primary surplus from the current 4.5% of GDP to 3.0% of GDP. These elements are fundamental for the required increase in the investment rate of Brazilian economy from the current 20% of GDP to 27% of GDP needed for a sustained growth of 5% per year.

We would like to thank everyone at Edward Elgar for their encouragement in the publication of this volume. Our particular gratitude and appreciation go to Warren Mosler of Valance Inc., who is also Senior Associate Fellow at the Cambridge Centre for Economic and Public Policy, University of Cambridge.

2. Is growth theory a real subject?

Franklin M. Fisher

It is, I believe, the job of an after-dinner speaker to be interesting, humorous, and provocative. I cannot guarantee the interesting part; I have hopes for the humorous requirement; but I am pretty sure about the provocative part of the job.

So let me get that part started right away. This is, I am informed, a conference about economic growth, and, at least in large part, a conference about the theory of economic growth. My problem, however (kindly sit up straight, now), is that I strongly (and correctly) believe that, properly considered, there *is* no such theory. Further, there may even be no such subject as economic growth, considered in some of the very usual ways.

Shall I sit down now before I am hustled away by the thought police? I do have some colleagues here to defend me. But I am probably safer here in Cambridge, UK, than in the United States – even though I will point out that such safety comes from an ideological rather than reasoned understanding of what I shall be talking about.

Modern growth theory makes very heavy use of a national aggregate production function – *whether it does so explicitly or not*. The problem is that, save under very restrictive conditions, aggregate production functions do not exist and that such non-existence calls into question the use of such concepts as 'total factor productivity', the 'labor-capital ratio', the 'natural rate of growth', and so forth, even the macroeconomic meaning of terms such as 'capital', 'labor', 'investment', and even 'output'. It is on that matter that I shall be reflecting this evening.

The debate over the existence of aggregate production functions goes back nearly 50 years to the start of the so-called 'Cambridge vs. Cambridge' debates.* (Geoffrey Harcourt has discussed the debates in detail a number of times – including his paper at this conference.) They began, of course, with Joan Robinson's insistence that there is no such thing as an aggregate capital stock. Curiously, however, the debates

* A remark: at least one of the Cambridges tends to be rather insular. When, in 1983, I referred to the 'Cambridge vs. Cambridge' debate, in a book on stability, the indexer for the Cambridge University Press indexed it under 'Cambridge against itself'. Fortunately, I caught that in proof and reminded the indexer that there is another Cambridge.

very largely went on in terms of *macro*economic issues. I say ‘curiously’ because one would suppose that the issues involved are purely technical, having to do with the question of whether it is possible to go from the *micro*economics of production at the firm level (or below) to the construction of a *macro* production function – a subject that began to be studied in the 1940s.

But the question was not seen as a technical one. Rather, ideological issues intruded. According to Jagdish Bhagwati, Joan Robinson once referred to the late Leif Johansen as a ‘running dog of the fascist bourgeoisie’ or some similar term of endearment. When Bhagwati said to her ‘How can you say that? Don’t you know that he is the secretary of the Norwegian Communist Party?’, she replied ‘but . . . but he believes in production functions!’

Indeed, to Joan Robinson and others at Cambridge, UK, the question of the existence of aggregate production functions, especially the existence of an aggregate capital stock, appeared crucial for the entire structure of neoclassical economics (and hence, the underpinnings of western capitalist ideology). But, of course, that is simply untrue. The existence of aggregate production functions is important for neoclassical *macro*economics – and, indeed, as we shall see, for growth theory as it is generally practiced. But aggregate production functions play no role whatever in neoclassical *micro*economics.

Nevertheless, the old debate went on mostly in *macro*economic terms. Among other things, one side used the fact that aggregate production functions appear to fit well to justify their use. The other used the fact that there were strange things about aggregate production functions – so-called ‘reswitching’ and ‘reverse capital deepening’ – to argue that aggregate production functions were wrong. And both argued in terms of macro theories of capital.

In fact, each argument was either wrong or somewhat beside the point – although, I think that, in terms of results, the Cambridge, UK side had the better position.

The argument that aggregate production functions fit well does not work. A number of authors – especially Jesus Felipé and John McCombie, in their powerful paper for this conference (Felipé and McCombie, 2006) and elsewhere, have shown that the good fit is simply the fit of an accounting identity and has nothing to do with the use of an aggregate production function.

On the other side, the very idea that reswitching or reverse capital deepening is a problem arises because one is thinking in terms of the properties that a *micro* production function has. Since the aggregation literature shows that an aggregate production function exists only under very, very special conditions, it should come as no surprise that, if you try to use one, it does not exhibit reasonable production-function properties. Arguing about the properties, rather than tackling the existence problem itself, seems a rather roundabout way to frame the debate. Certainly, as it turns out, the concentration on the nature of capital was largely misplaced.

I shall return to the aggregation issue in a moment, but I must remark that the old debate was not without its amusing moments. The best known of these occurred in the 1950s when the *Review of Economic Studies* published an article by Joan Robinson (Robinson, 1953–54), followed two years later by one by Bob Solow (Solow, 1955–56). Bob began his article by saying 'Mrs Robinson seems to have written her article in the way an oyster secretes a pearl – out of sheer irritation.'

Indeed, for a time after that, Robinson and Solow (who had never theretofore met) engaged in a spirited correspondence that Bob used to say he hoped someone would some day edit. That (for good or ill) has not occurred, but I do remember his writing: 'In calling me pigheaded, you ought to remember that what seems to you to be pigheadedness may seem to the pig to be sweet common sense.'

But enough digression. I return now to the aggregation literature itself.

I was certainly not the only one who worked on this problem. Earlier work was done by Lawrence Klein (Klein, 1946a, 1946b), Kenneth May (May, 1946, 1947), Shou-Shan Pu (Pu, 1946), and especially André Nataf (Nataf, 1948). Further, Terence Gorman (Gorman, 1968), as well as others, wrote when or after I did. (Indeed, Terence, hearing that I had written a paper on capital aggregation and not having a copy wrote his own – justly famous paper – to see what such a paper would say, and, in a real sense, wrote the paper dual to mine.) But I probably wrote more than anybody else.[1]

The issue involved is simple to state. Suppose that one wishes to aggregate the production functions of individual firms, and suppose that markets (or some other mechanism) allocate factors of production so that aggregate production is efficient. When can this be represented by an aggregate production function?

The answer is also relatively simple. *Almost never!* It is true that the job can sometimes be done: examples include cases in which relative prices remain constant, relative quantities remain constant, every firm has the same constant-returns-to-scale production function, or there is only one kind of capital, one kind of labor, and one kind of output to begin with – and all are allocated to firms to achieve efficiency. But, under more general circumstances, aggregate production functions will only exist under very restrictive conditions that surely do not hold in real economies.

Moreover, this fact has very little to do with the nature of capital. Indeed, it would remain even in production systems with no capital involved. There is not only a capital aggregation problem, but also a labor aggregation problem, and even an output aggregation problem.

This is a very serious set of results. It not only calls into question the use of aggregate production functions in growth theory (or any other kind of macromodel) but also raises the issue of what one really *means* by such terms as 'capital', 'investment', 'labor', 'productivity', or even 'output' when speaking of macro-concepts. Of course, as Humpty-Dumpty said, the question is who is to be master, oneself or the words, and you are surely free to define such concepts however you please. But you are surely not free then to use them as though they were related by a production function. Macroeconomic thinking that does so must be recast if it is to be more than microeconomic thinking writ large and misused. That applies to most of the papers at this conference.

It is important to understand the implications. This does not just mean that macroeconomists should not explicitly use aggregate production functions; it also makes illegitimate their implicit use. Every paper that talks about the macroeconomic 'supply side' in terms of the economy's capacity to produce is falling through the thin ice. Concepts such as 'total factor productivity', 'productivity growth', and 'natural rate of growth' are in deep freezing water, and even 'capital', 'labor', 'investment', and 'output' are terms that must be used with extreme care.

I shall conclude with a few more reflections.

When, in 1993, I collected my papers on this topic into book form (Fisher, 1993), I observed that I thought I had resolved the Cambridge vs. Cambridge capital debates. 'Pride goeth before a fall', and I should have known better. A few years later, I chanced to meet a graduate student at Columbia University who had been an undergraduate at Cambridge, UK. He asked me whether the existence of an aggregate capital stock was considered an important topic at MIT. I (rather disingenuously but truthfully) replied that it was not. He then informed me that at Cambridge, there was an entire course devoted to this subject, a course that he had taken. Questioning then revealed that he had no idea that I had ever written a word on the topic!

Indeed, my work, and that of Gorman and others, has nearly sunk without a trace. Macroeconomists go on doing the undoable, and, apparently, even here at Cambridge, UK, the arguments go on in the same old way. You will understand, then, why I became so enchanted with Jesus Felipé, who introduced himself to me some years ago by saying that he had written an article that began: 'I have decided to take Frank Fisher seriously.' Alas, others have not shown the same wise perspicacity, although Jesus and I (Felipé and Fisher, 2003), together with a few others like John McCombie, keep trying to tell the profession that there is deep trouble here.

One can surely ask why this is so. Perhaps the answer lies in the following story:

Nearly 25 years ago, I spent the academic year visiting Harvard (the second important university in Cambridge, MA). I taught the first-year graduate course in micro-theory. On the common wishful hope that one can allay the fears of first-year graduate students, the department had a pre-term meeting with them at which those teaching first-year courses made little presentations. By custom, *micro*economics went first (as, of course, it should). Not wishing the students to think that MIT professors were colorless personalities, I began by saying: ‘Pay no attention to what the people who come after me say. *Micro*economics is what economists really know about.’ There was nervous laughter. But when Ben Friedman, who taught the first-year *macro* course got up to speak, he began by saying ‘I agree totally with Frank Fisher. *Micro*economics *is* what economists really know about. But *macro*economics is what they *want* to know about. That’s what makes it so interesting.’

I think Ben’s view has a great deal to be said for it. But the fact that one would like to know about something is no excuse for ignoring the problems inherent in the methods that we use in attempting to gain that knowledge. That applies to growth theory – if, indeed, there is ever really to be such a subject.

NOTE

1. My papers are collected in Fisher (1993). The subject is surveyed in detail in Felipé and Fisher (2003)

REFERENCES

Felipé, J. and F.M. Fisher (2003), ‘Aggregation in Production Functions: What Applied Economists Should Know’. *Metroeconomica*, **54**, 208–62.

Felipé, J. and J. McCombie (2006), ‘The Tyranny of the Identity: Growth Accounting Revisited’, *International Review of Applied Economics*, **20**, 283–99.

Fisher, F.M. (1993), *Aggregation: Aggregate Production Functions and Related Topics*, Cambridge, MA: MIT Press.

Gorman, W.M. (1968), ‘Capital Aggregation in Vintage Models’, in J.N. Wolfe (ed.), *Value, Capital, and Growth: Papers in Honour of Sir John Hicks*, Edinburgh: University of Edinburgh Press.

Klein, L. (1946a), ‘Macroeconomics and the Theory of Rational Behavior’, *Econometrica*, **14**, 93–108.

Klein, L. (1946b), ‘Remarks on the Theory of Aggregation’, *Econometrica*, **14**, 303–12.

May, K.O. (1946), ‘The Aggregation Problem for a One-Industry Model’, *Econometrica*, **14**, 285–8.

May, K.O. (1947), 'Technological Change and Aggregation', *Econometrica*, **15**, 51–63.

Nataf, A. (1948), 'Sur la Possibilité de Construction de Certains Macromodèles', *Econometrica*, **16**, 232–44.

Pu, S. (1946), 'A Note on Macroeconomics', *Econometrica*, **14**, 299–302.

Robinson, J. (1953–54), 'The Production Function and the Theory of Capital', *Review of Economic Studies*, **21**, 81–106.

Solow, R.M. (1955–56), 'The Production Function and the Theory of Capital', *Review of Economic Studies*, **23**, 101–8.

3. What is endogenous growth theory?

Mark Roberts and Mark Setterfield

1. INTRODUCTION

Over the last 20 years, the term 'endogenous growth theory' has entered and become prominent in the parlance of both economists and policy makers alike. A keyword search for 'endogenous growth' in the *Journal of Economic Literature's* EconLit database reveals 153 records during the period 1985–90, increasing to 843 records 1991–95 and 2402 records 1996–2005, illustrating the increasing prominence of the phrase in academic literature.[1] In terms of policy making, few will forget (much less allow him to forget) then-Shadow Chancellor Gordon Brown's reference to 'post-neoclassical endogenous growth theory' in a 1994 speech. Moreover, the rhetoric of political speeches aside, the idea of endogenous growth seems to have had a very real impact on policy making over the last decade. For example, influenced by endogenous growth theory (or at least a particular variant of it), the UK Treasury has identified certain 'key drivers' of national and regional growth and has set specific performance targets relating to these drivers.[2]

But what exactly is endogenous growth theory? The purpose of this chapter is to investigate this question in detail. We begin, in Section 2, by identifying two definitions of endogenous growth theory and by contemplating both the intellectual pedigree of endogenous growth theory and the distinction between supply-led and demand-led growth processes. Section 3 is largely devoted to an examination of neoclassical endogenous growth theory, whilst Section 4 explores its Keynesian counterpart. Section 5 introduces a third definition of endogenous growth theory that permits identification of a subset of models in which growth is path dependent. Finally, Section 6 offers some conclusions.

2. ENDOGENOUS GROWTH THEORY: TWO DEFINITIONS

Economic growth involves the expansion of real output per capita and per worker over time. It is *sustained or steady* increases in real output per capita and per worker that are of primary interest – in other words, the study of growth focuses on the behaviour of the economy in the long run (periods that are longer than a single business cycle) rather than the short run.[3]

Growth as (narrowly) defined above is commonly understood to involve economy- and even society-wide processes of structural change and economic development (including, but by no means limited to, things like changes in the sectoral composition of output and employment, and changes in household fertility preferences). This 'grand vision' of growth as a process of social transformation pre-occupied Classical economists (including Smith, Ricardo and Marx), but since the genesis of modern macroeconomics in the first half of the twentieth century, the analysis of growth has, by and large, focused more narrowly on the interaction of relatively few, purely economic, variables associated with the expansion of real output.

From a contemporary vantage point, then, growth theory can be understood as the field of economics that seeks to analyse economic growth as (narrowly) defined above. *Endogenous* growth theory constitutes a particular branch of this endeavour and can be defined in one of two ways.[4] Either:

1. An endogenous growth theory is one in which the rate of growth is determined by the (equilibrium) solution of the growth model itself, rather than being imposed upon the model from without; or
2. An endogenous growth theory is one in which technical progress is explicitly modelled, rather than being treated as exogenously given 'manna from heaven'.

In fact, these definitions are not mutually exclusive – many models of endogenous growth are consistent with both.[5] But nor are the definitions equivalent, since some growth models abstract from technical change altogether but still give rise to endogenous growth by the first definition.[6] By the same token, other models explicitly theorize technical progress but the growth rate is still essentially imposed from without, in the sense that the ultimate determinants of growth are not obviously or easily amenable to change by virtue of changes in economic behaviour.[7]

Whilst the *term* 'endogenous growth theory' is of relatively recent origin (its widespread use having arisen only since the mid-1980s), the *idea* of endogenous growth theory – by either of the definitions provided above – has existed for some considerable time. For example, focusing only on the 'modern' history of growth theory,[8] both the actual and the equilibrium (warranted) rates of growth arise from solution of the model discussed (although not explicitly or 'formally' developed) by Harrod – a model that therefore satisfies the first definition of endogenous growth theory given above.[9] Meanwhile, models of growth embodying either Kaldor's technical progress function (Kaldor, 1957; see also Palley, 1996b) or Verdoorn's Law (see, for example, Dixon and Thirlwall, 1975) satisfy both the first and the second definition of endogenous growth theory above.[10]

This observation draws attention to two key points. First, the frequently-expressed idea that endogenous growth theory itself is 'new' (having originated during the 1980s) is really nothing more than a product of the relatively recent neoclassical discovery of endogenous growth, coupled with the neoclassical 'capture' of the history (and consequently the language) of growth theory that seems to have accompanied the revival of neoclassical interest in growth theory over the last two decades. Second, the distinction between supply-led (neoclassical) and demand-led (Keynesian) growth theory is every bit as important as the distinction between exogenous and endogenous growth theory, and the existence of demand-led or *Keynesian* endogenous growth theory must be taken into account if the full meaning and variety of endogenous growth theory is to be properly understood. In what follows immediately below, we devote ourselves to further discussion of these two key points.

3. THE NEOCLASSICAL 'CAPTURE' OF THE HISTORY OF GROWTH THEORY AND THE DEVELOPMENT OF NEOCLASSICAL ENDOGENOUS GROWTH THEORY

According to contemporary neoclassical accounts, developments in the field of growth theory involve little or nothing more than a simple linear evolution from the Solow (1956) model[11] (which dominated discussion during the 1950s and '60s) to the introduction of endogenous growth theory, which is identified with the pioneering contributions of Paul Romer (1986) and Robert Lucas (1988) during the 1980s. These developments, it is argued, were separated by a hiatus in growth theory during the 1970s, when the focus of macroeconomics turned away from aggregate *output* dynamics towards aggregate *price* dynamics and the analysis of inflation.[12]

Whether or not they constitute the 'be all and end all' of developments in growth theory *writ large*, there can be no doubt that the developments in *neoclassical* growth theory outlined above do involve a sea change in analysis, from exogenous to endogenous growth theory. Hence in the first generation of neoclassical growth theory pioneered by Solow (1956), the focus of attention is on the reconciliation of the actual, equilibrium and 'natural' rates of growth – the latter representing the potential rate of growth of output, as determined by the rate of growth of the labour force and technical progress.[13] In this model, the actual rate of growth will automatically converge towards an equilibrium rate that is, in turn, identical to the natural rate. But since both the rate of growth of the labour force and technical progress – and hence the natural rate of growth – are treated as exogenously given, so, too is the actual rate of growth of the economy that is 'explained' by this model.[14] Changes in economic behaviour – perhaps most significantly, changes in the rate of saving and hence the rate of accumulation of physical capital – can be shown to have no effect on the rate of growth that will be established in the long run, a rate of growth that is imposed upon the Solow model from without as an exogenous constant. In the words of Barro and Sala-i-Martin (2004, p. 18), 'we end up with a model of growth that explains everything but long run growth'.

What ultimately makes growth exogenous in the Solow model is a particular feature of the theory of production embraced by this model. Specifically, the canonical Solow model assumes that the marginal physical product of capital diminishes to zero in the limit.[15] As intimated above, the rate of accumulation of capital is explained within the Solow model – that is, it is determined endogenously – by the rate of saving. But because of the assumption it makes about the behaviour of the marginal physical product of capital, it is impossible to sustain long-run growth on the basis of capital accumulation alone in the Solow model: the contribution to total output of further additions to the stock of capital eventually dwindles to zero.[16] As such, it is impossible to connect the behavioural parameters of the model (such as the savings rate) that affect the process of accumulation to determination of the long-run growth rate.

The key innovation that renders growth endogenous in the second generation neoclassical growth theories of Romer (1986), Lucas (1988) and Rebelo (1991), and those who have since built upon these contributions, involves a change in the theory of production outlined above. Specifically, second generation neoclassical growth models – or what may be called neoclassical endogenous growth (NEG) models – assume non-diminishing marginal returns (at least in the limit) to accumulable factors of production (which include not just physical capital but also, in some models, human capital and/or the economy's stock of knowledge).[17] It is thus impossible to completely exhaust the growth potential latent in the accumulation of these factors, as a result of which the long-run growth rate *is* affected by the process of accumulation and, therefore, by the behavioural parameters of the model that explain this process. In other words, any economic behaviour that affects factor accumulation – including, for example, the savings rate – will now affect the value of the long-run equilibrium rate of growth, and a theory of endogenous growth (by the first definition of this term proffered above) emerges. Frequently, NEG models seek to justify their assumption of non-diminishing returns to accumulable factors by relating the process of factor accumulation to technical progress. For example, the accumulation of physical capital by individual firms may affect the productivity of the aggregate capital stock via external economies of scale or 'spillover' effects, as in Romer (1986). Alternatively, advances in the state of technology may be explicitly modelled in terms of a production function that describes the creation of new ideas, as in Romer (1990). The result is that in many NEG models, growth is ultimately endogenous in the sense of the second as well as the first definition of endogenous growth theory provided earlier.[18]

Not all neoclassical growth theorists are equally enthusiastic about these developments, however. Solow (1994, pp. 49–51), amongst others, argues that the assumptions about the structure of production necessary to sustain NEG theory leave it on a knife-edge: the marginal returns to accumulable factors must not only be non-diminishing (in order to avoid a regress into the territory of the Solow model), but also non-increasing (in order to avoid giving rise to unrealistically explosive growth outcomes). In other words, accumulable factors must ultimately display *exactly* constant marginal returns.[19] Moreover, by finding evidence of convergence, empirical studies such as Mankiw, Romer and Weil (1992) favour the assumptions of the Solow model (strictly *diminishing* returns to accumulable factors) over those of NEG theory.[20, 21] Within neoclassical macrodynamics, this scepticism has given rise to the emergence of what can be termed 'semi-endogenous' growth theory (see, for example, Jones 1995).[22] In models of this genus – and *unlike* the Solow model – the growth rate *does* depend on the (equilibrium) solution of the model and technical progress *is* explicitly modelled. But the parameters that ultimately determine the rate of growth are not obviously or easily amenable to change on the basis of the decisions of economic agents (including policy makers), whilst changes in the rates of factor accumulation (due to the choices of economic agents) do not affect the long-run growth rate.[23] In short – and very much *like* the Solow model – long-run growth is not sensitive to economic decision making that affects factor accumulation (such as changes in the saving rate), which instead ultimately affects only the *level* of real output per capita/worker. Semi-endogenous growth models can therefore be understood as an 'intermediate' class of models that are neither exactly like the Solow model (because they focus on modelling technical change and because the growth rate emerges from the solution of the model), nor exactly like NEG models (because factor accumulation does not drive growth and because the long-run growth rate is neither obviously or easily amenable to change as a result of economic decision making).

These neoclassical developments are all very well, but as noted earlier, endogenous growth theory by either of the definitions introduced at the beginning of this section clearly pre-dates the neoclassical ‘discovery’ of endogenous growth during the 1980s. Hence, as already noted, the concept of endogenous growth can be found in the work of, for example, Harrod and Kaldor, a fact that has resulted in these authors being identified as the progenitors of modern endogenous growth theory.[24] Why then, one might ask, is this fact not more widely recognized, and why is the term ‘endogenous growth’ frequently understood as a contemporary contrivance that applies exclusively to recent developments in neoclassical growth theory? Whilst it is beyond the scope of this survey to address these questions, it is pertinent to note that they bring us back to the second of the two key points raised earlier: the idea that any full and proper account of endogenous growth theory must acknowledge the distinction between supply- and demand-led growth, and the concomitant existence of *Keynesian* counterparts to the NEG theories described above.

4. SUPPLY-LED VERSUS DEMAND-LED GROWTH, AND THE FEATURES OF KEYNESIAN ENDOGENOUS GROWTH THEORY

Neoclassical growth theory (both first *and* second generations) focuses on modelling the expansion of *potential* output on the supply side of the economy, with aggregate demand assumed to passively adjust to accommodate potential output. In other words, it embodies Say’s Law: aggregate supply creates its own demand, and there is no prospect of ‘Keynesian problems’ arising from deficient aggregate demand at any point in time. As a result of this, the economy’s long-run potential growth path becomes its long-run *actual* growth path.[25] The essential vision of neoclassical growth theory may thus be characterized as follows:

$$y_p = f(X), f' > 0 \tag{1}$$

$$y = y_p \tag{2}$$

where y and y_p denote the rates of growth of actual and potential output per capita, respectively, and X is a vector of variables reflecting the growth of the quantity and/or productivity of factor inputs. Equation (1) constitutes little more than a dynamic aggregate production function, and as was made clear in the previous section, the variables X may be taken as exogenously given (as in the Solow model) or explained within the confines of the model itself (as in NEG theory). Equation (2) establishes the growth of potential output as the proximate determinant of the growth of actual output. Combining (1) and (2), then, we arrive at:

$$y = f(X) \tag{3}$$

which establishes that the ultimate determinant of the actual rate of growth is growth in the quantity and/or productivity of the factors of production.[26]

In contrast to the neoclassical, supply-led vision of growth, Keynesian demand-led growth theory focuses on modelling the expansion of *actual* output as a result of developments on the demand side of the economy. Aggregate supply is treated as adjusting, within limits, to accommodate demand-led changes in actual output – what Cornwall (1972) terms 'Say's Law in reverse' – through changes in capacity utilization and/or changes in the natural rate of growth resulting from induced factor accumulation, factor migration (both occupational and geographical) and technical progress.[27] The result is that potential output determined on the supply-side of the economy does not create an autonomous and necessarily binding constraint on the path of actual output in demand-led growth theory. Rather, long-run growth is essentially demand-determined in much the same way that Keynesians have always argued that the levels of real output and employment are essentially demand-determined in the short-run. The central vision of Keynesian growth theory may thus be characterized as follows:

$$y = g(Z), g' > 0 \tag{4}$$

$$y_p = y \tag{5}$$

where Z denotes a vector of variables determining the rate of growth of demand and all other variables are as previously defined.[28] In equation (4), the actual rate of growth of output is determined by factors affecting the rate of growth of aggregate demand. Note that demand-led growth models typically furnish explicit theories of the growth of aggregate demand, so that the rate of growth of actual output described by these models results from the solution of the models themselves. In this way, there is a close association between demand-led growth theory and the first definition of endogenous growth theory presented in Section 2.

The growth of aggregate demand is translated into the growth of actual output in the first instance through variations in the utilization rates of productive resources, the latter conceived as being typically under-utilized at any given point in time in the demand-constrained vision of Keynesian macroeconomics. But demand-led growth theory also allows for aggregate supply to respond to aggregate demand via the endogeneity of the natural rate of growth to the actual rate of growth. This is captured in equation (5), wherein the actual rate of growth is presented as the proximate determinant of the rate of growth of potential output. Combining equations (4) and (5), we arrive at:

$$y_p = g(Z) \tag{6}$$

according to which the ultimate determinants of the growth of potential output are factors affecting the growth of aggregate demand.[29] As intimated earlier, this relationship arises from the purported impact of changes in the demand-led actual rate of growth on a variety of phenomena traditionally associated with the supply-side of the economy, including factor accumulation, factor migration, and technical progress. Mention of the latter at this juncture draws to attention the close association between demand-led growth theory and the second definition of endogenous growth theory provided earlier.

The distinction between demand-led and supply-led growth only comes to light once we acknowledge the variety of growth theories that is otherwise obscured by the neoclassical 'capture' of both the history and terminology of growth theory. Moreover, this distinction, once uncovered, reveals an immediate affinity between demand-led growth theory and endogenous growth theory by *either* definition of the latter stated earlier. In other words, in addition to the NEG theories previously discussed, which are often *exclusively* associated with the notion of endogenous growth theory, we can also identify a family of *Keynesian* endogenous growth (KEG) models that satisfy either one or both of our definitions of endogenous growth theory. There are two main strands of KEG theory: Kaldorian and neo-Kaleckian. Neo-Kaleckian growth theory has its roots in Robinson's (1956) *Accumulation of Capital* and the two-sided relationship between investment and profit emphasized by Kalecki (1935), according to which investment creates profit (via an income-expenditure process) and profit creates investment (since profit expectations motivate investment, whilst realised profits finance investment). Rowthorn (1981) and Dutt (1984) are the progenitors of formal neo-Kaleckian models of growth; see Blecker (2002) for a survey of contemporary neo-Kaleckian growth theory. The focus of neo-Kaleckian growth theory is the relationship between growth and the functional distribution of income – specifically, the question as to how changes in the latter affect the former. Neo-Kaleckian models typically abstract from technical progress: these are endogenous growth models primarily in the sense of the first definition of endogenous growth introduced above.[30]

Kaldorian theory, meanwhile, is rooted in the various contributions to growth theory made by Nicholas Kaldor, and especially his post-1966 work on cumulative causation (see, for example, Kaldor, 1970, 1972, 1985, 1996). Dixon and Thirlwall (1975) are the progenitors of formal Kaldorian models of growth; see McCombie and Thirlwall (1994) for a survey of contemporary Kaldorian growth theory. Kaldorian theory focuses on the recursive interaction of economic growth and technical progress in an open economy context. Endogenous technical progress is understood to arise from a variety of sources, including the external economies of scale or 'spillover' effects that are also a feature of some NEG models.[31] In this way, Kaldorian models satisfy both of the definitions of endogenous growth theory stated earlier.

The distinction between Kaldorian and neo-Kaleckian growth theories outlined above is somewhat false. Both are of the same KEG genus, so that models from both strands can ultimately be regarded as *partial* representations of essentially the *same* underlying demand-led growth process. To see this more clearly, note that using only national income accounting identities and assuming a constant capital-output ratio, we can write:

$$q = \hat{w} + \left(\frac{1-\omega}{\omega}\right)\hat{r} \tag{7}$$

where q is the rate of growth of labour productivity as previously defined, $\hat{w}$ is the rate of growth of real wages, $\hat{r}$ is the growth of the rate of profit and ω is the wage share of income. Equation (7) makes the simple statement that productivity growth must break down into real wage growth and/or profit rate growth according to *some* distributional schema: the seemingly proprietorial themes of Kaldorian models (technical progress and hence productivity growth) and neo-Kaleckian models (the functional distribution of income) are thus intimately related.[32] It is significant in this regard that efforts are now being made by Kaldorians and neo-Kaleckians alike to provide generalized models of growth, distribution *and* technical change (see, for example, Bhaduri, 2006 and Naastepad, 2006).[33]

5. ENDOGENOUS GROWTH AS A HISTORICAL PROCESS

The definitions of endogenous growth theory introduced at the start of Section 2 serve a useful purpose in permitting a distinction between successive generations of neoclassical growth theory and, via a critique of the neoclassical 'capture' of the history and terminology of growth theory, facilitate the identification of a family of demand-led or KEG theories. But as useful as they are, the very notion of 'endogenous' growth consistent with *either* of these definitions can be criticized for embodying a conception of endogeneity that amounts to little more than 'pushing back the frontiers of exogeneity'. Specifically, according to these definitions, rendering growth 'endogenous' would seem to amount to little more than taking what was once treated as exogenous (technical progress, or even the growth rate itself) and describing it in terms of foreclosed explanation in terms of *other* exogenous givens – whether it be the supply-side determinants of factor accumulation and productivity (such as preferences for human capital formation or the discount rate) in NEG models, or the demand-side determinants of the expansion of aggregate demand (such as the animal spirits of investors or the elasticity of demand for exports) in KEG models.[34] For many economists, the process just described is the hallmark of progress in explaining phenomena in economic theory, and little or no further discussion is merited.[35] But other economists distinguish between determinate and indeterminate dynamical systems, where the former converge towards final (that is, fixed point) outcomes defined and reached independently of the path taken towards them (that is, their outcomes and behaviour are determined by exogenously given 'data'), whereas the latter do not, but are instead path dependent.[36] Indeed, indeterminate dynamical systems may be altogether bereft of any identifiably 'final' outcomes and generate only streams of continuously path-dependent output.[37] The distinction between determinate and indeterminate dynamical systems supports a different (and arguably 'deeper') conception of 'endogenous' growth as a *historical process*, wherein the growth rate today is influenced by the pace of growth in the past. In other words, growth is *endogenous to its own past history*, and 'the only truly exogenous factor is *whatever exists at a given moment of time*, as a heritage of the past' (Kaldor, 1985, p. 61: emphasis in original). On the basis of this conception of endogenous growth, we might identify a *third* definition of endogenous growth *theory* as the enterprise that seeks to analyse growth as a historically contingent process, eschewing traditional equilibrium analysis in favour of organizing concepts such as cumulative causation, lock-in, hysteresis and evolutionary change as the basis for growth modelling.[38] Examples of work in this genre of endogenous growth theory include Kaldor's (1972, 1985) verbal models of cumulative causation, together with formalisations of these models that do not involve determinate long-run equilibrium solutions (Setterfield, 1997, 2002b; Roberts, 2002, 2005b).[39] It seems reasonable to expect contributions consistent with this third definition of endogenous growth theory to multiply over time, as the volume of macrodynamic research that involves treating the economy as an evolving complex system increases.[40] Indeed, it is already possible to identify a variety of both Keynesian (demand-led) and neoclassical (supply-led) growth models that display chaotic dynamics. See, for example, Sportelli (2000) and McCombie and Roberts (2002) for chaotic Keynesian growth models, and Bohm and Kaas (2000) and Guo and Lansing (2002) for chaotic neoclassical growth models.

6. CONCLUSION

The object of this chapter has been to survey the field of endogenous growth theory with a view to drawing attention to the variety of different models of endogenous growth. Despite the neoclassical capture of the term ‘endogenous growth’ (and the history of growth theory more generally), it is important to realise that there exist both neoclassical and Keynesian endogenous growth theories, distinguished by their different characterisations of the growth process as being either fundamentally supply- or demand-led, respectively. Moreover, a ‘deeper’ conceptualization of endogenous growth as an inherently historical process permits identification of a special subset of endogenous growth theories that eschew traditional equilibrium analysis in favour of modelling growth as a path-dependent process. The main conclusion to be drawn from this chapter is that the essential variety of endogenous growth theories should have more bearing on discussions of growth, growth theory and the policy implications of theories of growth than it does at present.

NOTES

1. These search results were accurate as of July 2005.
2. These key drivers are: skills (human capital), investment (in physical capital), innovation, enterprise and competition. For more detail about the UK Treasury’s approach to raising both the national and regional growth rates, see HM Treasury (2000, 2001).
3. In the long run, the rates of growth of real output per capita and per worker can be expected to be equal. This is because $y \equiv q + e + p$, where y is the growth rate of real output per capita, q is the growth rate of real output per worker, e is the rate of growth of the employment rate and p is the rate of growth of the labour force participation rate. In the long run, both e and p will equal zero, which leaves us with $y = q$.
4. As will become evident in Section 5 below, it is also possible to entertain a *third* definition of endogenous growth theory that is substantively different from either of those offered below.
5. These include what will subsequently be identified as neoclassical endogenous growth models (for example, Aghion and Howitt, 1992, 1998; Grossman and Helpman, 1991; Romer, 1990) and Keynesian endogenous growth models (for example, Dixon and Thirlwall, 1975; Dutt, 2003; Palley, 1996b).

6. Once again, these include examples of both neoclassical endogenous growth theory (for example, Lucas, 1988; Rebelo, 1991) and Keynesian endogenous growth theory (Dutt, 1984; Rowthorn, 1981).
7. Models of this nature are sometimes referred to as ‘semi-endogenous’ growth models (see, for example, Jones, 2002) and are discussed in more detail below.
8. The term ‘modern’ growth theory is sometimes used to distinguish post-1930 developments in the field of economic growth from those associated with Classical economists, including Smith, Ricardo and Marx: see, for example, Jones, 1976, Chapter 1 for discussion of this usage. See Kurz and Salvadori (1998) for discussion of antecedents of endogenous growth theory in classical economics.
9. See Palley (1996a) for discussion of the formal development of Harrod’s growth dynamics.
10. See Dixon and Thirlwall (1975, p. 209) and McCombie and Thirlwall (1994, p. 464) for a demonstration of the links between Kaldor’s technical progress function and Verdoorn’s Law, of which Kaldor himself was a champion (Kaldor, 1966, 1970).
11. The Solow model is otherwise known as the Solow–Swan model in recognition of the virtually simultaneous discovery of the same model by Swan (1956).
12. See Barro and Sala-i-Martin (2004, pp. 16–21) for an exemplary account of this view. See also Setterfield (2002a, 2003a) for a critical assessment of the neoclassical ‘capture’ of the history of growth theory.
13. The relationship between the rate of growth of the labour force and technical progress on the one hand, and the natural rate of growth as defined in the text on the other, can be understood by examining the definition of the *level* of labour productivity: we can write $Q = Y/N$, where Q denotes labour productivity, Y is the level of real output and N is the level of employment. Re-arranging, we have $Y = Q.N$, on the basis of which it is straightforward to see that the actual level of real output can vary depending upon the level of labour productivity and the level of employment. But now suppose that Q is completely determined by the current state of technology, and that $N = L$, where L denotes the available labour force. Then we can write $Y_p = Q.L$, where Y_p denotes the potential level of real output – the maximum level of real output that can be produced given the current state of technology and the available labour force. (In practice, of course, Y_p may be lower than the value derived above, because it may not be possible to achieve the zero unemployment rate implicit in the assumption $N = L$. However, we overlook this complication – which does not, in fact, affect the result derived below – for purposes of simplicity.)

    Finally, taking logs and differentiating with respect to time on both sides of this expression, we arrive at $y_p = q + l$, where lower case variables denote the rates of growth of the upper case variables introduced above. This expression tells us that y_p – the growth of potential output or the ‘natural’ rate of growth – depends on technical progress (the determinant of q) and the rate of growth of the labour force, as claimed in the text. It should be noted that together with technical progress, social factors (including changes in the organization of production and/or the intensity of the labour process) may affect q, but this possibility is overlooked here for the sake of simplicity.
14. By equilibrium in this context we mean steady-state as opposed to market clearing. Markets continuously clear in the Solow model and, therefore, in this respect, the economy is in continuous equilibrium even when it is not in steady-state equilibrium.
15. In other words, the aggregate production function satisfies the so-called Inada conditions. See Barro and Sala-i-Martin (2004, Chapter 2).
16. To see this, consider the familiar equation for the accumulation of capital per worker in the Solow model, $\dot{k} = sQ - lk - \delta k$, where k is the capital-labour ratio, s is the saving rate, δ is the rate of depreciation of capital and Q and l are as previously defined. Assuming for simplicity that $l = \delta = 0$, we can rewrite the equation above as $(\dot{Q}/Q)\cdot(\dot{k}/\dot{Q}) = s$, which implies $q = s.MP_K$, where MP_K denotes the marginal physical product of capital. Clearly, if MP_K diminishes to zero as the stock of capital rises, then so, too, must q: growth cannot be sustained on the basis of capital accumulation alone.
17. Accumulable factors of production may encounter diminishing marginal returns in NEG theory, but the marginal physical product of accumulable factors must be bounded

from below by a positive constant to ensure that additions to the existing stock of these factors always have a positive impact on total output. In terms of the growth equation derived in the previous footnote, if $MP_K = c > 0$ even in the limit (where K is now defined broadly to include all accumulable factors of production), then the rate of growth $q = sc$ can be sustained on the basis of factor accumulation alone.

There are three basic types of NEG models: one-sector models in which capital accumulation has a direct effect on the output of individual firms, and an indirect effect by virtue of its enhancing the aggregate capital stock and hence the average product of labour (as in Romer, 1986); one-sector models with linear aggregate production functions (so-called 'AK' models) in which growth is directly linked to the process of capital accumulation (as in Rebelo, 1991); and two- or three-sector models in which the output of a knowledge-producing sector impacts the production function of the goods-producing sector of the economy (as in Lucas, 1988 and Romer, 1990). Despite this apparent heterogeneity, assumptions regarding the technical conditions of production that give rise to non-diminishing marginal returns to accumulable factors of production are common to all NEG models.

18. Indeed, at this remove, it is fair to say that endogenous technical progress is now a hallmark of NEG theory, the modelling of technical change through theories of R&D together with the introduction of imperfect competition (as a result of which firms are able to capture monopoly rents to their R&D activities) having exemplified the 'second wave' of NEG theories developed during the 1990s (see, for example, Romer, 1990; Grossman and Helpman, 1991; Aghion and Howitt, 1992, 1998). Indeed, some authors (for example, Solow, 1994, p. 51) regard the modelling of technical progress as *the* lasting contribution of NEG theory. See the ensuing discussion of 'semi-endogenous' growth theory below for further contextualization of this opinion.
19. Equivalently, the key differential equation that drives the dynamics of the model in question must be fundamentally linear. If the equation is less than linear (the diminishing marginal returns case) then growth becomes exogenous in the sense of being imposed upon the model from without, whilst, if it is greater than linear (the increasing marginal returns case) then growth becomes explosive.
20. The study by Mankiw, Romer and Weil (1992) is concerned with cross-country variations in income per capita levels and growth rates. However, their findings of convergence, which are taken as providing support for an augmented version of the Solow model, have been replicated for a wide range of regional samples. Particularly important in this respect is the work of Barro and Sala-i-Martin (1991, 1992, 2004, Chapter 11).
21. See, however, Felipe and McCombie (2005) for a skeptical view of the capacity of studies such as Mankiw, Romer and Weil (1992) to tell us anything useful about the structure of production, and hence the presumed nature of the growth process.
22. See also Jones (2002, Chapters 5 and 8) for an overview of semi-endogenous growth theory.
23. These models do not embody the assumption, common to NEG models, of constant marginal returns to accumulable factors. In other words, as with the Solow model, they are characterised by a key differential equation that is less than linear.
24. See, for example, Thirlwall (2000) on the notion that 'AK' models of neoclassical endogenous growth have simply rediscovered Harrod's constant capital–output ratio, Palley (1996b, pp. 123–5) on the notion that Kaldor's (1957) technical progress function anticipates contemporary neoclassical models of R&D (such as Romer, 1990), and Roberts and McCombie (2004) on the interpretation of Kaldor's (1976, 1996) two-sector agriculture–industry model as a model of endogenous growth.
25. Note that these characteristics of neoclassical growth theory are all implicit in the discussion in the previous section.
26. It should be noted that neoclassical growth theory – or more specifically, NEG theory – is certainly capable of creating a causal role for aggregate demand in the determination of long-run growth (see Palley, 2002a for general discussion of this point, and Blackburn, 1999 for an example of a specific NEG model in which aggregate demand conditions influence long-run growth). But the proclivity of neoclassical macroeconomics to claim

that supply creates its own demand means that, whilst variations in demand conditions *may* influence growth in *some* NEG models, they are far from being a central (much less a necessary) feature of NEG theory: growth can be, and usually is, analysed in entirely supply-side terms, as per the characterization in the text. So it is that Stern (1991, p. 123), in his survey of advances in neoclassical growth theory since the mid-1960s, opines that growth 'is about the accumulation of physical capital, the progress of skills, ideas and innovation, the growth of population, how factors are combined, managed and so on . . . [and] is therefore, principally, about the supply side'. Absent is any hint that demand may play an important role in either the development or the subsequent utilization of the productive forces he names. The absence of serious engagement with the demand side in neoclassical growth theory is also a prominent theme in Solow (2000, pp. 183–4); see also Roberts (2005a, pp. 12–14) for discussion of the supply-side orientation of both first and second generation neoclassical growth theory.

27. It is important to note that the adjustment of aggregate supply to aggregate demand is not always treated as being as passive and fully accommodating as the notion of 'Say's law in reverse' might seem to suggest. Thus, one of the concerns of demand-led growth theory is the precise elasticity of aggregate supply with respect to aggregate demand and, in particular, the conditions necessary to bring about a reconciliation of the rates of growth of actual and potential output in a steady-state equilibrium. See, for example, Cornwall (1972), Palley (2002b), Dutt (2006) and Setterfield (2006). There is, furthermore, a concern within demand-led growth theory about the emergence of possible constraints on the growth of aggregate demand, constraints that can arise on the supply-side of the economy (see, for example, McCombie and Thirlwall, 1994, Chapter 3).
28. Keynesian – and in particular, neo-Kaleckian – demand-led growth theories frequently emphasise not just the *rate of growth* of aggregate demand but also its *level* in the determination of the growth of actual output. More specifically, the current level of demand (in real terms) relative to the current level of real potential output – capturing the extent to which currently existing factors of production are utilized – is understood to be a determinant of the rate of growth (see, for example, Blecker 2002). For the sake of simplicity, this is overlooked in the stylised representation of demand-led growth theory provided above.
29. Again, the reader is reminded that the *level* of aggregate demand (relative to the potential real output of the economy) may also be important here.
30. Endogenous technical progress can be accommodated within the neo-Kaleckian framework, however. See, for example, Lavoie (1992, pp. 322–7) and Dutt (2003, pp. 87–93).
31. Kaldorian theory typically models technical progress by appeal to Verdoorn's Law, which relates the rate of growth of labour productivity to the rate of growth of output. Verdoorn's Law is something of a 'black box', with a taxonomy of factors including (but not limited to) spillover effects used to justify the relationship it embodies (see Toner, 1999).
32. It might be said that Kaldorian models focus on technical progress but neglect to describe how the fruits of this process are distributed, whilst neo-Kaleckian models focus on distribution but neglect to describe how the growing per capita surplus that is the source of distributional conflict is generated.
33. A contrast exists here with the development of neoclassical growth theory which, rather than seeking to reconcile the endogenous determination of growth, technical change and distribution, has evolved from a state in which exogenous technical change and growth facilitated the endogenous determination of distribution (as in the Solow model) to one in which theories of endogenous technical change and growth are facilitated by exogenous determination of distribution (NEG theory). See Kurz and Salvadori (1998) for extensive discussion of the notion that determination of distribution by means of extraneous technical parameters is a common feature of NEG models.
34. This theme has been raised before in earlier surveys of endogenous growth theory – see, for example, Setterfield (1993) and Fine (2000). Note that it renders moot the distinction between endogenous and semi-endogenous growth in neoclassical variants of endogenous growth theory. See also Kurz and Salvadori (1998, pp. 67–8) for a related, but qualitatively different critique, of the endogeneity of growth in NEG theory.

35. See, for example, Solow (2000, pp. 180–1).
36. The language used here is derived from Kaldor (1934).
37. In Kaldor's (1934) original language, these constitute the class of indefinite-indeterminate systems, whereas indeterminate systems that do eventually generate 'final' outcomes are definite-indeterminate.
38. Setterfield (1997, Chapter 1) identifies 'traditional equilibrium analysis' as a methodology associated with the application of determinate dynamical systems, as defined in the text.
39. It should be noted that it was this same basic vision of macrodynamics as a historical process that inspired Robinson (1956). See Setterfield (2003b) for an attempt to develop a neo-Kaleckian model consistent with this vision.
40. A complex system as envisaged here is one that, following Day (1994), does not converge onto a fixed point attractor or limit cycle, or endogenously generate continuously explosive output growth. Note the affinity between this definition of a complex system and the notion of an indefinite-indeterminate dynamical system as defined in note 37.

REFERENCES

Aghion, P. and P. Howitt (1992), 'A model of growth through creative destruction', *Econometrica*, **60**, 323–51.

Aghion, P. and P. Howitt (1998), *Endogenous Growth Theory*, Cambridge, MA: MIT Press.

Barro, R.J. and X. Sala-i-Martin (1991), 'Convergence across states and regions', *Brookings Papers on Economic Activity*, **1**, 107–82.

Barro, R.J. and X. Sala-i-Martin (1992), 'Convergence', *Journal of Political Economy*, **100**, 223–51.

Barro, R.J. and X. Sala-i-Martin (2004), *Economic Growth*, Second Edition, Cambridge, MA: MIT Press.

Bhaduri, A. (2006), 'Endogenous economic growth: a new approach', *Cambridge Journal of Economics*, **30**, 69–83.

Blackburn, K. (1999), 'Can stabilisation policy reduce long-run growth?', *Economic Journal*, **109**, 67–77.

Blecker, R. (2002), 'Distribution, demand and growth in neo-Kaleckian macromodels', in M. Setterfield (ed.), *The Economics of Demand-Led Growth: Challenging the Supply-side Vision of the Long Run*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Bohm, V. and L. Kaas (2000), 'Differential savings, factor shares, and endogenous growth cycles', *Journal of Economic Dynamics and Control*, **24**, 965–80.

Cornwall, J. (1972), *Growth and Stability in a Mature Economy*, London: Martin Robertson.

Day, R.H. (1994), *Complex Economic Dynamics, Volume 1: An Introduction to Dynamical Systems and Market Mechanisms*, Cambridge, MA: MIT Press.

Dixon, R. and A.P. Thirlwall (1975), 'A model of regional growth along Kaldorian lines', *Oxford Economic Papers*, **27**, 201–14.

Dutt, A.K. (1984), 'Stagnation, income distribution and monopoly power', *Cambridge Journal of Economics*, **8**, 25–40.

Dutt, A.K. (2003), 'New growth theory, effective demand and post-Keynesian dynamics', in N. Salvadori (ed.), *Old and New Growth Theories: An Assessment*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Dutt, A.K. (2006), 'Aggregate demand, aggregate supply and economic growth', *International Review of Applied Economics*, **20**, 319–36.

Felipe, J. and J.S.L. McCombie (2005), 'Why are some countries richer than others? A skeptical view of Mankiw–Romer–Weil's test of the neoclassical growth model', *Metroeconomica*, **56**, 360–92.

Fine, B. (2000), 'Endogenous growth theory: a critical assessment', *Cambridge Journal of Economics*, **24**, 245–65.

Grossman, G.M. and E. Helpman (1991), *Innovation and Growth in the Global Economy*, Cambridge, MA: MIT Press.

Guo, J-T., and K.J. Lansing (2002), 'Fiscal policy, increasing returns and endogenous fluctuations', *Macroeconomic Dynamics*, **6**, 633–64.

HM Treasury (2000), 'Productivity in the UK: the evidence and the Government's approach', http://www.hm-treasury.gov.uk/documents/enterprise_and_productivity/the_evidence/ent_prodevi_index.cfm, accessed 26 June 2006.

HM Treasury (2001), 'Productivity in the UK: 3- the regional dimension', http://www.hm-treasury.gov.uk/documents/enterprise_and_productivity/the_evidence/ent_prod3_index.cfm.

Jones, C.I. (1995), 'R&D based models of economic growth', *Journal of Political Economy*, **103**, 759–84.

Jones, C.I. (2002), *Introduction to Economic Growth*, Second Edition, New York: W.W. Norton.

Jones, H.G. (1976), *An Introduction to Modern Theories of Economic Growth*, New York: McGraw Hill.

Kaldor, N. (1934), 'A classificatory note on the determinateness of equilibrium', *Review of Economic Studies*, **2**, 122–36.

Kaldor, N. (1957), 'A model of economic growth', *Economic Journal*, **67**, 591–624.

Kaldor, N. (1966), *Causes of the Slow Rate of Economic Growth of the United Kingdom*, Cambridge: Cambridge University Press.

Kaldor, N. (1970), 'The case for regional policies', *Scottish Journal of Political Economy*, **18**, 337–48.

Kaldor, N. (1972), 'The irrelevance of equilibrium economics', *Economic Journal*, **82**, 1237–55.

Kaldor, N. (1976), 'Inflation and recession in the world economy', *Economic Journal*, **86**, 703–14.

Kaldor, N. (1985), *Economics Without Equilibrium*, Cardiff: University College of Cardiff Press.

Kaldor, N. (1996), *Causes of Growth and Stagnation in the World Economy*, Cambridge: Cambridge University Press.

Kalecki, M. (1935), 'A macroeconomic theory of the business cycle', *Econometrica*, **3**, 327–44.

Kurz, H.D and N. Salvadori (1998), 'The "new" growth theory: old wine in new goatskins', in F. Coricelli, M. di Matteo and F. Hahn (eds), *New Theories in Growth and Development*, London: Macmillan.

Lavoie, M. (1992), *Foundations of Post-Keynesian Economic Analysis*, Aldershot, UK and Brookfield, USA: Edward Elgar.

Lucas, R.E. (1988), 'On the mechanics of economic development', *Journal of Monetary Economics*, **22**, 3–42.

Mankiw, N.G., D. Romer and D. Weil (1992), 'A contribution to the empirics of economic growth', *Quarterly Journal of Economics*, **107**, 407–38.

McCombie, J.S.L. and M. Roberts (2002), 'The role of the balance of payments in economic growth', in M. Setterfield (ed.), *The Economics of Demand-Led Growth: Challenging the Supply-side Vision of the Long Run*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

McCombie, J.S.L. and A.P. Thirlwall (1994), *Economic Growth and the Balance of Payments Constraint*, London: Macmillan.

Naastepad, C.W.M. (2006), 'Technology, demand and distribution: a cumulative growth model with an application to the Dutch productivity slowdown', *Cambridge Journal of Economics*, **30**, 403–34.

Palley, T.I. (1996a), 'Aggregate demand in a reconstruction of growth theory: the macro foundations of economic growth', *Review of Political Economy*, **8**, 23–35.

Palley, T.I. (1996b), 'Growth theory in a Keynesian mode: some Keynesian foundations for new endogenous growth theory', *Journal of Post Keynesian Economics*, **19**, 113–36.

Palley, T.I. (2002a), 'Keynesian macroeconomics and the theory of economic growth: putting aggregate demand back in the picture', in M. Setterfield (ed.), *The Economics of Demand-Led Growth: Challenging the Supply-side Vision of the Long Run*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Palley, T.I. (2002b), 'Pitfalls in the theory of growth: an application to the balance-of-payments-constrained growth model', in M. Setterfield (ed.), *The Economics of Demand-Led Growth: Challenging the Supply-side Vision of the Long Run*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Rebelo, S. (1991), 'Long-run policy analysis and long-run growth', *Journal of Political Economy*, **96**, 500–21.

Roberts, M. (2002), *Essays in Cumulative Causation*, unpublished PhD thesis, Department of Land Economy, University of Cambridge.

Roberts, M. (2005a), 'Convergence and the Kaldorian approach to growth', mimeo, Cambridge Centre for Economic and Public Policy, University of Cambridge.

Roberts, M. (2005b), '"History matters": multiple equilibria versus intentional human agency', in M. Setterfield (ed.), *Interactions in Analytical Political Economy: Theory, Policy and Applications*, Armonk, NY: M.E. Sharpe.

Roberts, M. and J.S.L. McCombie (2004), 'Effective demand constrained growth in a two-sector Kaldorian model', mimeo, Cambridge Centre for Economic and Public Policy, University of Cambridge.

Robinson, J. (1956), *The Accumulation of Capital*, London: Macmillan.

Romer, P.M. (1986), 'Increasing returns and long-run growth', *Journal of Political Economy*, **94**, 1002–37.

Romer, P.M. (1990), 'Endogenous technical change', *Journal of Political Economy*, **98**, S71–S102.

Rowthorn, R.E. (1981), 'Demand, real wages and economic growth', *Thames Papers in Political Economy*, London: Thames Polytechnic.

Setterfield, M. (1993), 'Change or permanence? Growth and development in capitalist economies', *Review of Income and Wealth*, **39**, 217–23.

Setterfield, M. (1997), *Rapid Growth and Relative Decline: Modelling Macroeconomic Dynamics with Hysteresis*, London: Macmillan.

Setterfield, M. (2002a), 'Introduction: a dissenter's view of the development of growth theory and the importance of demand-led growth', in M. Setterfield (ed.), *The Economics of Demand-Led Growth: Challenging the Supply-side Vision of the Long Run*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Setterfield, M. (2002b), 'A model of Kaldorian traverse: cumulative causation, structural change and evolutionary hysteresis', in M. Setterfield (ed.), *The Economics of Demand-Led Growth: Challenging the Supply-side Vision of the Long Run*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Setterfield, M. (2003a), 'Supply *and* demand in the theory of long-run growth: introduction to a symposium on demand-led growth', *Review of Political Economy*, **15**, 23–32.

Setterfield, M. (2003b), 'Neo-Kaleckian growth dynamics and the state of long run expectations: wage- versus profit-led growth reconsidered', in N. Salvadori (ed.), *Old and New Growth Theories: An Assessment*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Setterfield, M. (2006), 'Thirlwall's Law and Palley's pitfalls: a reconsideration', in P. Arestis, J.S.L. McCombie and R. Vickerman (eds), *Growth and Economic Development: Essays in Honour of A.P. Thirlwall*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Solow, R.M. (1956), 'A contribution to the theory of economic growth', *Quarterly Journal of Economics*, **70**, 65–94.

Solow, R.M. (1994), 'Perspectives on growth theory', *Journal of Economic Perspectives*, **8**, 45–54.

Solow, R.M. (2000), *Growth Theory: An Exposition*, Second Edition, Oxford: Oxford University Press.

Sportelli, M.C. (2000), 'Dynamic complexity in a Keynesian growth-cycle model involving Harrod's instability', *Journal of Economics (Zeitschrift für Nationalökonomie)*, **71**, 167–98.

Stern, N. (1991), 'The determinants of growth', *Economic Journal*, **101**, 122–33.

Swan, T.W. (1956), 'Economic growth and capital accumulation', *Economic Record*, **32**, 334–61.

Thirlwall, A.P. (2000), 'The AK model of "new" growth theory is the Harrod–Domar growth equation: investment and growth revisited', *Journal of Post Keynesian Economics*, **22**, 427–35.

Toner, P. (1999), *Main Currents in Cumulative Causation: The Dynamics of Growth and Development*, London: Macmillan.

4. Is the natural rate of growth exogenous?*

Miguel Leon-Ledesma and A.P. Thirlwall

INTRODUCTION

It was Sir Roy Harrod who first formally introduced the concept of the natural rate of growth into economic theory in his famous paper 'An Essay in Dynamic Theory' (Harrod, 1939). The paper was essentially a dynamisation of Keynes's *General Theory* and asked the question: if the condition for a static equilibrium is that plans to invest equal plans to save, what must be the growth of income in a growing economy for plans to invest to equal plans to save to give a moving equilibrium through time? Moreover, is there any guarantee that this required rate of growth (which Harrod called the warranted growth rate) will prevail, and, if not, what will happen? The answer was that there is no guarantee, and if the two growth rates diverge there will be dynamic instability. If the actual growth rate exceeds the warranted rate there will be overcapacity utilisation and producers will feel they have done too little investment for steady growth. They will invest more, pushing the actual growth rate further above the warranted rate. Contrawise, if the actual growth rate is below the warranted rate there will be excess capacity. Producers will revise their investment plans downwards, pushing the actual growth rate further below the warranted rate.

Within this framework, Harrod's natural rate of growth fulfilled two functions. First, it set a ceiling to explosive growth, turning cyclical booms into slumps. Secondly, it was supposed to give a measure of the long-run growth rate to which economies will gravitate; what Harrod called the 'social optimum' growth rate made up of the growth of the labour force and the growth of labour productivity (or the growth of the labour force in efficiency units). The rate of growth of the labour force and labour productivity, however, were treated as *exogenously* given, as if immutable to actual conditions prevailing in the economy – boom or slump.

* This paper was first published in *Banca Nazionale del Lavoro Quarterly Review*, December 2000, and is republished here in revised form with the permission of the Editor.

Ever since, virtually the whole of mainstream growth theory has treated the natural rate of growth as exogenous, independent of the actual rate of growth. It was treated as exogenous in the neoclassical response to Harrod, as, for example, in the original work of Solow (1956) and Swan (1956), which is still highly influential in the way economists view the growth process. It was treated as exogenous (by and large) in the original Keynesian response to the neoclassicals, represented by the work of Kaldor (1957) and Joan Robinson (1956). Paradoxically, it is even treated as exogenous in 'new' *endogenous* growth theory. Growth, in this apparently 'new' theory, is only endogenous in the sense that investment matters for growth, because the neoclassical assumption of diminishing returns to capital is relaxed, not in the sense that the determinants of the long-run (natural) growth rate – labour force growth and productivity growth – respond to the growth rate itself. New, endogenous, growth theory does not model the demand side of the economy. Indeed, the main aim of many of the economists who do empirical work in the field, such as Barro (1991), seems to be to rehabilitate neoclassical growth theory by saying that the neoclassical growth model would hold, and particularly its prediction of convergence of per capita incomes across regions and countries, *if only* levels of education, research and development expenditure, and other factors that affect the productivity of capital, were the same across countries. In other words, the basic aggregate production function exhibits diminishing returns to capital, but the marginal product of capital does not fall as countries get richer, and the capital–labour ratio rises, because of various externalities.

It is also the case that the assumption of *exogeneity* of factor supplies and productivity growth permeates the whole of the mainstream growth accounting literature on the sources of growth that uses the aggregate production function, such as the pioneer studies of Abramovitz (1956), Solow (1957), Maddison (1970) and Denison (1967), and the more recent work of Alwyn Young (1995) on South East Asia and Hu and Khan (1997) on China. Young claims that there has been no growth miracle in the East Asian 'tiger' economies of Hong Kong, Taiwan, Singapore and South Korea, because most of the rapid growth can be explained by the growth of factor inputs and not by technical progress or total factor productivity growth. But the question is never asked, let alone answered, of *why* the growth of factor inputs – labour and capital accumulation – was so fast? The use of factor inputs is not in general exogenously determined. Rather, the growth of factor inputs is endogenous to demand, and in the case of the Asian 'tigers' the driving force was undoubtedly the growth of export demand. There is no way that these countries could have grown at the rate that they did without the rapid growth of exports to pay for the import requirements for growth. In this sense, their growth was miraculous.

Suppose the natural rate of growth is not exogenously given. Suppose it is *endogenous* to demand, or the actual rate of growth, as we are suggesting above. What implications does this have? It has two major implications. First, at the theoretical level, it has implications for the efficiency and speed of the adjustment process between the warranted and natural growth rates in the Harrod growth model. Second, it has implications for the way the growth process should be viewed, and why growth rates differ between countries: whether growth is viewed as *supply determined*, or whether growth is viewed as *demand determined* or determined by *constraints on demand* before supply constraints bite. The view we take here is that it is a mistake to think of a natural rate of growth exogenously determined. In other words, there is nothing 'natural' about the natural rate of growth (just as there is nothing 'natural' about the natural rate of unemployment)! Both the growth of the labour force and labour productivity growth are positively related to demand or the actual rate of growth. Later, we will test this hypothesis and show this is empirically the case for a sample of 15 OECD countries over the period 1961 to 1995.

First, however, we will formally define the natural rate of growth and discuss the theoretical consequences of the natural rate being endogenous. Secondly, we will give reasons why the natural rate is likely to be endogenous. Thirdly, we will suggest a simple technique for estimating empirically the natural rate of growth and testing for its endogeneity. Finally, we will present results for 15 OECD countries of the elasticity of the natural rate during periods of boom when the actual rate has exceeded the natural rate.

THE NATURAL RATE OF GROWTH AND THEORETICAL CONSEQUENCES IF IT IS NOT EXOGENOUS

Although it was Harrod in 1939 who first coined the term 'the natural rate of growth', as a matter of historical interest, Keynes had effectively anticipated Harrod's ideas two years earlier in his Galton Lecture to the Eugenics Society in 1937 on 'Some Economic Consequences of a Declining Population' (Keynes, 1937), where he expressed the worry that because of a falling population there would not be enough demand to absorb full employment saving. Consider, he says, an economy with a savings ratio of 8–15 per cent of national income, and a capital–output ratio of 4, giving a rate of capital accumulation which will absorb saving of approximately 2–4 per cent. With a constant capital–output ratio, this is the required growth rate, but can this growth rate be guaranteed? Historically, it appeared to Keynes that one-half of the increase in capital accumulation could be attributed to increased population; the other half to increased living standards (productivity growth). Now suppose population growth falls to zero. Since the standard of life cannot be expected to grow by more than one per cent per annum, this means that the demand for capital will only grow at one per cent while the supply grows at between 2–4 per cent – a clear and worrying imbalance, which would have to be rectified either by reducing saving or reducing the rate of interest to lengthen the average period of production (that is, to raise the capital–output ratio). This discussion is exactly analogous to Harrod's discussion of divergences between the warranted and natural rates of growth. The required rate of growth to absorb saving is the warranted rate of growth and the long-run growth rate determined by population (labour force) growth and rising living standards (productivity growth through technical progress) is the natural rate of growth. Harrod's dynamic theory is precisely anticipated by Keynes, and Keynes, like Harrod, treats the natural growth rate as exogenous.

Given the definition of the natural rate of growth as the sum of the rate of growth of the labour force and the rate of growth of labour productivity, it follows that the measured natural rate must be that rate of growth that keeps the unemployment rate constant. Otherwise, if the actual growth rate is above the natural rate, unemployment will fall; and if the actual growth is below the natural rate, unemployment will rise. Throughout the rest of this chapter we define and measure the natural growth rate of countries as the rate which keeps unemployment constant.

As all students of economic growth will know already, there was no mechanism in the original Harrod model for bringing the warranted and natural rates of growth into line with one another, with the implication that economies might experience perpetual secular stagnation (if the warranted rate exceeded the natural rate) or permanent inflation and structural unemployment (if the natural rate exceeded the warranted rate, as in most developing countries). But mechanisms that achieve equilibrium were soon invented. The Cambridge, Massachusetts school, represented by Robert Solow, Paul Samuelson and Franco Modigliani used the neoclassical production function and variations in the capital–output ratio to show that the warranted growth rate would adjust to the natural rate (assuming appropriate factor price adjustment and a spectrum of production techniques to choose from). The Cambridge, England school, represented by Nicholas Kaldor, Joan Robinson, Richard Kahn and Luigi Pasinetti used variations in the savings ratio, brought about by changes in the functional distribution of income between wages and profits, as the mechanism to bring about equilibrium. But both schools, which hotly debated this issue for over twenty years, have equilibrium growth proceeding at the *exogenously* given natural rate.

What happens, however, if the natural rate of growth is not exogenous? This has interesting consequences both for the short-run trade cycle model of Harrod, as well as the long-run equilibrium growth model. Remember that in the trade cycle model, if the actual growth rate diverges from the warranted growth rate in either direction, forces come into play which widen the divergence – but divergence is bounded by ceilings and floors. The ceiling is the natural rate of growth, because the level of output cannot exceed the full employment ceiling. But suppose that the natural rate increases with the actual rate of growth (because labour force growth and productivity growth are induced); this will perpetuate the cyclical upturn. We conjecture that this increases the possibility that the cyclical upturn is not brought to an end by an absolute ceiling, but by demand constraints associated with inflation, and balance of payments problems due to bottlenecks in the system. This may explain why cyclical peaks are often accompanied by excess capacity. In any case, the endogeneity of the natural rate will surely lengthen the cycle.

In the long-period model of divergence between the warranted and natural growth rate, the endogeneity of the natural rate will impede adjustment to equilibrium. If the warranted rate exceeds the natural rate, it means that the growth of capital exceeds the growth of the labour force in efficiency units, and the warranted rate must fall for equilibrium. In conditions of recession, however, the natural rate is also likely to fall as workers leave the labour force and productivity growth slows, impeding adjustment. Similarly, if the natural rate exceeds the warranted rate, this implies that the growth of the effective labour force exceeds the growth of capital, and the warranted rate must rise for equilibrium. In booms, however, the natural rate is also likely to rise as workers are attracted into the labour force and productivity growth accelerates, also impeding adjustment.

In general, the endogeneity of the natural rate of growth has serious implications for the notion of a *given* full employment production frontier which economies will gravitate towards. In practice, this frontier will continually shift with the actual growth rate.

IN WHAT WAYS IS THE NATURAL RATE ENDOGENOUS?

There are many mechanisms through which the natural rate of growth is likely to be endogenous to the actual rate of growth. Consider first the growth of the labour force or labour supply. Labour supply is extremely elastic to demand. When the demand for labour is strong, labour input responds in a number of ways. Firstly, participation rates rise. Workers previously out of the labour force decide to join the labour force. The participation rates of the young, the old and married women are particularly flexible and vary with the trade cycle. Secondly, hours worked increase. Part-time workers become full-time workers, and overtime work increases. Thirdly, and significantly for many countries across the world, labour migration takes place in response to booming labour markets. If countries are short of labour, they import it. Cornwall (1977) and Kindleberger (1967) document the important role that immigrant labour played in Europe during the 'golden age' of economic growth between 1950 and 1973. The migration of labour from Portugal, Spain, Greece and Turkey into Germany, France, Switzerland and northern Italy was not an exogenous movement, but fuelled by an excess demand for labour in the receiving countries because the growth of demand for output was so high. Similar stories could be told for other parts of the world.

Now consider the growth of labour productivity. There are several mechanisms through which labour productivity growth is endogenous to demand, and they are well documented. First, there are static and dynamic returns to scale associated with increases in the volume of output, and the technical progress incorporated in capital accumulation. With a constant ratio of capital to output, all technical progress is labour-augmenting. Some technical progress is autonomous, but a great deal is demand-driven, particularly process innovation. Secondly, there are macro increasing returns in the Allyn Young (1928) sense associated with the interrelated expansion of all activities. Suppose the market for a good expands, which makes it profitable to use more sophisticated machinery, which cuts costs. This not only reduces the price of the good (leading to further expansion of demand), but will also reduce the price of machinery if there are scale economies in its production, which makes it profitable to use machinery in other activities. The initial demand expansion leads to a series of changes, which propagate themselves in a cumulative way causing labour productivity to rise. Thirdly, there is the well-known phenomenon of learning by doing, whereby the efficiency or productivity of labour is an increasing function of a learning process related to cumulative output. The more widgets produced, the more adept labour becomes at producing them. Clearly the impact of learning will gradually diminish with cumulative output, but as long as product ranges change over time, the effect of learning on productivity growth will be a continuous process related to the expansion of output.

All the phenomena mentioned above are captured by the Verdoorn relation, or Verdoorn's Law, which posits a positive relation between the growth of output as the independent variable and the growth of labour productivity as the dependent variable (Verdoorn, 1949). In recent years, this relation has been tested extensively across countries (Kaldor, 1966; Michl, 1985), across regions within countries for both developed and less developed countries (McCombie and de Ridder, 1983; Fingleton and McCombie, 1998; Leon-Ledesma, 2000; Hansen and Zhang, 1996), and across industries (McCombie, 1985), and all find the relationship robust, with a central estimate of the Verdoorn coefficient of approximately 0.5. That is, an expansion of output demand by one per cent leads to a half-per cent increase in employment and a half-per cent increase in labour productivity induced by scale economies, embodied technical progress and learning by doing. It is no accident, therefore, that when growth slows down, productivity growth also slows down. The productivity growth slowdown after the shocks to the world economy in the 1970s was regarded as a puzzle by some economists, but can be readily understood in the context of models in which productivity growth is endogenous.

ESTIMATING THE NATURAL RATE OF GROWTH AND TESTING ITS ENDOGENEITY

Let us now turn to the question of how the natural growth rate of a country may be estimated, and test whether it is endogenous. The technique for estimation relies on a modification of the equations used for testing Okun's Law (Okun, 1962): a technique first suggested and applied by one of the present authors (Thirlwall, 1969). We saw earlier that by definition, the natural growth rate must be the growth rate that keeps unemployment constant. If we therefore relate changes in unemployment in a country to its growth rate, we can solve for the growth of output that keeps unemployment constant. In others words, let

$$\Delta\%u = a - b(g) \qquad (1)$$

where $\%u$ is the percentage rate of unemployment and g is the growth rate. Solving for g when $\Delta\%u = 0$ gives an expression for the natural rate of growth of $g_n = a/b$. The technique is simple, but there are certain problems. The estimate of the coefficient b may be biased downwards because of labour hoarding, which would exaggerate g_n. Equally, however, the constant term a may be biased downwards through workers leaving the labour force when g is low. It is difficult to know *a priori* what the relative strengths of the (offsetting) biases are likely to be.

An alternative procedure is to reverse the variables in equation (1) to give:

$$g = a_1 - b_1(\Delta\%u) \qquad (2)$$

Solving for g when $\Delta\%u = 0$ now gives an estimate for the natural rate of growth of $g_n = a_1$. This has statistical problems, since the change in unemployment is an endogenous variable, although it transpires empirically that this does not affect the results obtained from fitting equation (2).[1]

If this simple technique for estimating the natural rate of growth is accepted, the obvious way to test for endogeneity is to include a dummy variable into (say) equation (2) in periods when the actual growth rate is above the estimated natural rate and test for its significance, that is,

$$g = a_2 + b_2 D - c_2(\Delta\%u) \tag{3}$$

where D takes the value of 1 when actual growth is greater than the natural rate of growth and zero otherwise. If the dummy is significant, this must mean that the rate of growth in periods of boom to keep unemployment constant has risen. The actual growth rate must have been pulling more workers into the labour force and inducing productivity growth. The constant a_2 plus b_2 gives the natural rate of growth in boom periods. The interesting question is then how this estimate of the natural rate in boom periods compares with the estimate of the natural rate which does not distinguish between boom and slump. What is the elasticity of the natural rate in periods of boom?

EMPIRICAL RESULTS[2]

To test the model we take a sample of 15 OECD countries over the period 1961 to 1995. Both equations (1) and (2) were fitted to estimate the natural rate of growth over the whole period. In general, equation (2) gave the best results in terms of goodness of fit of the equations and the reasonableness of the results. In equation (2), the estimate of the natural rate of growth is given by the constant term (a_1), and this is reported for all countries in the first column of Table 4.1. The constant term was estimated as statistically significant in all 15 countries.[3] The estimates of the natural rate of growth all look reasonable for the countries concerned, and range from 2.5 per cent in the UK (the lowest) to 4.6 per cent in Japan (the highest). The average natural growth rate for the 15 OECD countries as a whole is 3.5 per cent.

When a dummy variable was added to equation (2) for years when the actual growth rate exceeded the estimated natural rate (equation 3) it was found to be significant for all 15 countries. The sum of the dummy plus the new constant (a_2) gives the natural rate of growth in boom periods, and is shown in column 2 of Table 4.1.

Table 4.1 *Estimates of the natural rate of growth and its endogeneity for 15 OECD countries, 1961–1995*

Country	(1) Natural Rate from Equation (2)(%)	(2) Natural Rate in Boom Periods (%)	(3) Increase in Natural Rate in Boom Periods, (2)–(1)	Percentage Increase in Natural Rate in Boom Periods, (3)/(1) ×100
Australia	3.999	5.713	1.714	42.9
Austria	3.136	4.956	1.820	58.1
Belgium	3.524	4.910	1.386	39.3
Canada	3.835	5.261	1.426	37.2
Denmark	2.942	4.782	1.840	62.5
France	2.827	3.934	1.107	39.2
Germany	3.505	4.709	1.204	34.3
Greece	4.509	7.671	3.162	70.1
Italy	3.344	5.910	2.566	76.8
Japan	4.567	8.720	4.153	90.9
Netherlands	3.282	5.315	2.033	62.0
Norway	3.972	5.009	1.037	26.1
Spain	4.062	6.093	2.031	50.0
UK	2.544	3.802	1.258	49.5
USA	2.991	3.664	0.673	22.5
Average	3.536	5.363	1.827	51.7

Source: Leon-Ledesma and Thirlwall (2002).

The natural rate is seen to increase considerably in all countries, but in some countries by more than others. Taking the countries as a whole, the average increase is 1.8 percentage points, which is to say that the actual rate of growth in boom periods has induced labour force growth and productivity growth by that amount. The countries where the sensitivity of the natural rate seems to be greatest are those where the reserves of labour are known to be highest, such as Greece and Italy (due to surplus labour in the south), and where output growth has induced impressive technical progress through learning and sectoral rationalisation, such as Japan. In general, the results show substantial elasticity of the labour force and productivity growth; certainly enough to suggest that the natural rate of growth is not exogenously given, but is very responsive to demand conditions in the economy. It is important to stress that these results are not measuring simply the *cyclical* effect of demand on output growth because this is captured by the coefficient c_2 in equation (3). The results are capturing the longer lasting effects that sustained demand expansion has had on the growth of productive potential over the period under study.

CONCLUSION

If supply or output potential responds to demand, this raises the crucial question of what it means to say that output growth is supply determined, or constrained by supply, which is the prevailing orthodoxy. Of course, it is true in a trivial sense that capital and labour are required to produce output, and how much output is produced will also depend on the level of technical efficiency, but the really important question is: why does the growth of capital, labour and technical progress differ so much between countries? The supply orientated, neoclassical production function approach to the analysis of growth cannot answer this question, and for the most part never asks it!

In our view, demand should assume a central role in growth theory and must play a major part in the explanation of growth rate differences between countries. For most countries, and particularly developing countries, demand constraints bite long before capacity is reached, and as we have shown, supply capacity is elastic. In an open economy, the major long-run constraint on demand is likely to be its balance of payments, but this is another story – originally outlined by one of the present authors elsewhere (Thirlwall, 1979). There is now substantial empirical support for this view, and interested readers are referred to McCombie and Thirlwall (1994, 1997, 2004) and a Symposium in the *Journal of Post Keynesian Economics* (1997).

NOTES

1. In Thirlwall (1969) both equations (1) and (2) were fitted to US and UK data over the period 1950 to 1967, and both procedures gave the same estimates of g_n: 2.9 per cent per annum for the UK and 3.3 per cent for the US – which seemed eminently reasonable.
2. The full regression results from which the various estimates in this section (and Table 4.1) are derived are available on request and can also be found in Leon-Ledesma and Thirlwall (2002).
3. Allowing for the endogeneity of $\Delta\%u$ using instrumental variables does not alter the results.

REFERENCES

Abramovitz, M. (1956), 'Resource and Output Trends in the United States since 1870', *American Economic Review* (Papers and Proceedings), May.

Barro, R. (1991), 'Economic Growth in a Cross Section of Countries', *Quarterly Journal of Economics*, May

Cornwall, J. (1977), *Modern Capitalism: Its Growth and Transformation*, London: Martin Robertson.

Denison, E. (1967), *Why Growth Rates Differ: Post-War Experience in Nine Western Countries*, Washington, DC: Brookings Institute.

Fingleton, B. and J. McCombie (1998), 'Increasing Returns and Economic Growth: Some Evidence from the European Regions', *Oxford Economic Papers*, January.

Hansen, J. and J. Zhang (1996), 'A Kaldorian Approach to Regional Economic Growth in China', *Applied Economics*, June.

Harrod, R. (1939), 'An Essay in Dynamic Theory', *Economic Journal*, March.

Hu, Z. and M. Khan (1997), 'Why is China Growing so Fast?', *IMF Staff Papers*, March.

Journal of Post Keynesian Economics (1997), *Mini Symposium on Thirlwall's Law and Economic Growth in an Open-Economy Context*, Spring.

Kaldor, N. (1957), 'A Model of Economic Growth', *Economic Journal*, December.

Kaldor, N. (1966), *Causes of the Slow Rate of Economic Growth of the United Kingdom*, Cambridge: Cambridge University Press.

Keynes, J.M. (1937), 'Some Economic Consequences of a Declining Population', *Eugenics Review*, April.

Kindleberger, C. (1967), *Europe's Postwar Growth: The Role of Labour Supply*, Cambridge, MA: Harvard University Press.

Leon-Ledesma, M. (2000), 'Economic Growth and Verdoorn's Law in the Spanish Regions 1962–1991', *International Review of Applied Economics*, January.

Leon-Ledesma, M. and A.P. Thirlwall (2002), 'The Endogeneity of the Natural Rate of Growth', *Cambridge Journal of Economics*, July.

McCombie, J. (1985), 'Increasing Returns and the Manufacturing Industries: Some Empirical Issues', *Manchester School*, March.

McCombie, J. and J. de Ridder (1983), 'Increasing Returns, Productivity and Output Growth: The Case of the United States', *Journal of Post Keynesian Economics*, Spring.

McCombie, J. and A.P. Thirlwall (1994), *Economic Growth and the Balance of Payments Constraint*, London: Macmillan.

McCombie, J. and A.P. Thirlwall (1997), 'The Dynamic Harrod Foreign Trade Multiplier and the Demand Oriented Approach to Economic Growth: An Evaluation', *International Review of Applied Economics*, January.

McCombie, J. and A.P. Thirlwall (2004), *Essays on Balance of Payments Constrained Growth: Theory and Evidence*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Maddison, A. (1970), *Economic Progress and Policy in Developing Countries*, London: Allen and Unwin.

Michl, T. (1985), 'International Comparisons of Productivity Growth: Verdoorn's Law Revisited', *Journal of Post Keynesian Economics*, Summer.

Okun, A. (1962), 'Potential GNP: Its Measurement and Significance', *Proceedings of the Business and Finance Statistics Section of the American Statistical Association*.

Robinson, J. (1956), *The Accumulation of Capital*, London: Macmillan.

Solow, R. (1956), 'A Contribution to the Theory of Economic Growth', *Quarterly Journal of Economics*, February.

Solow, R. (1957), 'Technical Change and the Aggregate Production Function', *Review of Economics and Statistics*, August.

Swan, T. (1956), ‘Economic Growth and Capital Accumulation’, *Economic Record*, November.

Thirlwall, A.P. (1969), ‘Okun’s Law and the Natural Rate of Growth’, *Southern Economic Journal*, July.

Thirlwall, A.P. (1979), ‘The Balance of Payments Constraint as an Explanation of International Growth Rate Differences’, *Banca Nazionale del Lavoro Quarterly Review*, March.

Verdoorn, P. (1949), ‘Fattori che Regolano lo Sviluppo della Produttivita del Lavoro’, *L’Industria*, No. 1.

Young, Allyn (1928), ‘Increasing Returns and Economic Progress’, *Economic Journal*, December.

Young, Alwyn (1995), ‘The Tyranny of Numbers: Confronting the Statistical Realities of the East Asian Growth Experience’, *Quarterly Journal of Economics*, August.

5. The representative firm and increasing returns: then and now

Stephanie Blankenburg and G.C. Harcourt

I

There is something archaic, yet modern about the tone and issues of the 1920s debates in the *Economic Journal* on the representative firm and increasing returns, often referred to as 'the cost controversy'. Then, as now, applied economists, 'realitics', Sir John Clapham called them, and theoretical economists ('analytics') were often poles apart, who neither properly understood or appreciated each other's roles and approaches. Then, as now, views differed on whether or not theory had to be directly applicable in explanations of 'real world' observations, and much misunderstanding occurred because the separation between logically coherent 'high theory' in its own domain and, a separate issue, its direct applicability, was not made by protagonists. Or, one side would be concerned with the former, the other with the latter, neither making this understanding explicit.

For this, Marshall must take much blame. Though the distinction was clear to him, in the arguments of his *Principles* it is purposefully blurred because he wished to be read by businessmen (*sic*). Its text often reads as a narrative about real-life happenings, admittedly explicitly confined mostly to normal periods but meant to call up in readers' minds their own observations and experiences. Yet Marshall's powerful theoretical mind provided an underlying but hidden structure, confined to footnotes and appendices, of explicitly set out theoretical pre-suppositions and arguments. It took the same subtlety of mind (and foxiness of character) as Marshall's to interpret the text.

At this time, the distinction between theory and practice was also being worked out, not least by Frank Knight in *Risk, Uncertainty and Profit* (1921). The classical notion of a freely competitive environment, which in the writings of Adam Smith and Karl Marx was the dynamic setting for distribution, accumulation and growth as well as a theory of price formation, was being refined into the rigorous but static characteristics of the price-taking model of today's microeconomics courses. This became a coherent, rigorous, logically watertight but basically un-illuminating, ineffectual explanation of the behaviour of actual firms, industries and economies. Though price-taking behaviour is hinted at in the *Principles*, its twin assumption of perfect foresight, which together define perfect as opposed to pure competition, was conspicuously absent from the *Principles* (but not from its underlying structure of a long-period stationary state).

Because it is often not clear whether the long or the short period or both are being analysed, there is not always a clear distinction made between analysis of production and price-setting, on the one hand, and accumulation and investment, including the choice of technique within the investment decision, on the other. When the shapes of cost curves were examined, it was not always clear whether a production or an investment decision was being analysed, that is, whether when a change in production was considered, the analyst was asking what may reasonably be expected of changes in costs, if any, in a given situation of the here and now; or, what happens to the level of long-period costs at different potential levels of production of the plants chosen to be invested in. These points became clear in Edward Chamberlin's discussion in Appendix B to the sixth edition of his *Theory of Monopolistic Competition* (1950 [1933]) of the long-period envelope cost curve. He overcame the misconceptions in Jacob Viner's classic statement of the issues, misconceptions associated with the contribution of his obstinate Chinese draftsman, to show that calling the long-period curve an envelope was a serious misnomer, because the curve had no separate existence from the possible short-period plant curves in a given situation, the relevant parts of which it is made up.

II

There was always a potential contradiction in Marshall's 'vision'. He had 'invented' supply and demand analysis to allow him to handle that elusive but vital variable *time*, by distinguishing between the market, short and long periods where what was locked up in the *ceteris paribus* pound was decided by the economist/analyst. He also saw societies as evolving interrelated organic entities, and he knew that the static analysis of his supply and demand functions and periods could not handle this in a fundamental way. Though he wanted 'economic science' to be fruit- as well as light-bearing, he was also committed to defending a competitive environment which was sustainable, despite substantial changes over time in institutions. He wanted his 'trees in the forest' analogy of competitive conditions to survive the emergence of monopolies and oligopolies, and the emergence, then dominance, of joint stock companies as the main form of industrial organisation. He knew of what he considered to be Cournot's error (as a description of the real world, not as a logical argument in its own domain), that increasing returns in firms and competition could not be enduring bedfellows. Once one firm got ahead of its rivals it could undersell them and take over the industry.

Marshall's answer to this was not couched in terms of Schumpeter's creative destruction – a temporary monopoly firm exploiting an innovation. This makes monopoly profits attract imitators in to compete them away – nor Joan Robinson's fish in a pond whereby bigger fish dominate – eat up – smaller ones. Rather he introduced the device of the representative firm, which allowed the real world facts of increasing returns, internal and external economies, to be accommodated without destroying the viability of competitive environments in the long run. The representative firm was a forerunner of modern representative agent, the purpose of which is to stand in for the behaviour of the group (for Marshall, the industry, for the moderns, the whole economy) while having no actual existence in reality.

There were minor policy recommendations allowed; they were spelt out in more precise analytical detail than Marshall would have done by his chosen successor, Pigou. By distinguishing between private and social costs and recognising the existence of decreasing and increasing returns, Pigou built on Marshallian foundations to establish carrot and stick measures to induce social optima of a limited, commonsense variety, the fare of *The Economics of Welfare*.

Here Clapham entered the fray with his teasing, emperor has no clothes, article, 'On empty economic boxes' (1922). Clapham was a no-nonsense economic historian with detailed knowledge of agricultural, industrial and financial firms and industries, and of the interrelated systems to which they belonged. He had read Marshall and Pigou, but felt the gap between their careful and precisely stated propositions (especially Pigou's) and the complexity arising from his own observations was virtually unbridgeable. Especially was this so when the preliminary step to advocating a subsidy or a tax was to classify specific industries as subject to either decreasing or increasing returns. He illustrated this by a visit to a hat factory; he asked in which of the boxes labelled increasing or decreasing returns industries would the specific industry to which the hat factory 'belonged' be found. He wondered whether the neat products of theory – homogeneous, competitively sold – could ever match the variety of actual products coming under the general heading of hats. He wondered whether the definition of an industry could ever be pinned down in practice; and so on. By doing so he contributed to the general debate, on both sides of the Atlantic, concerning the natures of applied and theoretical work. This involved a tightening up of the definition of perfect competition, taking it away from the looser but more dynamic concept of free competition of the Classicals, Marx and Marshall to the price-taking, perfect foresight model differentiated from pure competition, which only assumed price-taking. (Smith, Marx and Marshall would never have tried to analyse a world in which the future was known with certainty.) In the December 1922 *Economic Journal* (Clapham's article had appeared in September, things were published quickly then), Pigou (1922) replied to his fellow King's man. He thought that the categories 'analytic' and 'realistic' were themselves empty. Pigou distinguished between two sorts of knowledge. First, 'pure' knowledge about implications such as we find in logic and mathematics. The second is realistic knowledge, knowledge of subject matter presumed to be actual, the characteristics of the data studied by physicists. He classified knowledge into that which can or cannot give direct help in the practical conduct of affairs (always Pigou's emphasis).

If Clapham were to be taken literally, there was no point in analysing returns, because even if we filled the empty boxes, we would still receive no help in practice: A *non sequitur*, says Pigou, adding that by far the greater part of knowledge which history aims at is totally irrelevant in practice. Nevertheless, because knowledge *by itself* is of little value we must make more Jevonses, people at home in both fields. Meanwhile, we should substitute cooperation between the two types for quarrels on the basis of imperfect understanding of the deficiencies of one another's method.

III

Throughout the debates, explicit mention is made about whether the short or the long period or both is the appropriate context. An explicit distinction is made between the processes of production, on the one hand, and accumulation or investment, which allows new methods to be put in place in firms and industries, on the other. It must be said that these aspects were a source of much confusion and muddle which were not cleared up until after the 1930 *Economic Journal* symposium.

The watershed which settled these issues is Chamberlin's Appendix B (1950 [1933]) on the relationship between the short-period average cost plant curves and the long-period misnamed envelope average cost curve – the envelope curve is not an envelope, containing something else, but segments of (with continuity points on) succeeding plant curves showing the least-cost method of producing levels of output in the given situation. The reasons for the U shape of the plant curves are completely different from those responsible for the U shape of the long-period curve. The former is due to the application of the law of variable proportions to production with each possible plant; the latter is due to economies of scale outweighing diseconomies as we pass from one plant to another, and then to a reversal of the dominance between contending factors.

The construction of the long-period curve from its constituent plants tells observing economists (businesspeople and profit-seeking accumulators) the appropriate choice of technique for anticipated future levels of output. It is relevant for the investment decision now, and, once embodied, for later production decisions. In a full analysis, such as we get in Wilfred Salter's 1960 classic, the amount of investment and the change in capacity in individual firms depends on how many existing plants remain profitable – positive expected quasi-rents in any given situation – and how much production deemed to be profitable needs to come from the new plant and methods introduced in current investment expenditure associated with optimum points on the long-period curve. Seen thus, cross purposes and non-meeting of minds in the debates of the 1920s disappear.

IV

In the 1930 symposium, Piero Sraffa made clear where he stood on Marshall's theory:

> We seem to be agreed that [Marshall's] theory cannot be interpreted in a way which makes it logically consistent, and at the same time, reconciles it with the facts it sets out to explain. Mr. Robertson's remedy is to discard mathematics, and to suggest that [Sraffa's] remedy is to discard the facts; [. . .] I ought to have explained that . . . it is Marshall's theory that should be discarded. (Sraffa 1930, 93).

At the Corfu Conference on capital theory, Sraffa (1961) set out the standards he set himself (and the subject) concerning measurement, theory and criteria they should meet within the context of capital theory. He distinguished between measurement by statisticians '. . . only approximate . . . provided a suitable field for work in solving index number problems' and theoretical measures which 'required absolute precision. Any imperfections . . . knocked down the whole theoretical basis'. The definition of capital in theory must therefore be kept 'separate from the needs of statistical measurement . . . J.B. Clark, Böhm-Bawerk and others intended to produce pure definitions of capital, as required by their theories . . . contradictions . . . pointed to defects' (Sraffa 1961, 305–6).

Sraffa's 1925 and 1926 articles (the first pages of Sraffa 1926, constitute a summary of the complex, detailed arguments of the 1925 article) are principally attacks on the logical foundations of Marshall's theory of competitive value. (Often it is Pigou's version Sraffa has within his sights.) The theory depends on 'the fundamental symmetry existing between the forces of supply and demand', using the method of partial equilibrium ('particular equilibrium' in Sraffa 1926, 539), so that 'the essential causes determining the price of particular commodities may be [. . .] grouped together so as to be represented by a pair of intersecting curves of collective demand and supply' (Sraffa 1926, 535). Sraffa concludes that the only case which is *logically* consistent with the approach is one of constant costs 'in respect of small variations in the quantity produced' (ibid., 541). Constancy comes from an 'absence of causes which tend to cause the cost either to increase or diminish', not from an improbable 'accidental balancing of two opposite tendencies', as Sidgwick and Marshall would have it (ibid., 542).

The criticism is a search for conditions which allow the method logically to be applied. Sraffa reminds us that the 'laws' of returns historically belonged in different parts of the discipline, diminishing returns in the theory of distribution (rent), not in a discussion of *relative* values and prices, increasing returns in discussions of 'general economic progress', not in a discussion of 'increase[s] in the scale of production' (ibid., 537). Combining them, as modern theory did (and does) in a theoretical whole, was restrictive and contradictory. Especially was this true of increasing returns, for to be consistent with competitive theory (they could be a fact of life), the division of labour had to be 'limited to the case of independent subsidiary factories coming into existence as the production of an industry [increased]' (ibid., 537). Internal economies associated with the growth of individual firms had to be 'entirely abandoned [. . .] incompatible with competitive conditions' (ibid., 537–8), so 'external economies' were more and more emphasised.

Diminishing returns were associated with the existence of a fixed factor and in (then) modern theory (but not Classical Political Economy) with the short period. So, the more broadly an industry was defined, the more likely it was that diminishing returns would arise from the use of the one factor. Conversely the more homogeneous we made the commodity within an industry (move from agriculture to fruit) 'the greater will be the possibility that the forces which make for increasing returns will predominate' (ibid., 538). Time has similar effects, especially the time allowed for adjustments to occur and affect returns.

Most importantly, the independence assumption relating to demand and supply schedules is at risk: 'It is precisely in [the] category that . . . when a variation in the quantity produced by [an] industry . . . sets up a force which acts directly . . . upon its own costs [and] the costs of other industries [so that] the conditions of the "particular equilibrium" which it was intended to isolate are upset . . . that the applications of the laws of returns fall, in the great majority of cases' (ibid., 539). For diminishing returns, we are left with 'that minute class of commodities in the production of which the whole of a factor is employed' (ibid., 539). For increasing returns, we have those economies ('most seldom to be met with [in practice]') external to the firm but internal to the industry (ibid., 540). So, *if we are to use partial equilibrium analysis*, 'in normal cases the cost of production of commodities produced competitively . . . must be regarded as constant in respect of small variations in the quantities produced'. (ibid., 540–41).

In the 1926 article, Sraffa rejected general equilibrium as a way out, spawning instead the imperfect/monopolistic competition revolution permitting lower costs at higher rates of output because of demand constraints in imperfectly competitive markets.

V

Joseph Schumpeter's 1928 *Economic Journal* article on the instability of capitalism is profoundly cerebral. He asked: if we abstract from political and other shocks, including certain classes of technical advances, is a competitive economy a stable system in the sense that it tends to a repeatable level of operation? (Walras's ambition was to show that it was.) Schumpeter reveals his views on the nature of economic theorising and on what he means by stability. In the 'real world' a system could behave in an unstable way because of political shocks, for example a war or a natural calamity, yet be internally stable in its *ceteris paribus* operation. This distinction justified his view that while Walras was the role model for theorists, his vision of the impact of innovations in the accumulation carried out by his entrepreneur heroes and continuous technical change could be logically fitted into a theoretical analysis starting from a Walrasian system in equilibrium.

His article appeared in the same issue as Lionel Robbins's article on the representative firm. Whitaker, in a 1989 essay, judged it not to have been of lasting value but much notice was taken of it at the time. Robbins argues that for Marshall the representative firm is an afterthought, and that he never corrected the section of the *Principles* containing it, quoting Keynes in support.

His criticism turns on the meaning of 'average' implied by 'representative', and whether in equilibrium a representative or average entrepreneur (or manager) is a viable or necessary concept. He has more Hubert Henderson's entrepreneur or manager than Marshall's practice in mind. Robbins followed up this discussion with a more lasting contribution on the ambiguity of the notion of stationariness in economic analysis, a discussion which has relevance for misunderstandings in capital theory, even today.

VI

We come to the climax: the symposium organised by Keynes in the March 1930 *Economic Journal*. There were three contributors – Dennis Robertson, Sraffa and Gerald Shove. Keynes saw Robertson as a stout defender of old Marshall, representative firm in hand, Sraffa as extremely negative (so what is new?) and Shove as original and constructive, working within the Marshallian tradition but not accepting the representative firm as the only way of tackling the issues.

Robertson responded to Robbins's view that normal profit is needed to keep particular individuals in an industry, so it is determined by what they could get elsewhere, and is therefore not a unique figure for the industry (or economy as a whole). Robbins had criticised Henderson on these points. Robertson now defended him. Henderson was describing those on the road to elimination because of their inadequacies. At any moment of time, however, similar types enter, leaving the *industry* in equilibrium. Robertson reiterates the proposition that the concept of the representative firm is essential for understanding the theory of increasing returns, because it helps us to tackle Cournot's difficulty concerning increasing returns. He lists recent answers. First, Schumpeter (and others) deny the validity of a long-period falling supply curve (except as a record of historical events); it cannot be the series of conditional statements it is required to be. For Robertson this is a counsel of despair, which Marshall considered and rejected (so that's alright then). Marshall stressed that the *firm's* experience is not equivalent to the industry's experience, especially in the long period.

Sraffa resorts to a theory of monopoly which yields a determinate result (always Sraffa's aim for theory). Robertson reminds us that it is always all in Marshall, this time in footnote 1, p. 458, as Marshall's auxiliary argument on the economics of imperfect competition. Pigou's answer depends upon there being external economies of large-scale production *at the level of the industry* even though individual firms work under conditions of decreasing returns. Pigou accepts *internal* economies of large-scale production but Sraffa rules these out, not because they do not exist but because Marshall's partial equilibrium method could not then be used. For Pigou, the representative firm is one for which there is an optimum size for each scale of *aggregate* output, beyond which there are no further internal economies. (He also transformed the representative firm into the equilibrium firm.) For Pigou, but not for Sraffa, external economies are sufficient to ensure this.

Allyn Young (1928) extended the concept of external economies to include lower costs resulting from progressive division and specialisation of industries, the essential process by which increasing returns are realised.

Pigou suggested that these external economies allow increasing *internal* economies, leading to an increase in the size of individual firms as industries expand, due to growing specialisation. While both possible and plausible, it comes from neither Marshall nor observed fact. Robertson thinks it is the response to increasing demand for the industry's products by individual firms wanting to get the benefits of large-scale organisation and plant, which are known but not yet brought into existence by specialisation or by other ways. Marshall's trees of the forest explain this idea and allow increasing returns to exist in the industry and competition between firms to prevail. Individual firms are *not* representative firms and they must expand as the representative firm does when overall demand increases. No one firm will ever have a monopoly of the whole trade. How, asks Robertson, would Robbins make this point if he scrapped the representative firm?

Enter Sraffa: why, if internal economies were available were they not taken advantage of? If they are, why do they cease at (long-period) equilibrium – a point of constant returns! Sraffa's answer was to find the assumptions implicit in Marshall's theory; in the event, he thought Marshall's theory should be discarded.

Shove accepts Robertson's piety as one way out, and that his account of the representative firm in Marshall's theory is correct. Shove also sides with Robbins, arguing that the problem can be solved without introducing the representative firm. In showing why he drew on an 'unpublished study of the relations between cost and output on which [he had] been engaged for some years' (Shove 1930, 94).

First, we allow for those characteristics of long-period equilibrium which Marshall intended the representative firm to display. If an entrepreneur has a special aptitude for the work he/she is engaged in, his/her earnings will be much greater than the return he/she would get elsewhere, and that which others in the industry with the same *general* capacity, but not special advantages, receive. Rent to the former is compatible with equilibrium and admits of wide variations in the rate of profit received by individual entrepreneurs. Secondly, luck plays an important part in determining actual earnings; so equilibrium requires not the equality of *actual* earnings, but of the mathematical *expectation* of earnings for similar units in different uses. Thirdly, the profits of the firm (apart from luck) differ at different stages of its life, so there is no tendency to exhibit the rate of profit unless firms are at the same stage of development. (The same is true for skilled employees, whose earnings are low when learning, rise to a maximum when fully equipped, and decline in old age.) Fourthly, the equilibrium of the industry does *not* imply that all (*any*) of the individual firms are in equilibrium. Equilibrium requires that *aggregate* output is unchanged; some firms may be expanding, others declining, so as to offset each other overall. For equilibrium, a *general* expansion or contraction should not be profitable.

The representative firm is a brilliant device for displaying facts when we want to depict equilibrium as resulting from the rise and fall of individual firms. When we look at the ebb and flow of resources of all kinds from one use to another, a different method is needed.

'We must then say that the attraction exerted by a particular occupation on a particular unit is the mathematical *expectation* of earnings in that occupation for a unit of resources with *the character and aptitudes of this one* (thus allowing for (1) and (2) [above]); that this expectation is to be reckoned by summing *the series of its probable earnings* at each stage of its career (thus allowing for (3)); and that it must be calculated on the assumption that resources within the industry . . . are distributed between the firms in the most profitable way (thus allowing for (4)).' (Shove 1930, 96, emphasis in original).

Shove argues convincingly that there are numerous equilibrium points, each one corresponding to a specified starting point. (Path-dependence is not *that* new.) For Shove, the economic problem represented by the real world is more a question of sorting out and fitting into appropriate niches a vast number of heterogeneous individuals and activities, than of regulating and directing into proper channels large homogeneous streams of standardised productive agents. Robertson disagrees because he wants to build up schedules by aggregating the supply schedules of various factors of production. This can only be done if factors of production can be expressed in quantities alone, that is, in a measure independent of value!

For Shove, when we try to unravel the relations between industry size and the efficiency of the businesses which make it up, we must distinguish between, first, improvements in organisation due to an increase in the output of the industry as a whole; and secondly, improvements due to increases in the output of individual firms, industrial output remaining unchanged. Marshall rolled the two together. Expansion under the second head causes increases *and* decreases in efficiency, so the most efficient size is where they just offset each other. Under the first head, all changes which increase efficiency must be related to increases in industry size alone. Increases in the size of firms are, therefore, governed by increases in the size of the industry. (Shove cites Marshall, 1920, 459–60, for confirmation.)

Shove's most telling point is that *time* is missing from Pigou's and Robertson's diagrams (see Pigou 1928 and Robertson 1930). They overlooked the crucial point that the enlargement of the total volume of trade may increase the *speed* at which individual businesses can grow. His final *coup de grace* is that Marshall's trees of the forest analogy is no longer typical. Most firms survive indefinitely, rising and falling with changes in the conditions of trade (that is, over the cycle). Marshall had recognised this by changing 'is' to 'was' in later editions of the *Principles* when he could no longer ignore the rise of joint stock companies.

VII

Twenty years later, Schumpeter credited Keynes with 'a stroke of editorial genius' for having organised 'a symposium [. . .] on the matter [. . .] that is still eminently worth reading'. (Schumpeter 1994 [1954], fn. 53, 1046). This remains true of the symposium and of some earlier contributions. Whereas the debates *are* archaic from a purely technical point of view – the long-period U-shaped cost curve has achieved textbook status, the toolbox of microeconomics has been extended to accommodate imperfect competition – they also raise profound highly topical conceptual and methodological issues. The most important are Sraffa's plea for realism in economics, and Young's argument for the cumulative nature of economic progress.

Sraffa's critique of Marshall extends beyond his careful analysis of the limiting cases in which partial equilibrium analysis can accommodate an industry supply curve with variable costs to a fully-fledged criticism of marginalism. His central concern is with the artificial restrictions a symmetrical theory of prices based on marginalist reasoning imposes on the analysis of production and accumulation in that it generates, among other things, an *a priori* requirement of a functional relationship between costs and output. What he objects to is the consequent lack of realism of the marginalist 'theory of competitive value'.

For Sraffa, the initial criterion by which to judge the validity of any theory is logical consistency. If the theory can be made logically consistent only for a few exceptional cases, it lacks explanatory power. This sums up his approach to Marshall's *ceteris paribus* construction. While Sraffa pays express attention to Clapham's call for 'realisticness' (1925, 323), he rejects a view, potentially implicit in Clapham's 'empty boxes' critique, that would reduce theorising to a purely descriptive exercise, or that would regard any empirical observation that contradicts a theoretical hypothesis as sufficient to discard a logically coherent theory. As he explicitly recognises, that partial equilibrium analysis is useless for any practical purposes does not exclude the possibility that this is purely due to given scientific limitations in handling complexity. So the realism of the approach might increase, *without the need to fundamentally question or abandon essential premises such as the symmetry assumption*, once the subject has advanced sufficiently 'to extend the field of investigation so as to examine the conditions of simultaneous equilibrium in numerous industries' (1926, 541).

In this sense, Sraffa's critique of partial equilibrium analysis is an internal one, and also one that excludes a naive empiricist understanding of realism. The external dimension of his critique of marginalism focuses, not on faithfulness to empirical detail, but on the role of *a priori* reasoning in Marshall's and marginalist analysis: a theory is unrealistic, not because it is incomplete, but because the movements it can accommodate and the categories it can handle *are generated by the requirements of the theory itself* not by limitations of knowledge. They are figments of the mind, purely imaginary thought experiments with no roots, however remote, in objective reality, be this collective human history or the physical world. Thus Sraffa's complaint that 'the "external economies" peculiar to an industry, which make possible the desired conciliation between scientific abstraction and reality, are themselves a purely hypothetical and unreal construction, is something that is often ignored' (1925, 347). Similarly, he concludes his discussion of the hypothetical possibility of general equilibrium analysis to resolve the problem of cost–output interdependencies by remarking that, for increasing returns, we are left with 'the impossibility of confining within statical conditions the circumstances from which they originate' (1926, 541).

While some concepts of marginalist theorising are explicitly subjectivist – 'the "demand function" is based on an elementary and natural hypothesis, that of decreasing utility' (Sraffa 1925, 325) – functional relationships to explain production are 'the result of a much more complicated set of hypotheses' (ibid.). Such functional relationships – Marshall's supply curve, but also Wicksteed's, according to Sraffa, ill-conceived distinction between 'descriptive' and 'functional' curves – are misleading by pretending to represent objective reality when they are just as much a figment of the imagination as 'utility': Wicksteed's 'functional' curves are just as dependent on the observer's point of view as his 'descriptive' curves, and not derived from any 'universal law' of decreasing productivity (Sraffa 1925, 335–8). Marshall's supply curve, as all marginal reasoning, is a *thought experiment* depending on the fictitious and arbitrary assumption of piecemeal continuous change, not a line that traces out real historical time, or even actual experimental results (Sraffa 1925, 358, fn. 80).

Such constructions are truly 'empty boxes à la Sraffa', not scientific abstractions that, in due course, can be filled with concrete contents, even if at the price of some omission of empirical detail. They are *a priori* idealisations that can, for this reason, never be filled. Sraffa makes little secret of his view that such idealisations serve the purpose of ideology: just as the 'laws' of increasing and of decreasing productivity were originally formulated in specific historical contexts and pertained to differing areas of economic theorising – the 'theory of distribution' and 'the analysis of production' (Sraffa 1925, 324) – so the attempt to unify them, irrespectively of their intellectual and material historical roots, under the single banner of a 'law of non-proportional productivity' (ibid.) is motivated by an attempt to justify a particular view of the world: that 'tranquil view which the modern theory of value presents to us' (Sraffa 1926, 536) of stable markets and egalitarian capitalism.[1]

The essential point concerning Young's 1928 contribution is the difference between his treatment of increasing returns and 'external economies' and that provided by Marshall, and the difference of both their notions from Pigou's 'externalities'. Marshall and Young had Adam Smith in mind when defining and specifying the concept of increasing returns, the idea that an increase in the extent of the market allows for a reorganisation of the productive process that results in an increase in labour productivity. Apart from Young's crucial insight that the division of labour also acts to expand markets, there are two main differences between Young and Marshall: first, Young explicitly rejects Marshall's view that external economies arise within an industry. Instead, it is crucially important to recognise that such economies operate at an economy-wide level, i.e., that 'industrial operations be viewed as an interrelated whole' (1928, 539). Secondly, while Marshall explicitly excludes 'any economies that may result from substantive new inventions' (Marshall 1920, 460), thereby leaving open the door to a static interpretation of increasing returns (that is, marginal increasing returns to scale arising from equi-proportional factor changes and fixed input costs), Young makes technological change a cornerstone of his analysis of dynamic (non-marginal) increasing returns to scale. Ultimately, Young's version of Sraffa's realist critique of marginalism emphasises (with Adam Smith) the *complementary relation* between factors of production and replaces the principle of partial factor variation with that of the cumulative nature of economic progress: not only are increasing returns of a dynamic and inter-sectoral nature (including technological change, whether of a piecemeal or a radical kind, variable input costs and non-proportional factor changes), but diminishing returns to capital are unlikely ever to set in, because capital investment embodies technical change. Neither Marshall's half-heartedly 'dynamic' version of 'external economies' nor Young's much more radical interpretation of them have much in common with Pigou's ingenious, but totally static, definition of an 'externality' as representing a divergence between private and social costs.

VIII

The 1920s controversy effectively ended the reign of Marshallian economics as the dominant paradigm of economic thought. The discussion moved on in three main directions: the exploration of the potential of Walrasian (and Paretian) general equilibrium theory (Arrow–Debreu–Hahn), Sraffa's alternative asymmetric theory of prices, distribution and reproduction, and endogenous theories of innovation, growth and development based on the principle of cumulative causation (Kaldor, Myrdal, Schumpeter). While Sraffa's 1960 contribution was largely ignored by the mainstream, neoclassical general equilibrium theory was undermined by a number of developments, including the results of the capital theory controversies of the 1950s–1970s, the defeat of the microfoundations project based on aggregative econometrics exemplified in the impossibility of removing the arbitrariness of aggregate excess demands and the failure to provide a proof of the existence of a general equilibrium position for the case of imperfectly competitive economies (Rizvi 1994).

More recently, the 'new classical macroeconomics' has taken over. Its perhaps most remarkable feature is its collective oblivion of the impasse reached by general equilibrium theory in the 1970s, and its return to pre-1970s theoretical constructs that had been confined (mostly) to the dustbin of the history of economic analysis. Pasinetti's assessment of the fate of the capital theory controversies equally applies to the other developments mentioned above: 'Amnesia on such a vast scale can only be explained by more appropriate terms, such as "suppression" or "repression" or "removal". This is, perhaps, one of the most intriguing examples of that process described by Kuhn [. . .], through which dominant "normal" science suppresses, and thus ignores, the cases of contradiction and anomaly it bears within.' (Pasinetti 2000, 412). The result is a renewed surge in 'empty box' reasoning in Sraffa's sense, the use of purely imaginary, sometimes logically inconsistent, concepts and tools 'as if they were part and parcel of everyday economic reality, not the slightest doubt being shown about them' (Pasinetti 2000, 416).

A good example of this tendency is the New Endogenous Growth Theory (NEGT). The obvious case of the use of aggregate capital apart, we wish to briefly raise a less well recognised case of 'empty box' reasoning, that is central to the NEGT's conceptualisation of 'the knowledge factor'. NEGT is the latest attempt to unpack the Solowian 'growth residual', its aim being to provide an endogenous explanation of technological progress and economic growth within a general equilibrium framework. A principal claim to originality is that NEGT can accommodate increasing returns, and 'knowledge' is accorded a central role in the explanation of modern technical progress and growth dynamics.

Analytically speaking, the main characteristic of NEGT is the absence of diminishing returns to capital (for example, Romer 1994, 13–14; Solow 1994, 49). In the Solow–Swan model, diminishing returns to both factors of production (physical) capital and labour, together with constant returns to scale, are instrumental for the outcome of zero growth per capita in the long period. In contrast, in the NEGT, the rate of profits no longer tends to fall and, consequently, the major conclusion of the Solow–Swan model is reversed: an increase in the saving rate *can* raise the growth rate of the economy permanently, and the thrifty entrepreneur/head of households has been restored to his/her rightful place as the pillar of efficient, thriving and free market societies.

This basic message of the NEGT is most clearly encapsulated in the basic AK models in which reproducible inputs are the main source of endogenous growth (such as Frankel 1962; Romer 1986; Lucas 1988). Assuming linearity in the differential equation for the production of the factor capital (now meaning all accumulable factors of production) derived from the standard Solow–Swan production function and the equation for capital accumulation, these models can be interpreted in two ways: first, the absence of diminishing returns to capital can be attributed to the elimination of all non-accumulable factors of production from the production function. If, in the Solow–Swan model, diminishing returns to capital are the consequence of taking all other determinants of aggregate output as given implying that labour is non-accumulable in the sense that an increase in output will require a more intensive use of physical capital, the AK models redefine labour as human capital, merging it with physical capital into a single factor. In effect, this assumes an unlimited supply of high quality labour, and hence an exogenously given constant real wage and constant rate of profits, independent of the amount of capital employed. Income distribution is thus technologically determined, and if the technology employed uses only self-reproducing inputs, perpetual motion is generated with its rate depending solely on the determinants of saving behaviour and the investment-saving mechanism (see Kurz and Salvadori 1998; also Rebelo 1991; Solow 2000). Secondly, AK models are a special case of Arrow's 'learning-by-doing' model (Arrow 1962) with a unity elasticity of learning, the difference being that the source of increasing returns at the aggregate level is specified by linking it to learning processes rather than attributing it to capital accumulation *per se*. Either way, there is an important drawback from the viewpoint of neoclassical distribution theory: in discarding the scarcity assumption (decreasing demand curves for capital and labour and partial factor variation), the long-term growth rate no longer equals the rate of growth of the labour supply and factors need no longer be rewarded at their marginal productivity. Consequently, the compatibility of the long-term growth path with full capacity utilisation through market forces is no longer guaranteed.

The next generation of NEGT models remedied this. They complement the basic analytical structure of the AK models by providing a microeconomic market imperfections story of technical change, where such imperfections affect the rate of growth, not only the level of output of an economy. Once it is assumed that the consumption (saving) choices of a representative agent engaged in (infinite or overlapping-generation horizon) intertemporal utility maximisation directly affect the rate of change of productivity, there is no limit to the factors that may affect this choice. Even so, the so-called 'knowledge factor' has been central to the NEGT's explanation of economic growth through technical change in the so-called R&D models (Uzawa 1965; Romer 1990; Aghion and Howitt 1992, 1998; Grossman and Helpman 1991). These models reintroduce the scarcity assumption by distinguishing between accumulable (human capital, education, ideas and design, and so on) and non-accumulable factors of production (standard physical capital and labour) For simplicity, we call the accumulable factor 'knowledge'. This factor of production is *non*-scarce because it is independent of the supply of labour. Knowledge is assumed to be a self-generating technology, entering the aggregate production function as a constant of some form (a constant share of labour time dedicated to knowledge generation, a constant 'amount' of knowledge or a constant increase in labour-augmenting technology generated by innovations). This assumption is combined with microeconomic stories that render technical progress endogenous, not through the formal abolition of decreasing returns to (combined) capital, but in the sense that it is 'a special resource-using, profit-seeking activity with its own technology' (Solow 1997, 17).

Whichever the microeconomic story told – horizontal innovation (Romer, Grossman and Helpman) or vertical innovation (Aghion and Howitt), for example – there remains an inconsistency: Knowledge *as a factor of production* is self-generating, that is, non-scarce in these models. As with the AK models, the purely *analytical* assumption of an absence of diminishing returns to the accumulable factor is what ensures a balanced growth path in the long period. The (micro)*economic* explanation of balanced growth dynamics in R&D models is, however, solely concerned with increasing returns to scale. Significantly, this shift in emphasis from the 'theory of rent' to the 'theory of economic progress' involves a re-interpretation of *knowledge as a (semi-) private good.* R&D models of endogenous growth explain increasing returns to scale as arising from *externalities*, caused by market imperfections in the process of innovation. Differently from the AK models (where the nature of the externality is not specified), the externality is now identical, in nature, with the accumulable factor of production: knowledge does not just generate externalities, it *is* an externality. The notion of externalities – unlike the broader notion of external economies, that is, dynamic (non-marginal) increasing returns to scale – has meaning only within the static neoclassical benchmark notion of efficient market allocation. Something is an externality because it is external to the market, that is, it cannot be transacted through the market and will therefore not be 'properly' compensated. Efficient market allocation is, in turn, a function of two factors: that goods be private goods (that is, their use is non-rivalrous) and that goods and factors be scarce. The analytical interpretation by NEGT of knowledge as a factor of production suggests that knowledge should be conceptualised as a pre-existing 'public fund' or 'free good' (for each period of production) on which standard factors of production can draw. What is emphasised by the microeconomic conceptualisation of innovation processes is that knowledge is not only non-rivalrous, but also *partially excludable*. This implies that market allocation is at least possible, if not efficient: 'The feature that makes a good collective rather than private [. . .] is the possibility of simultaneous enjoyment of the good, not the possibility of preventing others' enjoyment. The first issue deals with the efficiency of market allocation, the second with its feasibility. Market allocation of public goods may indeed be feasible, but that does not make it efficient' (Marglin 1984, 467–8).

This emphasis on the possibility of 'privatising' knowledge is difficult to reconcile with its presumed ability to generate (at least) constant returns to the accumulable factor, since introducing private property rights to allow for the private appropriation of returns from knowledge will generate artificial scarcity, certainly with regard to the use and reproduction of knowledge, probably also with respect to its accessibility. This 'paradox' of the R&D models is easily overlooked in the fog created by the concept of 'externality' that functions as a 'bridge' between the analytical dynamics of endogenous growth (knowledge as a factor of production) and the static analysis of allocative mechanisms (knowledge as a good). In the end, neither the AK nor the R&D models provide a substantial explanation of *dynamic* increasing returns to scale. In the first case, no attempt is made and the validity of neoclassical value (and distribution) theory is simply assumed. In the second case, the notion of externalities allows NEGT to ignore the implications arising from its formal formulation of balanced growth dynamics in favour of an economic interpretation that opens the way to safeguard, if not competitive value theory, at least the institutional framework of a private market economy.

IX

For this reason, ‘knowledge externalities’ are as much an ‘empty box’ as Marshall’s external–internal economies: They are *generated as an a priori requirement of the theory*, not derived from the observation of objective conditions, historical or physical in nature. They serve absolutely no explanatory purpose, but are very convenient to maintain the ideological illusion, not of the symmetry of demand and supply directly, but of a balanced growth path achieved by private market economy.

There is, in this respect, a certain parallel between the role played by ‘knowledge externalities’ in NEGT and the concept of the ‘representative firm’ in Marshall’s (and his defenders’, in the 1920s debates) attempt to safeguard marginalist analysis in the presence of increasing returns. For our purpose, there is no need to decide whether the ‘representative firm’ was meant to encapsulate the growth path of a firm or average (normal) industry expenses for a given aggregate volume of production in competitive conditions (especially as Marshall maintained his usual ambiguity on this question). What matters is that Marshall considered this concept particularly relevant for an explanation of industries operating with increasing returns (*Principles*, 1920, 376) and that it allowed him to ignore the impact of *radical* innovations at the *aggregate* (industry) level, and thus to reconcile firm disequilibrium with aggregate (industry) equilibrium by *assuming* that technical change is of such a gradual and piecemeal nature at the aggregate level that competitive equilibrium prevails. The device of blurring the difference between knowledge as a factor of production and as a good employed by NEGT achieves exactly the same: it allows NEGT to ignore the cumulative nature of embodied technical progress.[2] The analytical ‘bridge’ between static and dynamic analysis is provided by the concept of ‘knowledge externalities’. Differently from Marshall, increasing returns are not the result of a proportional increase of all factors in a single industry, but of an economy-wide scale effect arising from the greater efficiency of given amounts of non-accumulable factors of production due to the larger use of a self-generating (accumulable) factor of production called knowledge. In neither case are Youngian *dynamic* increasing returns or, more precisely, the processes underlying accumulation, explained. As both Sraffa and Young pointed out, ‘[w]ith the extension of the division of labour, the representative firm, like the industry of which it is a part, loses its identity’ (Young 1928, 538; see also Sraffa 1930, 91–2). The analytical purpose of the ‘representative firm’ and of ‘knowledge externalities’ is that they admit precisely those changes and movements that are *a priori* compatible with the theory. The cumulative nature of embodied technical progress is not among these.

Young and Schumpeter took the view that static analysis was simply different from dynamic analysis and chose to concentrate on the latter. Sraffa refused to leave the terrain of the theory of value to the neoclassicals. The result has been the emergence of two different research agendas: Sraffa went on to provide a consistent, general and asymmetric theory of prices, distribution and reproduction that allows for an encompassing analysis of dynamics. Kaldor developed Young's suggestions for a theory of embodied technical progress and capitalist accumulation. Not surprisingly, it is Schumpeter who has been 'revived' by NEGT since, of these three authors, he remained the most ambivalent with regard to the compatibility of his theory of development with marginalist analysis. To understand contemporary growth dynamics, its roots in the changed social fabric of late capitalism as well as the changed material basis of technical progress, it is time to bring back together Sraffa, Young and Kaldor.

NOTES

1. See Sraffa 1926, 541, fn. 1, also the introductory pages to his 1928–31 lectures on the *Advanced Theory of Value*. The unpublished manuscript is held in the Wren Library, Trinity College, Cambridge (Sraffa Papers, file D2/4).
2. Aghion and Howitt (1998, 102) say as much in a footnote on their treatment of (physical) capital accumulation.

REFERENCES

Aghion, P. and P. Howitt (1992), 'A model of growth through creative destruction', *Econometrica*, **60** (2), 323–51.

Aghion, P. and P. Howitt (1998), *Endogenous Growth Theory*, Cambridge, MA: MIT Press.

Arrow, K.J. (1962), 'The economic implications of learning by doing', *Review of Economic Studies*, **29** (3), 155–73.

Chamberlin, E.H. (1950 [1933]), *The Theory of Monopolistic Competition: A Re-orientation of the Theory of Value*, Cambridge, MA: Harvard University Press.

Clapham, J.H. (1922), 'On empty boxes', *Economic Journal*, **32**, 305–14.

Frankel, M. (1962), 'The production function in allocation and growth: a synthesis', *American Economic Review*, **52**, 995–1002.

Grossman, G. and E. Helpman (1991), 'Quality ladders in the theory of growth', *Review of Economic Studies*, **106** (2), 43–61.

Keynes, J.M. (1930) (ed.), 'Increasing returns and the representative firm. A symposium (with contributions by D.H. Robertson, P. Sraffa and G.F. Shove)', *Economic Journal*, **40**, 79–116.

Knight, F.H. (1921), *Risk, Uncertainty and Profit*, Boston and New York: Houghton Mifflin.

Kurz, H. and N. Salvadori (1998), 'What is new in the "New" growth theory? Or: old wine in new goatskins', in F. Coricelli et al. (eds), *Growth and Development: Theories, Empirical Evidence and Policy Issues*, London: Macmillan.

Lucas, R. (1988), 'On the mechanics of economic development', *Journal of Monetary Economics*, **22** (1), 3–42.

Marglin, S. (1984), *Growth, Distribution and Prices*, Cambridge, MA: Harvard University Press.

Marshall, A. (1920), *Principles of Economics*, 8th edition, London: Macmillan.

Pasinetti, L. (2000), 'Critique of the neoclassical theory of growth and distribution', *Banca Nazionale del Lavoro Quarterly Review*, **53** (215), 383–433.

Pigou, A.C. (1922), 'Empty boxes: a reply', *Economic Journal*, **32**, 458–65.

Pigou, A.C. (1927), 'The laws of diminishing and increasing costs', *Economic Journal*, **37**, 188–97.

Pigou, A.C. (1928), 'An analysis of supply', *Economic Journal*, **38**, 238–57.

Rebelo, S. (1991), 'Long run policy analysis and long run growth', *Journal of Political Economy*, **99** (3), 500–21.

Rizvi, S.A.T. (1994), 'The microfoundations project in general equilibrium theory', *Cambridge Journal of Economics*, **18**, 357–77.

Robbins, L. (1928), 'The representative firm', *Economic Journal*, **38**, 387–404.

Robertson, D.H. (1924), 'Those empty boxes (with a comment by A.C. Pigou and rejoinder by D.H. Robertson)', *Economic Journal*, **34**, 16–31.

Robertson, D.H. (1930), 'The trees of the forest', *Economic Journal*, **40**, 80–89.

Romer, P. (1986), 'Increasing returns and long-run growth', *Journal of Political Economy*, **94** (5), 1002–37.

Romer, P. (1990), 'Endogenous technological change', *Journal of Political Economy*, **98** (5), 71–102.

Romer, P. (1994), 'The origins of endogenous growth', *Journal of Economic Perspectives*, **8** (1), 3–22.

Salter, W.E.G. (1960), *Productivity and Technical Change*, Cambridge: Cambridge University Press, 2nd edition, 1966.

Schumpeter, J.A. (1928), 'The instability of capitalism', *Economic Journal*, **38**, 361–86.

Schumpeter, J.A. (1994 [1954]), *A History of Economic Analysis*, London: Routledge.

Shove, G.F. (1928), 'Varying costs and marginal net products', *Economic Journal*, **38**, 258–66.

Shove, G.F. (1930), 'The representative firm and increasing returns', *Economic Journal*, **40**, 94–116.

Solow, R.M. (1994), 'Perspectives on growth theory', *Journal of Economic Perspectives*, **8** (1), 45–54.

Solow, R.M. (1997), *Learning from Learning by Doing. Lessons for Economic Growth*, Stanford, CA: Stanford University Press.

Solow, R.M. (2000), 'The neoclassical theory of growth and distribution', *Banca Nazionale del Lavoro Quarterly Review*, **53** (215), 277–328.

Sraffa, P. (1925), 'Sulle relazioni fra costo e quantità prodotta', *Annali di Economia*, **2**, 277–328. English translation by John Eatwell and Alessandro Roncaglia in L.L. Pasinetti (ed.) (1998), *Italian Economic* Papers, vol. 3, Bologna, Il Mulino and Oxford: Oxford University Press, pp. 323–63.

Sraffa, P. (1926), 'The laws of returns under competitive conditions', *Economic Journal*, **36**, 535–50.

Sraffa, P. (1930), 'A criticism', *Economic Journal*, **40**, 89–93.

Sraffa, P. (1960), *Production of Commodities by Means of Commodities. Prelude to a Critique of Economic Theory*, Cambridge: Cambridge University Press.

Sraffa, P. (1961), 'Comment', in F.A. Lutz and D.C. Hague (eds), *The Theory of Capital*, London: Macmillan, pp. 305–6.

Uzawa, H. (1965), 'Optimum technical change in an aggregate model of economic growth', *International Economic Review*, **6**, 18–31.

Young, A. (1928), 'Increasing returns and economic progress', *Economic Journal*, **38**, 527–42.

6. A dynamic framework for Keynesian theories of the business cycle and growth

Pedro Leão

1. INTRODUCTION

Keynesian explanations of the cycle are essentially based on the interaction between the multiplier and the accelerator (for a comprehensive exposition of different Keynesian theories of the cycle, see Sherman 1991). However, and as we shall see, those explanations fail to account for the *self-sustained* nature of the booms and recessions we observe in the real world – processes where output growth (decline) in one period *automatically* leads to output growth (decline) in the following period, and so forth.

In this chapter, we recast the multiplier–accelerator model of the cycle in a new dynamic framework. This is inspired by Harrod's theory of economic growth (see Harrod 1939, 1948).[1] The results are twofold. First, the resulting model provides a satisfactory explanation for the observed self-sustained nature of booms and recessions. Second, our dynamic framework suggests that at the root of booms and recessions may intriguingly lie the fact that *a change in investment has a greater effect on aggregate demand than on aggregate supply.*

The paper is organized as follows. Section 2 explains why we do not find the multiplier–accelerator account of booms satisfactory. In Section 3 we present our dynamic framework. Sections 4 and 6 apply this framework to show the forces that move output up and down along booms and recessions. In Sections 5 and 7 we argue that several factors – mentioned by the Keynesian cycle literature – may at some point in a boom or a recession invert the dynamics generated by our framework, and thereby account for the upper and lower turning-points of the cycle. Section 8 makes a review and concludes.

2. A CRITIQUE OF THE MULTIPLIER–ACCELERATOR DESCRIPTION OF BOOMS

The Keynesian description of booms is based on the multiplier–accelerator developed by Samuelson (1939). Starting from three equations – a consumption equation dependent on (lagged) income, an investment equation based on the accelerator, and an equation of aggregate equilibrium – this model arrived at a reduced form second-order difference equation for output.[2] The problem was that, for reasonable values of the parameters, the solution of that difference equation yielded an explosive cyclical path for output – instead of a neat cyclical movement. The conclusion was then that the multiplier–accelerator model could not tell the whole story about the cycle, – but that it could at least account for the self-sustained growth of output along economic expansions. Sherman (1991, 150–51) summarized the basic 'self-sustained growth argument': 'a small increase in investment leads through the multiplier . . . to a larger increase in national income. The increase in national income leads through the accelerator to a certain amount of new net investment. *The new investment leads to more national income, which leads to more investment, and so forth . . .*' (emphasis added)[3]

However, this argument may be criticized in two ways: (i) It is true that an increase in income leads through the accelerator to a certain amount of new investment. But there is no guarantee that new investment leads to more national income. If the new value of investment is lower than the preceding one, *income will fall*; and (ii) Moreover, even if national income rises there is no guarantee of a subsequent increase in investment. If national income rises but by less than it had been rising in the recent past, by the accelerator *investment will fall.*

In short, the output path resulting from the interaction between the multiplier and the accelerator is *indeterminate*. Needless to say, this conclusion is consistent with the previously cited Samuelson's mathematical solution of the multiplier–accelerator model – an explosive cyclical path for output, that is, larger and larger fluctuations of output along time.

3. THE DYNAMIC FRAMEWORK

3.1. Dynamic Equilibrium Between Aggregate Demand and Agregate Supply

Our dynamic framework stresses the dual character of investment – it affects both demand and capacity – as the key force of the business cycle. This is consistent with one of the most important stylized facts of the

Table 6.1 Cyclical amplitudes of GNP, M1 velocity and investment (%)

	Average, 4 cycles 1921–38		Average, 4 cycles 1949–70		Average, 3 cycles 1970–82		Average, 2 cycles 1982–01	
	Exp.	Cont.	Exp.	Cont.	Exp.	Cont.	Exp.	Cont.
GNP	21.2	− 16.4	17.9	− 1.5	12.1	− 3.5	37.5	− 0.9
M1 velocity	9.3	− 13.3	16.7	− 1.4	14.0	− 1.5	13.9	− 3.0
Gross Private Investment	55.4	− 49.3	23.5	− 9.5	29.8	− 28.0	70.0	− 5.3

Source: Leão (2005, table 2, p. 121).

cycle, – the fact that investment is the most variable component of aggregate output (see Table 6.1).

For simplicity, we assume a closed economy with no taxes, no government expenditure, no autonomous consumption, and interest rates set by the central bank.[4] In this setting, aggregate demand depends uniquely on the level of autonomous investment and on the multiplier:

$$Y^{d} = (1/s) \cdot I \tag{1}$$

where Y^{d} is aggregate demand, $1/s$ is the Keynesian multiplier (s is the marginal propensity to save) and I is autonomous investment.

Besides setting the level of demand, investment also increases the productive capacity of the economy (aggregate supply). If we assume production functions with fixed coefficients, abundant labour supply, no capital depreciation, and no lags in the effects of investment on capacity, then aggregate supply in one period depends only on the capital received from the previous period and on new investment:

$$Y^{s} = 1/v \cdot (K_0 + I) \tag{2}$$

where Y^{s} is aggregate supply, K_0 is the capital received from the previous period and $1/v$ is the productivity of capital (in the United States, $1/v$ is roughly equal to 1/3 (Sherman 1991, 139)).

What must happen if aggregate demand is to grow in line with aggregate supply along time? Using equations (1) and (2), Domar (1947) provided the answer:

$$\Delta Y^{d} = \Delta Y^{s}$$
$$(1/s) \cdot \Delta I = (1/v) \cdot I$$
$$\Delta I/I = s/v$$

Thus *s*/*v* is the growth rate of investment that guarantees the dynamic equilibrium between demand and supply. Furthermore, given the *proportionality* between investment and aggregate demand, when investment rises at *s*/*v* aggregate demand also grows at *s*/*v* – and the same happens with aggregate supply and output. Hence, *s*/*v* is called the required rate of economic growth (g^r).

3.2. The Principle of Instability I: A Change in Investment has a Greater Effect on Demand than on Supply

One important reason why the economy is unstable and we have booms and recessions may be the following curious fact. Since $1/s > 1$ in equation (1), an increase in investment leads to an *amplified* increase in aggregate demand. By contrast, in equation (2) $1/v < 1$ and therefore an increase in investment leads to a *lessened* increase in aggregate supply. Hence, we can say that an increase in investment has a greater effect on aggregate demand than on aggregate supply. In Figure 6.1, this implies that the slope of the aggregate demand curve is higher than that of the aggregate supply curve.

As we will see, this result may be at the root of instability. For example, if individual firms lack capacity and try to suppress it by raising investment, *they generate a greater increase in demand than in supply – and end up with an even larger deficit of capacity.* And, since this will cause a further rise in

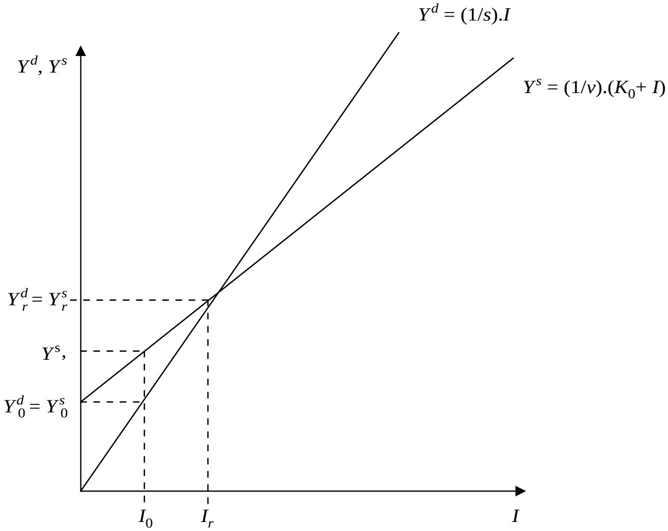


Figure 6.1 Changes in investment, aggregate supply and aggregate demand

investment, the system will diverge more and more from equilibrium with demand increasingly farther above supply.[5]

We now look closer at Figure 6.1 in order to explore further the relationship between investment, aggregate supply and aggregate demand. This figure assumes equilibrium between aggregate demand and supply in period 0, $Y^{d}_{0} = Y^{s}_{0}$. In the next period, the relation between aggregate demand and aggregate supply will depend on the level of investment. If investment remains constant at I_0, then aggregate demand will also stay constant at Y^{d}_{0} (see equation 1), but aggregate supply will rise to $Y^{s'}$ (because investment adds capacity). The result is that *a constant level of investment leads to excess capacity.*[6]

However, since the effect of investment on supply (the productivity of capital, $1/v$) is smaller than the effect of investment on demand (the multiplier, $1/s$), as investment increases from I_0 up to I_r the gap between capacity and demand gradually disappears. Excess capacity is eliminated if investment rises up to I_r. *Yet, if investment increases too much (beyond I_r) the paradoxical result is lack of capacity.*

As can be seen in Figure 6.1, there is a unique growth rate of investment, (I_r -I_0/I_0), that guarantees equilibrium between the pace of demand and the pace of supply. This is obviously Harrod–Domar's required rate of growth, s/v.

3.3. The Principle of Instability II: Expectations, Changes in Investment, and Booms and Recessions

Keynesian theories of the business cycle do not believe that 'supply creates its own demand' but rather the other way around: changes in demand are the prime movers of supply and output along the business cycle. These theories should therefore be able to address two questions: (i) What are the mechanisms that drive up demand and output in a self-sustained way along economic booms? (ii) Why do those mechanisms sometimes breakdown and aggregate demand and output become stagnant or start declining?

A major lesson of our framework is that *everything depends on the rate of change in investment – and thus on the expectations* of firms about demand growth. As we will see, if demand is expected to grow above the required rate ($g^{e} > g^{r}$), the resulting (big) increase in investment will generate an effective increase in demand that will be *self-sustained* – and will give rise to a cumulative expansion. However, if for some reason expected demand growth becomes lower than the required rate ($g^{e} < g^{r}$), the resulting (small) increase in investment will lead to an effective increase in demand that will *not* be sustained – and will give rise to a recession.

4. BOOMS: UPWARD INSTABILITY RESULTING FROM *ge* > *gw*

4.1. A Simple Model of Upward Instability

We start from equilibrium, at $Y_0^d = Y_0^s$ in Figure 6.2.[7] If we suppose that demand is expected to grow *above* the required rate, from Y_0^d to $(Y_1^d)^e$, then, by the accelerator, investment rises from I_0 to I_1. This leads, through the multiplier, to an increase in aggregate demand to Y_1^d – *which is higher than the initially expected increase in demand* and higher than the actual increase in supply. Hence, initial expectations are surpassed and there is an undesired increase in the rate of capacity utilization.[8]

Assuming the simplest form of adaptative expectations, $g^e = g^{-1}$, the expected growth in demand will be *revised upwards* moving even farther above the required rate. In Figure 6.3, demand will now be expected to grow from Y^d_1 to (say) $(Y^d_2)^e$ and, by the accelerator, there will be a rise in investment, from I_1 to I_2. Through the multiplier, this will then generate an increase in demand from Y^d_1 to Y^d_2, again larger than initially expected and again leading to an undesired increase in the rate of capacity utilization.

In short, when we start with an expected growth rate greater than the required rate, the result may be increasing rates of capacity utilization and self-sustained growth.[9] As we will see, this requires that the actual growth path of output is not obstructed by lack of labour supply.

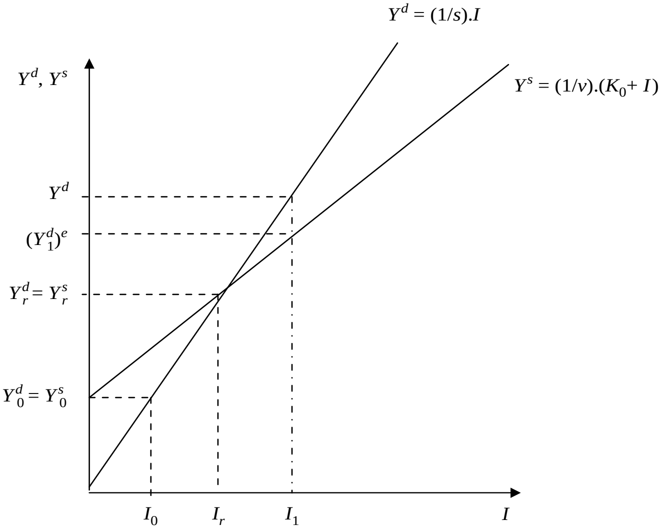


Figure 6.2 Upward instability

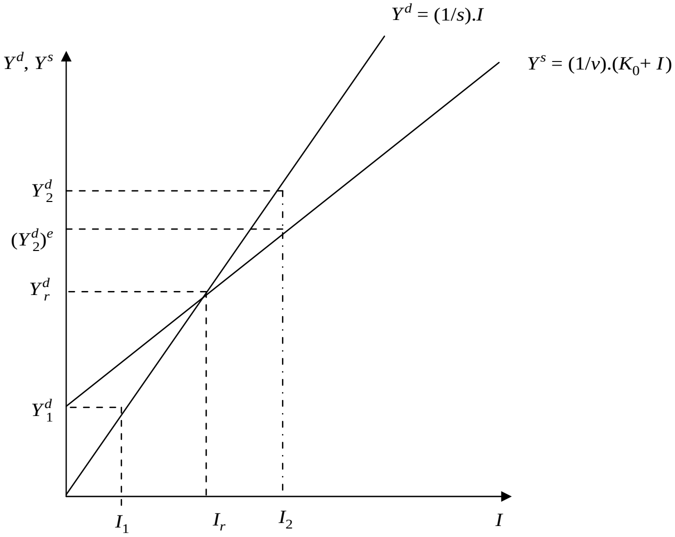


Figure 6.3 Upward instability (continued)

Finally, note that in our dynamic framework it is possible to recast the multiplier–accelerator description of the self-sustained process of expansion avoiding the two criticisms made in the previous section. This will be achieved if we modify the cited argument (Sherman 1991, 150–51) in the following way: 'a small increase in investment leads through the multiplier . . . to a larger increase in national income. The increase in national income leads through the accelerator to a certain amount of new net investment. [*If the increase in income was larger than the required rate*] the new investment leads to [an increase in] national income [higher than the required rate], which leads to [a new increase in] investment, and so forth . . .'

4.2. Checks on Upward Instability

The upward instability along booms we have just described depends on two consecutive mechanisms – the response of investment to expected increases in demand [$I^N = v.(\Delta Y^d)^e$], and the bigger effect on demand than on supply of changes in investment ($1/s > 1/v$). Both mechanisms – and thus the upward instability – are less strong than our simple model suggests.

First, investment is in practice less sensitive to growth expectations than the simple accelerator suggests. This happens because of two reasons. On the one hand, firms do not instantaneously adjust their capital stock to short-run changes in demand, but instead make gradual adjustments over

Table 6.2 Cycle relatives, average of four cycles, 1949–1970

Series/stages of the cycle	Expansion				**Peak**	Contraction			
	1	2	3	4	**5**	6	7	8	**9**
Capacity utilization, manufacturing (82)	**91.5**	97	103.3	104.1	**104.1**	101.3	97.4	92.9	**90.8**
Consumption/ national income (231/220)	**103.2**	101.2	98.5	98.6	**99.4**	100.2	101.2	102	**102.4**
Employee compensation/ national income (64)	**100.1**	98.4	98.7	100.8	**102.3**	102.9	103.1	102.9	**102.9**
Hourly wages, real (346)	**92.4**	94.7	99	103.1	**105.2**	105.4	105.7	106.2	**106.3**
Product per hour (358)	**93.4**	96.1	99.9	102.3	**103.3**	103.1	103.5	104.3	**104.5**
Prime interest rates charged by banks (109)	**85.1**	86.2	94.4	110.4	**123.8**	123.7	119.3	114.1	**108.3**

Notes: The cycle relatives are the original data for a variable divided by the average of that variable over the whole cycle. The division of each cycle in nine stages is that of Mitchell. Stage 1 is the initial trough of the cycle, stage 5 is the cycle peak and, finally, stage 9 is the final trough. By definition, each of these three stages is one quarter long. The expansion period (excluding stages 1 and 5) is then divided into three equal time periods, called stages 2, 3 and 4. Similarly, the contraction period (excluding stages 5 and 9) is divided up into three periods of equal length, called stages 6, 7 and 8. For a detailed discussion of the methodology, see Sherman (1991, pp. 11–13).

Source: Sherman (1991, appendix D). NBER cycles.

several years (Jorgenson 1971). On the other hand, there are significant components of investment which are largely independent of short-run changes in demand – namely, replacement investment, investments governed by long-term expectations (for example, a hydroelectric dam), and investments related to technical innovations.

Second, not only do expectations about demand growth lead to smaller increases in investment than the simple accelerator suggests, but also these very increases in investment lead to *lower* increases in actual demand than implied by our simple theory of demand, $Y^d = (1/s).I$. Why? One reason is that the marginal propensity to consume and the multiplier tend to decline along business expansions.[10] A second reason is that there are

Table 6.3 Cycle relatives, average of five cycles, 1970–2001

Series/stages of the cycle	Expansion				**Peak**	Contraction			
	1	2	3	4	**5**	6	7	8	**9**
Capacity utilization, manufacturing (82)	**95.2**	98.1	102.8	106.1	**105.1**	102.3	98.6	95.6	**91.8**
Consumption/ national income (231/220)	**101**	100.2	99.4	98.4	**98.3**	99	100.8	101.9	**103.3**
Employee compensation/ national income (64)	**101.1**	99.8	99.3	99	**98.9**	100	101.6	101.8	**101.9**
Hourly wages, real (346)	**97.8**	98.7	100.5	100.9	**99.4**	99.1	99.6	99.6	**100**
Product per hour (358)	**96.6**	98.8	100.7	101.4	**100.9**	100.3	99.6	99.7	**100**
Prime interest rates charged by banks (109)	**87.5**	84.9	87	111.3	**142.9**	133.7	141.9	119.8	**105.2**
Raw materials, Producer Price index (331)	**83.3**	87.8	92	107.1	**117**	118.8	115.4	121.6	**118**
Consumer Price Index (320)	**87.2**	90.2	94.9	102	**109**	110.5	114	117.4	**118.2**

Source: Calculations based on Sherman (1991, Appendix E) and on Federal Reserve Data (FRED). NBER cycles.

other types of autonomous expenditure apart from investment, namely autonomous consumption (*C*) and government expenditure (*G*) and autonomous net exports (*X*-*M*). Hence, aggregate demand is better described by $Y^d = m.(C + I + G + X\text{-}M)$, and increases in investment lead to increases in demand through a multiplier, *m*, much smaller than 1/*s*.

5. THE UPPER TURNING-POINT

We have already seen that $g^e > g^r$ leads to a self-sustained expansion of aggregate output. However, if g^e ever happens to be lower than g^r it is likely that the boom will stop and be followed by a contraction.

Assume an output expansion is going on and that in period t output is given by $Y^{d}_{t} = Y^{s}_{t}$ in Figure 6.4. Now suppose that, for some reason, g^{e} falls below g^{r}. With demand now expected to grow only from Y^{d}_{t} to $(Y^{d})^{e}$, the accelerator investment rises from I_{t} to I_{t+1}. Then, through the multiplier, actual demand increases to only Y^{d}_{t+1}, lagging behind expected demand – $(Y^{d})^{e}$ – and actual supply (Y^{s}_{a}).

More importantly, expectations about demand growth will be revised downwards. Assuming once more $g^{e}=g^{-1}$, *demand will now be expected to grow at a lower rate and, as a result of the accelerator, investment will suffer a rapid contraction.* This marks the end of the expansion and the beginning of the contraction, for, through the multiplier, the fall in investment leads to a reduction in aggregate output. In conclusion, the upper turning-point is brought about by a reduction in the expected rate of demand growth to a value below the required rate.[11]

But why should g^{e} fall below g^{r}? In the remainder of this section, we present various reasons why that may happen at some point during an expansion, thereby initiating a contraction of output.

In so doing, we integrate into our framework several explanations given in the Keynesian cycle literature for the onset of economic crises. We start with the explanations based on the ceiling hypothesis: (i) the ceiling of full-employment of labour (Hicks 1956); and (ii) the inflation-barrier (Robinson 1962). Afterwards, we move on to arguments that dispense with

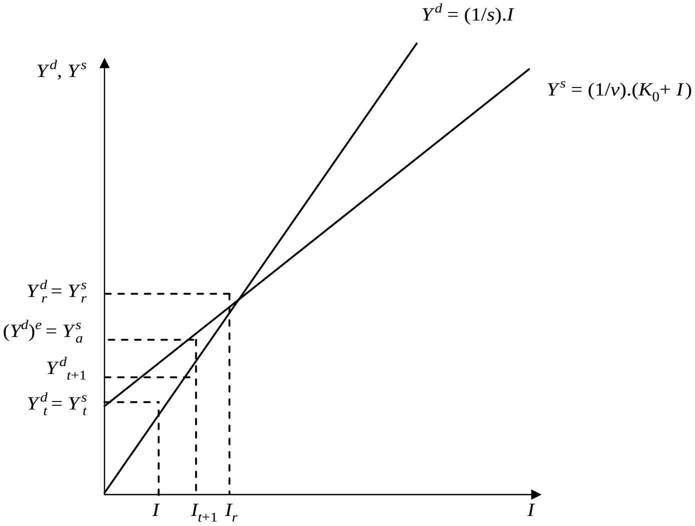


Figure 6.4 The upper turning-point

that hypothesis, namely (iii) the decline in the propensity to consume along expansions (Hobson 1922) and (iv) the saturation of important innovations (Schumpeter 1934) or the burst of speculative bubbles.

5.1. Upper Physical Limits

The ceiling of full-employment of labour

If the actual growth rate of output during an expansion is higher than the natural rate ($g > g^n$), sooner or later output will reach the physical ceiling of full employment (g^n is given by the sum of the labour population growth rate and the pace of labour-saving technological progress). If at the same time $g^r > g^n$, our framework allows us to draw the following surprising conclusion. When full-employment is reached, *the result cannot not be a mere slowdown in the actual rate of growth to the natural rate along the ceiling. Instead, output is pushed into a downward movement* (see Figure 6.5).

Why? Suppose an output expansion is going on, and that in period *t* we have $Y^d_t = Y^s_t$ in Figure 6.6. Assume also that demand is expected to keep on growing at the actual rate, close to the required rate, and that this is higher than the natural rate (that is, $g^e = g \approx g^r > g^n$). Given the expected increase in demand from Y^d_t to Y^d_r, by the accelerator, firms would like to raise investment from I_t to I_r. However, as investment rises from I_t up to only I_{t+1}, output increases by g^n, from Y_t to Y_{t+1}, reaching an upper

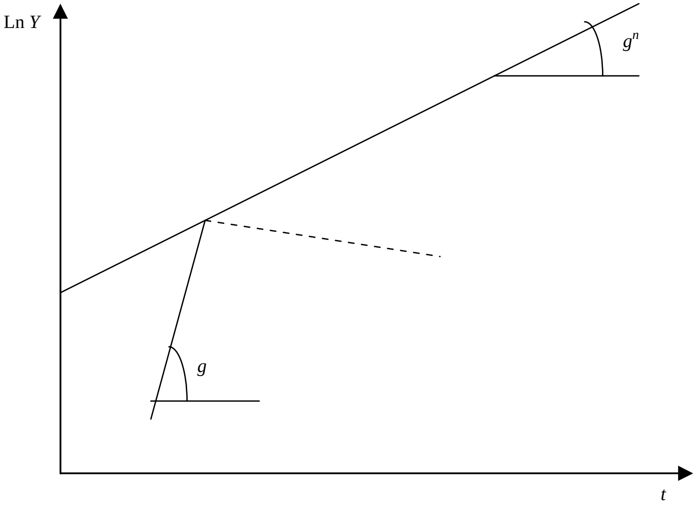


Figure 6.5 Actual growth vs natural growth

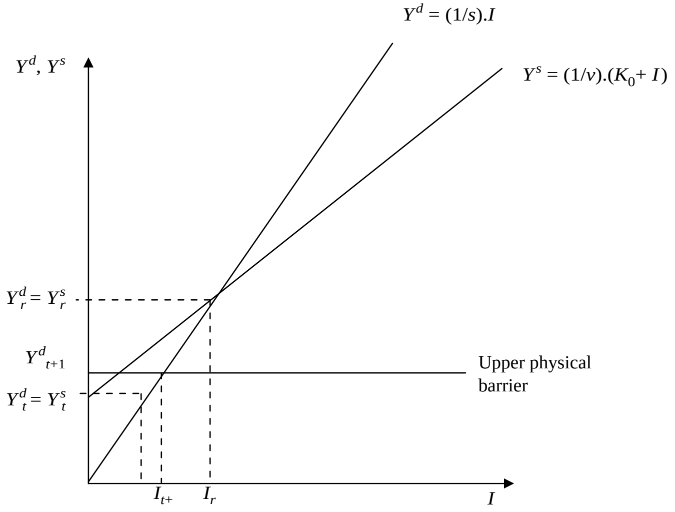


Figure 6.6 The full employment ceiling hypothesis about the upper turning-point

physical barrier (full employment of labour). From that moment on, neither investment nor output can rise any further.

Thus, instead of rising from Y_t to Y_r, *output slows down its growth to a rate below the required rate*, by growing from Y_t to only Y_{t+1}. This, in turn, leads to a corresponding downward revision of expectations.

The slowdown of expectations about demand growth *below the required rate* subsequently pushes investment down by the accelerator. Through the multiplier, this then causes a contraction in demand and output.[12]

The inflation barrier

If, during a boom, the actual growth rate of output exceeds the natural rate and, after a certain point, the economy comes close to full employment, nominal wages may start rising more than productivity. Raw materials prices tend to rise in these times as well. In these circumstances, production costs rise and feed into higher prices of final goods. Fearing a wage–price spiral, the central bank is then likely to raise the interest rate.[13]

Higher raw materials prices and interest rates will subsequently have a negative effect on demand and profits. Why? Consider first the effect of higher raw materials prices on demand. Since most raw materials are bought in foreign markets, the rise in their prices generates a flow of income – and thus of monetary demand – out of the United States. Hence, at the microeconomic level, the downward sloping demand curves (faced by producers operating in monopolistic or imperfectly competitive markets) do not expand *in proportion* with the costs of raw materials. In this context: (i) If producers are (continuous) profit maximizers, they will not *fully* incorporate the increase in raw materials costs into the prices of their finished products , in order to avoid an excessive fall in their demand; and (ii) If producers instead use some variant of the full-cost pricing principle, the rise in raw materials costs will be fully passed into prices, but demand and profits will decline by more.

The main point, however, is that in both cases *profits and demand will be hit by the increase in raw materials prices.*

The increase in interest rates leads to similar results. The rise in interest rates transfers income from debtors to savers, and thus has a negative impact on aggregate demand. Hence, the demand curves for firms' products do not expand in line with their rising interest costs. As a result, profits and demand will be negatively affected.[14]

How does the negative effect of rising raw materials prices and interest costs on demand and profits lead from prosperity to depression? Suppose an output expansion is going on at the required rate, and that in period t we have $Y^d_t = Y^s_t$ in Figure 6.4. By the accelerator, the slowdown in the growth of demand, say from Y_t to $(Y^d)^e$, would reduce the increase in investment to I_{t+1}. *If we add the fact that profits are also being hurt*, an even greater slowdown in investment is likely.[15] The result will be a contraction in output through the mechanisms already described.

5.2. Non-Ceiling Hypothesis About the Upper Turning-Point

As is well known, the major problem with ceiling explanations of the upper turning-point is that economic activity at the peak is often below full-capacity. In what follows we present alternative explanations for the upper turning-point.

A decline in the propensity to consume

Suppose an output expansion is again going on at $Y^d_t = Y^s_t$ in Figure 6.7 and that demand is expected to grow from Y^d_t to Y^d_r. According to the accelerator, firms raise investment from I_t to I_r. However, if the propensity to consume falls significantly – as often happens during booms – that increase in investment will operate through a lower multiplier and, as a result, the growth of demand will slow down from Y_t to, say, only Y_{t+1} (instead of Y_r). This slowdown in the growth rate of demand to a rate below the required rate ends the boom through the mechanisms that have been described before.

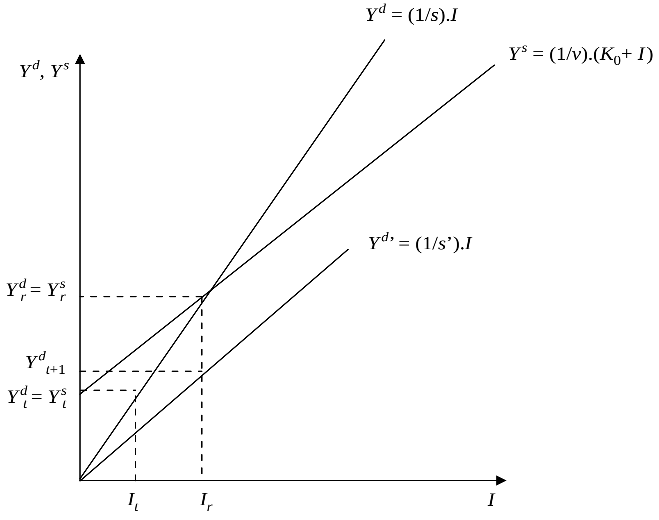


Figure 6.7 The non-ceiling hypothesis about the upper turning-point

The saturation of important innovations and the burst of speculative bubbles

If the market for an important innovation eventually becomes saturated, a sudden collapse of investment in the related sectors will ensue. On the other hand, if a stock-market or a real-estate bubble – which often develop from the optimism that goes along with economic booms – happens to burst, the resulting negative effects on both wealth and expectations may lead to a slowdown in consumption and investment demand. How does this lead from prosperity to depression?

Once more, suppose an output expansion is going on at $Y^{d}_{t} = Y^{s}_{t}$ in Figure 6.6 and that demand is expected to grow from Y^{d}_{t} to Y^{d}_{r}. The accelerator says that firms will raise investment from I_t to I_r; however, if an important innovation reaches saturation or a speculative bubble bursts, investment is better seen as slowing down from I_t to, say, I_{t+1} (instead of I_r). This will then trigger the contraction in the usual way.

6. A BRIEF COMMENT ON THE NATURE OF CONTRACTIONS

As we have seen, the upper turning-point is the result of a decline in investment which leads, by the multiplier, to a fall in output. This reduction in output subsequently causes, by the accelerator, a rapid collapse in the types of investment that are influenced by short-run changes in demand. However, after that collapse the contraction dynamics based on the interaction between the accelerator and the multiplier comes to a *halt*: investment stops falling (because the types of investment governed by short-run changes in demand are already nil), and hence the multiplier ceases to dictate further declines in output.

It logically follows that the periods of falling output (crises) tend to be relatively short-lived, and rapidly followed by periods characterized by autonomous expenditure and output stagnated at relatively low levels (depressions). This conclusion is supported by the empirical evidence (Sherman 1991, 14–15).

7. A BRIEF COMMENT ON THE REVIVAL

The revival usually coincides and is explained by a resumption of investment from the low levels it attains during depressions. This increase in investment may in turn be initiated by four different facts: (i) the eventual emergence (and bunching in time) of important innovations; (ii) the reappearance of replacement investment due to the fact that, as time goes by, depreciation eventually reduces the capital stock to below the level required to meet existing demand; (iii) enhanced prospects for the profitability of investment due to decreasing production costs (falling interest rates and raw-materials prices) during contractions; and (iv) finally, the increase in the propensity to consume and in aggregate demand due to the rise in the share of wages in national income.[16]

Irrespective of its cause, if the revival of investment is greater than the required rate an output expansion along the lines already described will ensue.

8. REVIEW AND CONCLUSION

In our dynamic framework (recall Figure 6.1), if investment remains constant from one period to the next, then aggregate demand will also stay constant, but aggregate supply will rise because investment adds capacity. Hence, a constant level of investment leads to excess capacity. However, since a change in investment has a bigger effect on demand than on supply, up until a certain point an increase in investment reduces the excess of capacity; beyond that point an increase in investment leads to lack of capacity.

In this setting, if firms expect only a 'small' increase in demand ($g^e < g^r$), investment will not increase enough. It will generate an increase in effective demand smaller than initially expected, and smaller than the actual increase in capacity. As a result, the rate of capacity utilization will fall and the expected increase in demand will become even smaller. This will lead, in turn, to a rapid fall in the types of investment that are influenced by short-run changes in demand, and to corresponding falls in output: an economic crises.

By contrast, if firms expect a 'big' increase in demand ($g^e > g^r$), investment will rise too much. It will generate an increase in effective demand bigger than initially expected, and bigger than the actual increase in capacity. Hence, the level of capacity utilization will rise and the expected increase in demand will become even bigger. This will lead again to a big increase in investment, so that a self-sustained process of expansion will ensue.

NOTES

1. This is the main contribution of the chapter. We shall argue that Harrod's dynamic system, which has been usually applied to the theory of growth, should instead be used as the *basic framework* in the analysis of the business cycle.
2. The accelerator essentially says that, if firms are producing at the normal rate of capacity utilization, then an expected increase in aggregate demand will be met by additions in capital (net investment):

$$I^N = v(\Delta Y^d)^e \tag{3}$$

where I^N is net investment and v, the accelerator coefficient, is the amount of new capital required to produce one extra unit of output.

As is well known, the fact that net investment is a function of the *change* rather than the *level* of demand has the following implication. *An increase in demand does not necessarily imply an increase in net investment*: (i) If demand rises but by less than it had risen in previous periods, then net investment *falls off*; (ii) There will only be an increase in net investment if demand growth *accelerates*; and (iii) Note finally that positive net investment presupposes demand growth: a stagnant demand requires no additional capital and thus leads to zero net investment.

3. For an application of this argument to the profit squeeze and underconsumption models of the cycle, see Sherman (1991, 203, 256–57).
4. The latter implies that the money stock and/or its velocity of circulation adjust endogenously to changes in aggregate demand and output. For reasons why these endogenous adjustments are likely to occur see Kaldor (1986), in the case of money, and Pollin (1991) and Leão (2005), in the case of velocity. Evidence on the variability of velocity along the cycle is presented in Table 6.1.
5. By contrast, if a change in investment had a greater effect on supply than on demand (that is, $1/v > 1/s$) this instability would not exist. In that case, if firms lacked capacity and tried to suppress it by raising investment, supply would rise more than demand and the lack of capacity would sooner or later disappear.

6. Joan Robinson (1969, 92) also arrived at this result: ′the tragedy of investment is that it can never remain at a constant level. For . . . the level of demand for goods will be the same. But all the time capital is accumulating . . . the rate of profit consequently falls off . . . and new investment will appear less attractive to entrepreneurs'.
7. In the remainder of the paper, we will assume that aggregate supply corresponds to the level of output obtained with a certain desired rate of capacity utilization (for example, 80% of full capacity), not with full capacity.

   Note also that our model uses the accelerator as the central explanation of the level of investment. However, we are aware that the investment process is very complex and that the accelerator has many limitations (for a detailed discussion, see Sherman 1991, 141–42). In particular, the accelerator is more incomplete than a function that makes investment dependent on *changes in profit rates*, as proposed by Sherman (1991, 251). This is because the accelerator ignores the effects on investment of changes in costs of raw materials, finance, taxes and labour. Hence, we will use the accelerator carefully, and complete it whenever changes in costs are not fully incorporated into prices and lead to changes in profits and investment.
8. Why do we have this surprising result? The reason is that ′too optimistic′ expectations about demand growth ($g^e > g^r$) lead to an increase in investment such that – since its effect on capacity is lower than on demand ($1/v < 1/s$) – firms end up with supply adjusting to expected demand, but lagging behind actual demand.
9. For empirical evidence on rates of capacity utilization along booms, see Tables 6.2 and 6.3.
10. The decline of the propensity to consume along booms is often attributed to the fall in the share of wages in national income that usually takes place in these periods. For empirical evidence, see Tables 6.2 and 6.3.
11. Note that this argument is different from that of the simple multiplier–accelerator model. Without fully endorsing it, Sherman (1991, 141) summarizes the multiplier–accelerator explanation for the upper turning-point: ′Suppose that demand grows rapidly which causes a certain level of investment. If the rate of growth of demand slows, then the level of investment must decline. But an actual decline in investment will cause a recession . . . Thus *a theory may explain business cycle downturns . . . merely by showing why aggregate demand will slow its growth*, and it is not necessary to prove that aggregate demand will decline before investment declines (emphasis added).'

    The difference *is* that, according to our argument, the explanation of downturns requires not merely a slowing down of demand growth – but a slowing down of demand growth to a level *below the required rate.*
12. In this setting, we may speculate that the unusually long US expansion of the 1990s may have been permitted by the fact that technological innovations and population growth have created a natural rate of output growth as high as the actual rate ($g^n \approx g$) – thereby preventing output from ever reaching the full-employment ceiling.
13. The referred movements of nominal wages, productivity, raw materials prices, prices of finished goods and interest rates occurred frequently during the expansions of the 20th century (see Tables 6.2 and 6.3).

    A notable exception was the behaviour of raw materials prices during the expansions of the 1950s and 1960s when they remained unchanged (Sherman 1991, 211).
14. This argument is especially relevant because, towards the end of expansions, firms' costs become particularly sensitive to increases in interest rates, – especially in short term rates. This happens because along booms (i) firms' debt/equity ratio rises; (ii) the periods allowed for debt repayment become shorter; and (iii) the ratio of liquid assets to short-run debt falls (see Wolfson 1986).
15. The effect of profits is (partly) distinct from the effect of demand on investment. Indeed, ‘profits influence investment not only by providing the motive for it [like demand] but also through providing the means. An important part of investment is financed out of retained profits. Moreover, the amount that a company puts up of its own finance influences the amount it can borrow from outside'. (Robinson 1962, 86).

16. For evidence on these movements of raw materials prices, interest rates, the share of wages and the propensity to consume along the contractions of the 20th century, see Tables 6.2 and 6.3.

REFERENCES

Domar, E. (1947), 'Expansion and Employment', *American Economic Review*, **37**, 34–55.

Harrod, R.F. (1939), 'An Essay in Dynamic Theory', *Economic Journal*, **49**, March.

Harrod, R.F. (1948), *Towards a Dynamic Economics*, London: Macmillan.

Hicks, J.R. (1956), *A Contribution to the Theory of the Trade Cycle*, 3rd edn, Oxford: Oxford University Press.

Hobson, J. (1922), *The Economics of Unemployment*, London: George Allen and Unwin.

Jorgenson, D. (1971), 'Econometric Studies of Investment Behaviour: a Survey', *Journal of Economic Literature*, **19**, 1111–47.

Kaldor, N. (1986), *The Scourge of Monetarism*, 2nd edn, New York: Oxford University Press.

Leão, P. (2005), 'Why Does the Velocity of Money Move Pro-cyclically?', *International Review of Applied Economics*, **19**(1), 119–35.

Pollin, R. (1991), 'Two Theories of Money Supply Endogeneity: Some Empirical Evidence', *Journal of Post Keynesian Economics*, **13**(3), 366–96.

Robinson, J. (1962), *Essays in the Theory of Economic Growth*, London: Macmillan.

Robinson, J. (1969), *Introduction to the Theory of Employment*, 2nd edn, London: Macmillan.

Samuelson, P. (1939), 'Interactions between the Multiplier Analysis and the Principle of Acceleration', *Review of Economics and Statistics*, May, **21**, 75–8.

Schumpeter, J.A. (1934), *The Theory of Economic Development*, Cambridge, MA: Harvard University Press.

Sherman, H.J. (1991), *The Business Cycle: Growth and Crises under Capitalism*, Princeton, NJ: Princeton University Press.

Sherman, H.J. and D. Kolk (1996), *Business Cycles and Forecasting*, New York: HarperCollins.

Wolfson, M. (1986), *Financial Crises*, New York: M.E. Sharpe.

7. A Keynesian model of unemployment and growth: theory

John Cornwall*

A. INTRODUCTION

The main objective of the paper is to extend the basic model of Keynes's *General Theory* to explain medium- and long-run economic performance in developed capitalist economies. In this way we seek to deepen our understanding of the macroeconomic processes that account for differences in macro performance over time and between economies at similar stages of their economic development. It naturally starts from a conviction that Keynes's original model of short-run economic fluctuations is the appropriate foundation for further macroeconomic research. In the process of extending the traditional Keynesian model we frequently compare our views with those of the mainstream macroeconomic theory.

The natural candidate for comparison is an extended version of New Keynesian macroeconomics, now popular in the textbooks (Blanchard and Melino, 1999; Ragan and Lipsey, 2005; Mankiw and Scarth, 2001). In the 1980s, it started as a well-received microeconomic research programme, whose main goal was to replace perfectly competitive neoclassical markets with imperfectly competitive Keynesian ones. This was followed by the incorporation of non-accelerating inflation rate of unemployment (NAIRU) analysis for modelling the short-run and eventually by the adoption of a long-run growth analysis, although this was not integrated with the short-run analysis. These subsequent developments, together with the earlier micro analysis, we can designate as mainstream macroeconomic theory, to be compared with our extended Keynesian macro theory.

To better grasp the essential differences between the two views, the theoretical models are presented in the form of a limited number of propositions, or a core, achieved by selecting from each of their theoretical

* The author is indebted to the Office of the Vice-President Academic & Provost, the Dean of the Faculty of Science, and the Department of Economics, all of Dalhousie University, for providing financial support for this research. Thanks are due to Mark Setterfield and Wendy Cornwall for helpful comment and criticism.

frameworks those parts which make it 'usable'. For example, the selection from our extended Keynesian framework includes propositions that explain the broad historical tendencies of macroeconomic development, the 'stylized facts' in Kaldor's terminology. In addition, our core theory includes the propositions that provide a coherent theoretical basis to derive useful policy principles.[1] A similar process is used to select the mainstream core. Before undertaking these tasks, the next section takes up some of the broad historical tendencies we expect any macroeconomic core theory to be capable of explaining. Given the casual disregard for the historical record in so much of current macroeconomic theory, a consistency with the historical record is stressed throughout the paper. Following a background discussion of historical developments, in Section C we initiate the comparative discussion beginning with the alternative short-run theories.

B. THE STYLIZED FACTS

Our study deals with the advanced capitalist economies in the period following World War II until the end of the century for which comparable data are available. Table 7.1 summarizes short-run and longer-run macroeconomic developments of eighteen advanced capitalist economies from 1960 until the end of the century, the 1990s being the last short-run period for which comparable data are available.[2] The total period divides into five short cycles in GDP, with common end points for each cycle indicated in the top row of the table. Rates of unemployment, inflation and labour productivity growth are given for each of the countries in our sample during each short-run GDP cycle period. A noticeable feature in the table is the strong negative correlation between unemployment and productivity growth rates from one GDP cycle to the next in all economies. Considering the entire post World War II period, the countries separate into three groups, those with low rates of unemployment throughout the entire period with the exception of one short-run period at the end of the century (the low unemployment countries); those with relatively high rates of unemployment throughout the entire period (the high unemployment countries); and a third group in which low unemployment until the mid-1970s was followed by high unemployment thereafter (the low–high unemployment countries).

The data also show two lengthy episodes distinguished by their differing performance: the Golden Age of low unemployment and rapid growth (1960–73), followed by what we designate the Age of Decline (1974–2000), an episode of high unemployment, slow growth and high rates of inflation in the 1970s. The first episode includes the first two short-run GDP cycles and the second includes the last three successive within-episode cycles. The overall average rate of unemployment for the 18 economies during the Golden Age was 2.3 per cent. Adopting 3 per cent as the full employment rate of unemployment (although choosing a rate twice the actual average rate would not alter the conclusions), Table 7.1 brings out clearly the exceptional unemployment and inflation records in the first episode for all but four of the economies, the high unemployment group. This was also a period of rapid expansion of the welfare state, especially outside the United States, leading to a marked reduction in poverty rates and the inequality in income distributions.

The period from the mid-1970s to the end of the century was one of economic decline. Accelerating inflation rates in the late 1960s and early 1970s were followed by persistent inflation problems in spite of restrictive aggregate demand policies and rising unemployment rates. From the first short-run cycle in the table to that of the 1990s short-run cycle, the overall average unemployment rate rose approximately 350 per cent. Within these average figures there were exceptions, as most of the time the five low unemployment countries continued to experience low unemployment rates. In general, rates of growth of labour productivity traced out a rapid growth-slow growth pattern from the Golden Age to the Age of Decline.

Such a marked historical shift in macro performance from the Golden Age to the Age of Decline did not go unnoticed among macroeconomists (for example, Blanchard, 1997a, 1997b; Solow, 1997, 2000; and Cornwall and Cornwall, Chapters 10 and 11, 2001). Nevertheless its prominence failed to induce mainstream macro theorists to include the medium run as a third analytical category when modelling the post World War II era.[3] Two additional stylized facts also stand out. From one short-run period to the next, in every economy there was a strong negative correlation between unemployment rates and the growth rates of productivity. Second, the capitalist economies were spared a repeat of the Great Depression. While there was a noticeable increase in unemployment rates from the first medium-run episode to the second, macroeconomic performance after the mid-1970s did not indicate a major disconnect between forces driving the rates of growth of aggregate demand and aggregate supply. Finally, changes in the average unemployment rate from one short-run period to the next indicate mild cyclical movements. Most of the remainder of the paper discusses the short-run, medium-run and long-run models we use to explain these stylized facts and compares our explanations with those of the mainstream core.

Table 7.1 Annual average unemployment rates (U)[a], rates of consumer price inflation ($\dot{p}$)[b] and productivity growth rates ($\dot{q}$)[b] for 18 countries (%)

Years	1960–1967			1968–1973			1974–1979			1980–89			1990–2000		
	U	$\dot{p}$	$\dot{q}$	U	$\dot{p}$	$\dot{q}$	U	$\dot{p}$	$\dot{q}$	U	$\dot{p}$	$\dot{q}$	U	$\dot{p}$	$\dot{q}$
Low unemployment															
Austria	2.0	3.6	4.9	1.8	5.2	5.1	1.8	6.2	2.7	3.3	3.8	1.2	3.9	2.4	1.4
Japan	1.4	5.7	8.5	1.2	7.1	7.5	1.9	9.9	2.8	2.5	2.5	2.7	3.2	1.0	1.3
Norway	2.0	3.9	3.8	1.7	6.9	2.3	1.8	8.7	2.5	2.8	8.3	1.9	4.8	2.5	2.3
Sweden	1.6	3.8	3.9	2.2	6.0	2.9	1.9	9.8	0.5	2.6	7.9	1.6	7.1	3.3	2.4
Switzerland	0.0	3.4	2.9	0.0	5.6	3.0	0.4	4.0	0.6	0.6	3.3	0.3	3.1	2.3	0.6
Unweighted Average	1.4	4.1	4.8	1.4	6.2	4.2	1.6	7.7	1.8	2.4	5.2	1.5	4.4	2.3	1.6
High unemployment															
Canada	4.8	2.4	2.6	5.4	4.6	2.5	7.2	9.2	0.6	9.4	6.5	0.9	9.3	2.2	1.3
Ireland	4.9	4.0	4.1	5.6	8.9	4.6	7.9	14.9	3.4	14.3	9.2	3.6	11.3	2.6	3.5
Italy	4.8	4.0	6.3	5.7	5.8	4.9	6.6	16.1	2.6	8.0	11.1	2.0	10.6	4.0	1.7
United States	4.9	2.0	2.6	4.6	5.0	1.0	6.8	8.5	0.5	7.3	5.5	1.2	5.6	3.0	1.9
Unweighted Average	4.9	3.1	3.9	5.3	6.1	3.3	7.1	12.2	1.8	9.8	8.1	1.9	9.2	3.0	2.1
Low-high															
Australia	2.2	2.2	2.7	2.0	5.6	2.2	5.1	12.2	1.8	7.5	8.4	1.0	8.4	2.7	1.8
Belgium	2.0	2.8	3.9	2.5	4.9	4.9	7.1	8.4	2.4	9.8	4.9	2.2	8.5	2.2	1.6
Denmark	1.6	6.2	3.0	1.0	6.3	3.0	6.1	10.8	1.2	8.1	6.9	0.7	7.1	2.2	1.9
Finland	1.6	5.6	4.0	2.6	5.8	5.8	5.1	12.6	1.7	5.4	7.1	2.6	11.7	2.3	2.6
France	1.6	3.6	4.9	2.6	6.1	4.3	4.5	10.7	2.4	8.8	7.3	2.2	11.1	1.9	1.3
Germany	0.6	2.7	4.1	1.0	4.6	4.0	3.2	4.6	2.7	5.8	2.9	1.5	7.7	2.5	1.6

Netherlands	1.0	3.6	3.7	1.5	6.9	4.4	5.4	7.2	2.1	7.9	2.8	–0.3	5.4	2.5	0.8
New Zealand	0.1	3.3	1.2	0.3	7.4	2.7	0.8	13.8	–1.6	4.6	11.8	0.5	7.8	2.1	0.8
United Kingdom	2.7	3.6	2.7	3.3	7.5	3.2	4.7	15.6	1.2	9.8	7.4	1.9	8.0	3.6	1.8
Unweighted average	1.5	3.7	3.4	1.9	6.1	3.8	4.7	10.7	1.5	7.5	6.6	1.4	8.4	2.4	1.6
Overall Average	2.2	3.7	3.9	2.5	6.1	3.8	4.4	10.2	1.7	6.6	6.5	1.5	7.5	2.5	1.7

Notes:

[a] OECD (2002) and OECD (1999), Table 2.19, Standardized Unemployment Rates. Data for 1960–64 are from the OECD–CEP data set (see Bagliano et al., 1991). For Austria, Denmark and Switzerland, and for New Zealand prior to 1974, standardized rates are not available; unemployment as a percentage of the total labour force is used instead.

[b] OECD (2002) and OECD (1999), Table 7.10, Consumer price inflation and Table 3.7, Real GDP per person employed.

C. SHORT-RUN THEORY

1. The Mainstream Core

Our explanation of the five short-run cyclical movements depicted in Table 7.1 is better understood by contrasting two theories of the short-run outlined in Solow (1997). The first is characterized by the adoption of supply-side equilibrium analysis or simply NAIRU analysis, which was to become a defining feature of the mainstream macro core. Beginning in the late 1980s, the growing acceptance of NAIRU analysis among macroeconomists and policy makers reflected a belief that traditional Keynesian theory had failed. It could not provide a general theoretical explanation of the 'Great Inflation' of the late 1960s–early 1970s, and the subsequent stagflation, relying as alleged on unconvincing *ad hoc* arguments. In contrast, proponents of NAIRU analysis maintained they offered a theoretical explanation of short-run fluctuations when their theory was combined with New Keynesian micro findings stressing market imperfections. This was a special case of an expectations-augmented Phillips curve in which the expected rate of inflation variable was assumed to have a coefficient of one, that is, real wage bargaining was assumed. This endowed the short-run theory with a unique short-run equilibrium, the unemployment rate at which the rate of inflation was constant.

In its original formulation, NAIRU theory was embedded in the neoclassical competitive framework of flex-price labour and product markets and invisible hands. A well known key property of this version of the NAIRU theory was that it came with an indicator of the achievement of its unique equilibrium, when the economy was at full employment. Another favourable property of this NAIRU equilibrium was its stability. Not only did any rate of unemployment different from the NAIRU set in motion accelerating or decelerating rates of inflation, but when out of equilibrium the price mechanism (the invisible hand) automatically generated the necessary increases or decreases of aggregate demand to ensure a steady convergence back to the full employment NAIRU equilibrium. Finally, this equilibrium was one determined solely on the supply side. Aggregate demand played a passive role in New Keynesian short-run analysis. To put this in more technical terms, the NAIRU acted as a 'strong attractor'. Any disturbance of the equilibrium set in motion mechanisms bringing the system back to its unique equilibrium.

2. An Evaluation of NAIRU Theory

Our main concern in this section is whether NAIRU analysis provides an understanding of the forces generating the short-run movements in macro activity and insights for developing useful policy principles. Certainly, the original NAIRU models were formulated in unrealistic terms and are not acceptable. The continuance of high rates of involuntary unemployment in the Age of Decline led to growing dissatisfaction with the competitive model as a support system, and to major reformulations. While this involved discarding the neoclassical competitive framework (as lacking in descriptive value) and replacing it with imperfectly competitive fix-price labour and product markets, the real wage bargaining assumption was retained. Discarding the competitive framework introduced an element of realism into the analysis, but did nothing to resolve the problem posed by the NAIRU being unobservable. Reliance on econometric estimates of the NAIRU has its own problems. For example, econometric work revealed that with only minor changes in model specification, sample period used and definitions of the variables, a wide range of NAIRU estimates are generated (Setterfield *et al.*, 1992). In all other cases, changes in short-run movements in output and unemployment would be difficult to interpret, and NAIRU theory would not provide a guide to policy. Nevertheless there has been an inclination to interpret any increase (decrease) in the rate of inflation and fall (rise) in the rate of unemployment as occurring within the context of a fixed NAIRU. Similarly, any change in the rates of inflation and unemployment in the same direction is attributed to a shift in the NAIRU (Ball, 1992).

These interpretive and policy problems are magnified by the absence of an invisible hand guaranteeing convergence to the NAIRU equilibrium. Thus, studies to determine whether there existed adjustment mechanisms ensuring convergence found that the usually cited mechanisms, Pigou and Keynes effects, were not powerful enough to ensure convergence (Tobin, 1993).[4] For this reason alone, movements in output and unemployment cannot be interpreted unambiguously as part of a convergence process back to the NAIRU; they may be part of the system's endogenous dynamics.

Further empirical and theoretical work, and the continued high rates of involuntary unemployment extending into the 1970s and beyond, raised questions about the appropriateness of the real wage bargaining assumption. Theoretical work has challenged the argument that anything other than real wage bargaining imputes irrational behaviour to wage bargainers (Sawyer, 2001). Thus, the notion that labour is always in a position to determine its real wage has been challenged by those stressing the role of power in wage determination. Others have stressed that at low rates of inflation, wage bargainers will be little concerned with current rates of price inflation. Only at high rates of inflation does labour bargain to protect its real wages (Akerlof *et al.*, 2000). This latter view has been formalized in a two-track model of inflation, in which the coefficient of the expected price inflation variable varies with the current rate of price inflation (Eckstein and Brinner, 1972; Cornwall, 1994).[5] Over a range of high unemployment rates the Phillips curve is downward sloping, but eventually turns vertical at a low threshold rate of inflation. Finally current writings argue that the NAIRU is subject to hysteresis, varying with recent movements in the unemployment rate. In this case it can no longer perform the role of an attractor (Ball, 1992).

In summary, according to NAIRU analysis the movement from one short-run period to the next in Table 7.1 should be understood as a movement through successive NAIRU constrained adjustment processes. From what has been said, this interpretation of events cannot be accepted. Our conclusion is that NAIRU analysis provides little assistance in understanding the forces generating short-run movements in macro activity and in formulating useful policy principles. It cannot be a sound foundation for theory.

3. An Extended Keynesian Core

Our preferred explanation of the short-run movements in Table 7.1 is strictly Keynesian, a decision ultimately based on the unacceptability of NAIRU analysis, with its assumptions of real wage bargaining and supply determined equilibrium analysis. This rejection is not to argue that high rates of inflation in the real world can never be a serious problem with adverse implications. Indeed, inflation has had this impact and the constraining influence of inflation on aggregate demand plays an important role in our short-run core model. However, it does so within a traditional Keynesian core which includes fix-price markets and a negatively sloped Phillips curve. Beginning at some high rate of unemployment, there will be a range of unemployment rates in which policy makers are allowed some choice between inflation–employment points. Here aggregate demand determines equilibrium. However at some low rate of unemployment–high rate of inflation combination, the inflation rate becomes unacceptable to the authorities and this places a constraint on further increases in aggregate demand. Indeed at this particular unemployment rate, involuntary unemployment may even be substantial, in which case it is necessary to think of the economy being subject to at least two possible constraints: one the inflation constraint and the other a physical constraint based on resource limitations. Another possible limitation on aggregate demand is a balance of payments constraint, whereby further increases in aggregate demand lead to unacceptable payments deficits and legal restrictions on stimulative fiscal policies. Following custom, all of these constraints are referred to here as supply constraints.

With the concept of constraints on aggregate demand in mind, an extended short-run Keynesian core theory can be formulated in which 'fixed-price' markets are embedded in a macro system in which performance is driven by aggregate demand. In dynamic formulations, some form of Keynesian income–expenditure mechanism would be assumed, such as an accelerator–multiplier interaction, to model short-run dynamics. Thus in Table 7.1 unemployment (and output) performance from one short-run period to the next would be modelled as a succession of short-run damped cycles in GDP. Each cycle traces the interactions of an income–expenditure model in a Keynesian system subject to possible aggregate demand constraints.

D. A MEDIUM-RUN KEYNESIAN THEORY OF AGGREGATE DEMAND AND UNEMPLOYMENT

As pointed out in Section B, the five short-run periods in Table 7.1 divide naturally into two longer historical episodes: the first two short-run periods which we designated the Golden Age, and the latter three the Age of Decline. We consider the presence of medium-run episodes among the most prominent features of macro development of the post World War II period. While mild short-run fluctuations in activity were a common feature in both episodes, a comparison of the two episodes reveals at least one important difference between them, a pronounced difference in levels of aggregate activity. For example, overall average unemployment rates for the eighteen economies rose from 2.3 per cent in 1960–73 to 6.6 per cent in the 1974–2000 episode. Remarkably, mainstream macroeconomists have not adjusted their theory to recognize or account for this stylized fact common to all eighteen economies listed in Table 7.1. To correct for this avoidable error and explain performance in the Age of Decline and the Golden Age, the traditional Keynesian core must be extended.

This starts with the recognition of the need for a political economy theory of aggregate demand in which aggregate demand policies, and therefore the levels of aggregate demand and unemployment, are endogenously determined. This is to be contrasted with the exogenous treatment of fiscal policy in Keynes's *General Theory*. Specifically, the theory is formulated so as to answer two questions. Why, for a quarter of a century following World War II, were the fiscal authorities on average willing to pursue aggregate demand policies sufficient to achieve full employment in most economies? Second, why during the subsequent episode were the authorities unwilling to provide, again on average, the aggregate demand stimuli required to prevent a deterioration of macroeconomic conditions? The traditional

Keynesian core provides no answer to either question; evidently something important has been omitted from the core.

To answer these questions, we extend the traditional Keynesian core theory and model the missing medium-run policy responses as the outcome of an interaction between the supply of and demand for full employment policies. The strength of demand for full employment policies is determined by the distribution of political and economic power among organized interest groups. The party control theory of economic policy is the most prominent of the models focussing on the demand side and will be adopted here. It offers a political economy explanation of fiscal policy choices and differences in unemployment rates in terms of the relative strength of right-wing and left-wing parties, a reflection of the relative distribution of power between capital and labour (for example, Kalecki, 1971; Hibbs, 1987; Alesina *et al.*, 1997). According to this theory, labour is more willing than capital to trade price stability for lower unemployment; this preference is registered at the ballot box through its choice of political parties. From one episode to the next, differences in aggregate demand policies and unemployment are traced to shifts in political power within the economy.

However, the impact of the distribution of political power can account only for the strength of demand for aggregate demand policies. The policies supplied by the authorities depend upon whether or not there are constraints that limit their policy options. For example, if full employment levels of aggregate demand are seen to generate unacceptable inflation or unsustainable payments deficits, less than full employment aggregate demand policies will be adopted. In the absence of such constraints on aggregate demand, the political economy theory of policy assumes the policy authorities will provide the necessary stimulus to achieve full employment, provided it is demanded by the electorate. When undesirable side effects of full employment are present, they are traced to unfavourable institutions, such as an adversarial industrial relations system or an absence of institutions fostering strong export growth. An absence of such constraints on aggregate demand can be traced to a favourable set of the same institutions, allowing the authorities to foster full employment goals.[6]

E. A LONG-RUN KEYNESIAN MODEL OF AGGREGATE DEMAND AND UNEMPLOYMENT

The political economy theory of aggregate demand policy gives a deeper understanding of differences in unemployment performance shown in Table 7.1. The Golden Age was an episode characterized by the absence of constraints on aggregate demand polices in most of the capitalist economies and the presence of an historically strong power position for labour relative to capital. As a result, both the willingness of the authorities to supply full employment policies and strong political demands for full employment policies were present, leading to low unemployment rates. In the episode that followed the Golden Age, most of the economies found themselves faced with institutional constraints on expansionary aggregate demand policies. Full employment proved to be incompatible with other national goals, such as politically acceptable rates of inflation or sustainable balance of payments positions. As well, in the Age of Decline there were additional constraints on pursuing full employment policies. Among the new constraints were the introduction of an international monetary system which deregulated international capital flows and increased flexibility of exchange rates. These contributed to a proliferation of restrictive policies and the rise of unemployment rates almost everywhere. The Age of Decline was also an episode in which labour's power was weak compared to its relatively strong position in the Golden Age, and this led to weak demands for low unemployment policies.

When these medium-run episodes are considered in sequence, they form an historical period that can be modelled as a long-run theory of aggregate demand and unemployment. Each of the two episodes is a medium-run in the economy's long-run development, beginning and ending with a marked change in unemployment rates. Each episode is therefore characterized by a given set of key institutions and distribution of power. A sustained radical alteration in performance signals the arrival of a new episode, characterized by new institutions or a new power distribution or both, and a major shift in the dominant policy stance of the authorities. For example, the institutional shift from a cooperative industrial relations system to an adversarial one or a radical change in the international monetary system can lead to incompatibility of full employment with acceptable inflation rates or sustainable payments positions or both, and restrictive policies that would end a medium-run boom episode. A radical shift in the distribution of power is also a possible source of radical shift in macroeconomic policy and performance.[7] While we have concentrated on the eighteen capitalist economies as a group, we emphasize that an acceptable explanation of unemployment trends in most of the individual economies listed in Table 7.1 can be modelled by our model of long-run aggregate demand and unemployment.

F. ENDOGENIZING THE GROWTH OF AGGREGATE SUPPLY

The focus until now has been on aggregate demand and its growth, yet growth involves a supply side as well. In the discussion of the stylized facts in Section B it was pointed out that the rise in unemployment rates from the Golden Age to the Age of Decline indicates a failure of aggregate demand to keep pace with potential aggregate supply. Fortunately, this shortfall of demand was not so great as to lead to something like the Great Depression of the 1930s. This would suggest mechanisms at work in the post-World War II period preventing a major divergence between rates of growth of demand and potential supply. Our second extension of the traditional Keynesian core argues that indeed there was a mechanism limiting divergence in these growth rates. Contrary to the traditional treatment of aggregate supply in growth theory, this mechanism was simply that the rate of growth of aggregate supply was importantly influenced by the rate of growth of demand.[8] To model this we endogenize the usual production function determinants, treating each as strongly influenced by aggregate demand.

In support of endogenizing supply, consider an economy operating at the full employment rate of unemployment and a standard rate of utilization of the capital stock. First, consider the case in which inputs in the production process are 'given' and allow a small sustained increase in the rate of growth of aggregate demand. The following selection of responses from a real world catalogue of supply responses is assumed. According to Okun's law, the higher rates of growth of real demand will induce responses on the supply side in the form of higher rates of labour force participation, overtime shifts and a shift from part time to full-time work by some of those already employed. This demand-induced expansion of employment will be accompanied by higher utilization rates of existing physical capital. In addition, the induced higher growth rate of output will lead to higher growth rates of productivity because of 'learning by doing' effects. The point is that even when the stocks of capital, labour and technology are assumed 'given', the elasticity of supply with respect to aggregate demand will be positive. Demand influences the supply of output and productivity, even in the short run.

Next, consider an economy experiencing a period of prolonged strong growth in aggregate demand, such as the Golden Age. The cumulative effect of the increased capital utilization rates and tighter labour markets can be expected to have a pronounced and lasting impact on factor supplies and technical progress. For example, a sustained higher rate of growth of aggregate demand will pull labour out of low income elasticity-high productivity growth sectors, such as agriculture, to satisfy the now more rapid rates of growth of demand for labour in the industrial and service sectors. If this source of 'surplus labour' proves insufficient, the growing number of job vacancies will be filled by importing more labour from abroad. In this scenario, labour supply is endogenous and not part of the 'givens'.

The sustained additional stimulus to the growth rate of aggregate demand will also have a sustained positive effect on the growth rate of capital and the rate of introduction of new technologies, much of which will be technology transfer from the industrial leader(s). As Domar pointed out over half a century ago, investment generates output through the multiplier but also leads to increases in maximum output and labour productivity by increasing the capital stock. By generating high growth rates of investment, capital and technical progress, high growth rates of aggregate demand also contribute to high rates of growth of maximum output and productivity. Furthermore, as during the Golden Age, the induced high productivity effects of strong aggregate demand dampens inflationary pressure and helps sustain the strong demand policies.

Finally, it goes without saying that influences other than aggregate demand pressures affect the supply side categories and were contributing to the differences in rates of growth of output and, as shown in Table 7.1, differences in rates of growth of productivity among countries and over time. Nevertheless, the important implication of arguments presented in this section is that strong sustained aggregate demand was a necessary condition, not only for the low unemployment rates, but also for the high growth rates of productivity and output during the Golden Age. Similarly, stagnant demand conditions have been a key factor accounting for the poor macroeconomic performance of the last two and a half decades.

G. A KEYNESIAN MODEL OF UNEMPLOYMENT AND GROWTH

The discussion in the previous sections can be summarized as follows. First, in Sections D and E, we outlined how certain institutions and the distribution of power determined the dominant medium-run aggregate demand policy. This determined the overall strength of aggregate demand and the average unemployment rates in each episode, and therefore in the long-run. Second, in the previous section we endogenized the supply side of our core model by assuming that the usual production function categories are importantly influenced by aggregate demand. This is a key feature in the construction of a medium-run and long-run theory of aggregate supply. It complements the medium-run and long-run Keynesian models of aggregate demand and unemployment discussed in Sections D and E.

With aggregate demand and the unemployment rate determined, and with due account taken of the exogenous factors influencing performance, the average growth rate of aggregate supply in each episode is determined.

Together the medium-run and long-run unemployment and output performances can be modelled as an interaction of aggregate demand and aggregate supply, giving rise to a two-stage recursive process. In this sequence the extended Keynesian core incorporates a mechanism in which changes in the long-run growth of aggregate demand drive the economy by inducing changes in the long-run growth of aggregate supply. This contrasts sharply with the dominant role of aggregate supply in mainstream, Neoclassical and New Growth theory, in which aggregate demand adjusts passively to aggregate supply. Our core model also provides an explanation of the absence in the post-World War II period of a breakdown of the economy such as the Great Depression of the 1930s. The other growth theories make no attempt to explain this stylized fact.

H. SUMMARY

We have summarized two theoretical frameworks in the form of a limited number of propositions or a core. This was done in order to better focus on their relative merits and on key issues at the same time as it advanced our views. The first criterion for the inclusion of a proposition in a core theory was that it make clearer the theory's explanation of the stylized facts of macro economic development. The second criterion adopted in constructing a core was that the selected propositions provide a theoretical basis for deriving useful policy principles. In concluding, we pose the question 'Which core does best?' and then evaluate the relative success of the core theories in achieving their assigned tasks. This is done for the short-run, medium-run and long-run periods.

Evaluating the merits of the mainstream short-run analysis is essentially determining the explanatory power and usefulness of NAIRU analysis in explaining events and deriving helpful policies. We concluded in Section C that NAIRU analysis provided little assistance in understanding short-run cycles or in formulating policies to achieve superior unemployment–inflation outcomes. The alternative short-run Keynesian theory rejected NAIRU analysis, and in our view correctly modelled short-run performance as the outcome of a Keynesian income–expenditure mechanism subject to constraints. The important issue in short-run analysis is whether or not shocks to an ongoing dynamic process activate disturbance-amplifying tendencies. Exploring these possibilities is better understood by modelling short-run movements in the just described Keynesian fashion and expanding the analysis by adding on additional features. For example, in a longer paper we would have given attention to the kind of mechanism needed for controlling speculative 'bubbles'.

Determining the relative merits of the extended Keynesian and mainstream core theories in explaining medium- and long-run events and generating useful policies is more straightforward. By way of contrast, mainstream macroeconomics has restricted its analysis to a framework and core, in which only short-run and long-run categories are recognized. As a result, it has no theory to explain one of the outstanding stylized facts of the times, the alternating medium-run episodes, and has had to rely on a large number of *ad hoc* conjectures. Mainstream macroeconomists also insist on modelling the long run as a steady state balanced growth process. We consider it inappropriate to model a long-run period of major shifts in institutions, power structures (and technologies), let alone an historical period comprising a succession of two radically different performance records, by an exponential. Indeed mainstream macroeconomists have often been careless, introducing an exponential to model the long-run tendencies when the choice of exponentials is not well supported by history (Solow, 2003).

A key policy implication of our medium-run and long-run analysis is that strong aggregate demand pressures have beneficial employment effects and also stimulate the rate of growth of productivity and aggregate supply. Policies to solve the unemployment problem are therefore doubly beneficial. Econometric studies (W. Cornwall, this volume) support the policy implications of our theoretical model. Current mainstream growth theories treat the growth of aggregate supply as independent of the rate of growth of aggregate demand, and in so doing overlook a source of policy-induced higher growth.

Another key policy implication of our analysis is that stimulative aggregate demand policies are only a necessary condition for reducing high unemployment and slow growth. Among the additional necessary conditions for recovery is the prior removal of constraints on stimulative aggregate demand policies. This entails the presence of 'full employment friendly' institutions that permit the simultaneous achievement of other macro goals, such as acceptable rates of inflation at full employment. In their absence 'useful' high employment policies require the simultaneous implementation of special types of institutional change policies.[9]

The data in Table 7.1 are useful for evaluating which policies are full employment friendly. Note the economies fall into three groups: the 'low unemployment economies' with full employment in the Golden Age (1960–73) and in most of the subsequent short-run cycles; the 'high unemployment economies' with high unemployment rates both during and after the Golden Age; and the 'low–high unemployment economies' with full employment during the Golden Age followed by high unemployment. Case studies (Cornwall, 1994, Chapter 5) provide evidence that in the Golden Age full employment economies tended to be those with above average amounts of government intervention in their labour markets and with strong unions. Furthermore, countries such as Canada and the United States, with weak unions and deregulated markets performed the worst.

Mainstream macroeconomists have had little to offer in the way of explaining differences in unemployment rates among countries in the Golden Age. Rather the mainstream theory policy principle that evolved focussed on explaining the decline in macro performance between the Golden Age and the Age of Decline. This attributed the deterioration to widespread increases in government intervention in the market and to the rising power of unions. Comparing countries, this policy principle asserted that, other things being equal, the greater the displacement of economies from the competitive ideal the greater their deterioration in performance, and the greater the need for radical corrective policies. Again the data in Table 7.1 and the evidence cited in the last paragraph cast serious doubt on the correctness of these policy principles.

I. CONCLUSIONS

When answering the question ‘which core does best?’ it is difficult to escape the conclusion that mainstream macroeconomic theory is caught in an unproductive supply-determined equilibrium framework. Its short-run theory is dominated by the NAIRU and its long-run theory is essentially the adoption of an exogenous full employment supply constraint. An extra good dose of Keynes is the suggested remedy.

We say this in full awareness that our extended Keynesian analysis has its own weaknesses. As the term core suggests, we do not offer a complete or detailed theory, simply a foundation to support the detailed rebuilding of Keynesian macroeconomics. For example, our medium-run theory of aggregate demand and unemployment uses a very simple political economy theory of aggregate demand policy. Here, a research programme emphasising case studies can provide the detail needed to expand and refine the analysis. Indeed the growth of comparative studies of capitalist economies, focusing on ‘which economies work best’, is cause for optimism in this regard. For now, our focus has been to establish a broad direction for future research with a Keynesian heritage.

NOTES

[1] What constitutes the usable core in macroeconomic theory was the theme of a recent symposium ‘Is there a Core of Practical Macroeconomics that We Should all Believe?’, in the *American Economic Review*. See especially papers by Blanchard (1997) and Solow (1997).

2. Comparable data are not available for all eighteen countries prior to 1960. The year 2000 marks the end of the last complete cycle. The common end points for the periods shown in the table are those published in *OECD* (2002).
3. Note that together the two medium-run episodes viewed in succession constitute a long run.
4. The standard New Keynesian short-run model has been updated in the 'new consensus' theory of inflation targeting. For a general critique of NAIRU analysis which questions the stability of equilibrium in such models, see Sawyer (2001) and Setterfield (2005).
5. A related model in which the coefficient in the expected rate of inflation variable varies with the current rate of unemployment has been developed by Duesenberry (1991).
6. For a more formal account of the theory see W. Cornwall, this volume. For an early explanation of policy outcomes stressing both demand and supply influences see Gordon (1975).
7. Kalecki's (1971) famous model of the political business cycle offers an alternative explanation of the interruption of an episode of strong aggregate demand, that is capital's ability to force governments to enact high unemployment policies.
8. An exception would seem to be New Growth Theory and its treatment of technical progress. In some models technical progress is endogenously determined on the supply side. However, as the text makes clear, the focus in our extended Keynesian model is on technical progress, as well as rates of growth of capital and labour, being determined on the demand side.
9. The continuous output of comparative studies focusing on 'which economies work best' attests to the importance of institutions, especially in the labour market.

REFERENCES

Akerlof, G., W. Dickens and G. Perry (2000), 'Near-Rational Wage and Price Setting and the Long-Run Phillips Curve', *Brookings Papers on Economic Activity*, No. 1.

Alesina, A. and N. Roubini with G. Cohen, M. Setterfield, L. Osberg and D. Gordon (1997), *Political Cycles and the Macroeconomy*, Cambridge, MA: MIT Press.

Bagliano, F.-C., A. Brandolini and A. Dalmazzo (1991), 'OECD–CEP data set (1950–1988)', *Working Paper No. 118*, London, Centre for Economic Performance, June.

Ball, L. (1992), 'Aggregate Demand and Long-run Unemployment', *Brookings Papers on Economic Activity*, No. 2.

Blanchard, O. (1997a), 'Is There a Core of Usable Macroeconomics?', *American Economic Review Papers and Proceedings*, No. 2.

Blanchard, O. (1997b), 'The Medium Run', *Brookings Papers on Economic Activity*, No. 2.

Blanchard, O. and Melino, A. (1999), *Macroeconomics, First Canadian Edition*, Scarborough: Prentice Hall.

Cornwall, J. (1994), *Economic Breakdown and Recovery: Theory and Policy*, Armonk, NY: M.E. Sharpe.

Cornwall, J. and W. Cornwall, (2001), *Capitalist Development in the Twentieth Century: An Evolutionary-Keynesian Analysis*, Cambridge: Cambridge University Press.

Cornwall, W. (1999), 'The Institutional Determinants of Unemployment', in M. Setterfield (ed.), *Growth, Employment and Inflation: Essays in Honour of John Cornwall*, London: Macmillan.

Duesenberry, J. (1958), *Business Cycles and Economic Growth*, New York: McGraw-Hill.

Duesenberry, J. (1991), 'A New American Inflation?', in J. Cornwall (ed.), *The Capitalist Economies: Prospects for the 1990s*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Eckstein, O. and Brinner, R. (1972), *The Inflation Process in the United States: A Study Prepared for the Joint Economic Committee, US Congress*, Washington, DC: U.S. Government Printing Office.

Gordon, R.J. (1975), 'The Demand and Supply of Inflation', *Journal of Law and Economics*, **18**(3), Winter.

Hibbs, D. (1987), 'Political Parties and Macroeconomic Policy', in D. Hibbs, *The Political Economy of Industrial Democracies*, Cambridge, MA: Harvard University Press.

Kalecki, M. (1971), 'Political Aspects of Full Employment', in *Selected Essays on the Dynamics of the Capitalist Economy*, Cambridge: Cambridge University Press.

Mankiw, N.G. and W. Scarth (2001), *Macroeconomics, Canadian Edition*, New York: Worth Publishers.

OECD (1999), *Historical Statistics 1960–1999*, Paris: OECD.

OECD (2002), *Historical Statistics 1970–2000*, Paris: OECD.

Ragan, C. and R. Lipsey (2005), *Macroeconomics*, Toronto: Addison Wesley.

Sawyer, M. (2001), 'The NAIRU: A Critical Appraisal', in P. Arestis and M. Sawyer (eds), *Money, Finance and Capitalist Development*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Setterfield, M. (2005), 'Central Bank Behaviour and the Stability of Macroeconomic Equilibrium: A Critical Examination of the New Consensus', in P. Arestis, M. Baddeley and J. McCombie (eds), *The New Monetary Policy: Implications and Relevance*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Setterfield, M. et al (1992), 'Searching for a Will o'Wisp: An Empirical Study of the NAIRU in Canada', *European Economic Review*, **1**.

Solow, R. (1997), 'Is There a Core of Usable Macroeconomics that We Should all Believe In?', *American Economic Review Papers and Proceedings*, **87**(2).

Solow, R. (2000), 'Toward a Macroeconomics of the Medium Run', *Journal of Economic Perspectives*, **14**(1), Winter.

Solow, R. (2003), 'General Comments on Part IV', in P. Aghion (ed.), *Knowledge, Information and Expectations in Modern Macroeconomics: In Honour of Edmund Phelps*, Princeton, NJ: Princeton University Press.

Tobin, J. (1993), 'Price Flexibility and Output Stability: An Old Keynesian View', *Journal of Economic Perspectives*, **7** (Winter).

8. A Keynesian model of unemployment and growth: an empirical test

Wendy Cornwall*

A. INTRODUCTION

This chapter presents a simple econometric test of the core theory proposed by John Cornwall in Chapter 7 of this volume. Building on a Keynesian foundation, and with the requirement that it must have the capacity to explain the behaviour of real economies, the theory emphasises the central importance of aggregate demand. First, the level and growth of aggregate demand are viewed as the outcome of policy choices, which are endogenous to this model. Second, growth of aggregate demand induces aggregate supply growth, which is therefore also endogenous.[1] These are the two essential components of the model that are to be tested here.

The chapter has two main parts. The first (Section B) introduces and discusses two equations implied by the theory. These determine aggregate demand and aggregate supply growth, represented here by unemployment rates and productivity growth respectively. Unemployment rates are the outcome of aggregate demand policy, which is endogenously determined via a constrained optimisation process. The determinants of the constraints and of the preferences to be optimised are labour market and other institutions, and the distribution of power. These are therefore ultimately determinants of the unemployment rate. Productivity growth depends on the result of the aggregate demand policy chosen, through its effect on unemployment and other demand variables such as investment spending. These and other determinants of productivity growth, and the routes by which they act, are also examined in this part of the chapter.

The second part of the chapter (Section C) presents the standard econometric analysis used to test the model's ability to explain the changing unemployment and productivity performances of a group of OECD

* The author wishes to thank Mark Setterfield and Mark Roberts for their helpful comment and criticism.

economies since World War II. The sample of sixteen economies includes the G8 countries, and eight smaller longstanding members of the OECD. It covers the period 1960–2000, which encompasses the Golden Age years 1960–73 and the subsequent years of inferior economic performance.[2] These two episodes, distinguished by their starkly different performance, include five short-run cycles; two are in the Golden Age while the remaining three comprise the second episode. The data for each of these cycles are averaged to remove cyclical variation, yielding five observations for each of the sixteen economies; these are pooled for estimation. In this section the data are defined and discussed, as are the estimation results.

B. FORMALIZING THE CORE MODEL

1. A Formal Model of Aggregate Demand and Unemployment

The model used here was originally used by Lipsey (1965) to provide definitions of various types of employment, such as demand-deficient and structural, that would permit formal testing and prediction. The feature of the model essential to our analysis is that it represents actual unemployment as the consequence of policy-makers exercising their preferences, constrained by the menu of choices presented by the Phillips curve. This formulation allows us to bring together two distinct strands of analysis relevant to this study. First, party control theory focuses on the demand for full employment policies. The strength of this demand is attributed to the distribution of power between organized interest groups, notably labour and capital. Fiscal policy choices are explained as the result of the relative strength of the left-wing and right-wing political parties that represent their interests (see Kalecki, 1971; Hibbs, 1987; Alesina *et al.*, 1997). A fundamental assumption is that labour will accept a larger increase in inflation to achieve a given drop in unemployment than will capital. The econometric results from this work are weak, relying on the inclusion of past values of the dependent variable to produce high R-squares, but failing to show a strong relationship between power and unemployment rates. This theory deals only with the demand side of macroeconomic policy choice. It does not consider the supply side, that is the set of available policy options, which depends upon the constraints imposed by competing policy targets. The most commonly studied competing target variable is inflation. The second strand of analysis focuses on these constraints, summarized by the position of the Phillips curve, and traced to institutions of the labour market, largely confined to those governing collective bargaining. An OECD (1997) study considers a large number of articles of this type, and retests their models with updated data, as well as specifying a new model covering more dimensions of the bargaining structure. Its authors find very little evidence to support a link between measures of economic performance and the institutions of collective bargaining. This weakness, and that of party control theory, flows from each considering only part of the model. The point to be made is that *both* the strength of demand *and* the constraints facing the policy-maker must be considered, and that is the approach taken here.

This allows us to formalize the core theory of Chapter 7, which endogenized macroeconomic policy as the outcome of both demand and supply, stressing the roles of institutions and the distribution of power. In terms of the model, these are determinants of the slope of the policy-makers' indifference curves and of the slope and position of the Phillips curve. Consider first the demand for full employment policies. The political preference function can be written as

$$M = M(\dot{p}, U; V_1), \text{ with } M_p, M_u > 0$$

where M measures the disutility of pairs of unemployment (U) and inflation ($\dot{p}$) rates. It is assumed to yield strictly concave indifference curves, and its parameters are determined by the political and institutional variables in vector V_1. These parameters vary with the political party in power, and left-leaning governments are expected to attach a relatively greater weight to unemployment than will right-of-centre governments, yielding steeper indifference curves. Preferences are also influenced by custom and tradition, leading to different weights, and so different slopes, even among countries with similarly left- or right-wing governments.

The supply of full employment policies, that is the set of choices open to governments, is constrained by the prevailing Phillips curve.[3] This is assumed to be downward sloping in the relevant range, so that there is a long-run trade-off between unemployment and inflation that governments can exploit. The Phillips curve can be written as

$$\dot{p} = f(U; V_2)$$

where V_2 is a vector of variables that influence its slope and position. Labour market institutions are of special interest, but to examine the medium- and long-runs, we stress institutions having a broad and persistent influence on labour market behaviour, such as the prevailing industrial relations system, and whether it is cooperative or adversarial.[4] International demand

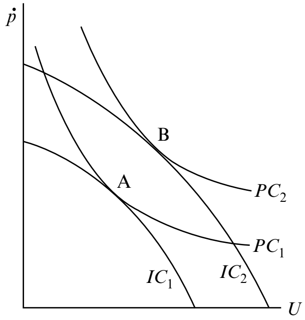


Figure 8.1 Optimising political preferences

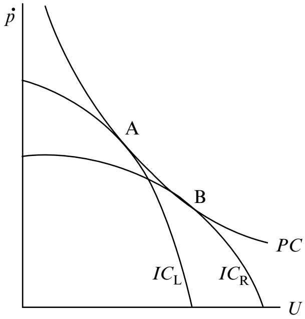


Figure 8.2 Alternative political preferences

conditions are also likely to affect the position of the Phillips curve via their impact on investment and productivity.

Each unemployment outcome results from the government attempting to optimize its preference function subject to the existing Phillips curve. Since the preference function measures disutility, the indifference curve closest to the origin is preferred. In Figure 8.1 this is shown at point A, the point of tangency between the Phillips curve (PC_1) and the indifference curve (IC_1). If the Phillips curve shifts to PC_2, there is greater disutility at the optimum point B.

The effect of alternative preference functions is shown in Figure 8.2, where the steeper indifference curve (IC_L) depicts the effect of a more left-wing government than curve IC_R. Given the prevailing Phillips curve, optimisation occurs at point A, with lower unemployment and higher inflation than at point B, which would be the choice of a right wing government.[5]

2. Connecting Aggregate Demand and Aggregate Supply

Each observed (optimal) unemployment rate can be represented by a reduced form equation

$$U = U(V_3) \tag{1}$$

in which the vector V_3 contains predetermined and exogenous variables from V_1 *and* V_2, including the political and institutional variables that are of particular interest to our model. To make the connection between the institutions and power structures that determine aggregate demand policy and unemployment rates, and the demand-determined supply performance of the core model, we use the following productivity growth equation

$$\dot{q} = f(SD, B; V_4) \tag{2}$$

where $\dot{q}$ is average productivity growth of the economy, *SD* is a measure of its stage of development, *B* represents the opportunity for catching up, and V_4 is a vector of demand variables. The first two variables influence productivity growth as a consequence of the economy's level of economic development, in contrast to the effect of current demand which is a prime interest in this chapter.

The model that yields this equation is explored fully in Cornwall and Cornwall (2001, Chapter 10). It has three sectors that have different productivity levels and growth rates. This level of disaggregation is sufficient to show how changes in sectoral output shares brought about by development govern the allocation of labour and determine the growth rate of average productivity. These long-run changes create a three-phase growth path as the agricultural, industrial and service sectors in turn dominate output and employment. As labour shifts among the sectors, the economy's average productivity growth rate changes, accelerating during industrialization and slowing as the services sector expands. This generates a component of the economy's productivity growth rate that varies with its stage of development. The inclusion of a 'stage of economic development' variable allows separation of this component from the contribution of current demand variables to growth.

Also related to development is the technology an economy currently uses relative to the best technology that is available. The variable *B* is included to account for any growth bonus from catching up to a technological leader. The larger is the technology gap, the greater is the potential growth bonus. Its actual size depends also on the rate of modernization, that is on investment in the new technologies. If we consider two economies with equal investment–GDP ratios, the one with the lower productivity level will have faster productivity growth; it skips more capital vintages, raising the output of every worker provided with the new technology by a larger proportion than in the less backward economy. Thus a given ‘effort’, for example, amount of investment, yields different results when the technology gap is different. This must be accounted for to assess more accurately the effects of the demand variables.

We turn now to the demand variables and their role in the demand-determined growth model presented above. The data show the growth rates of two components of aggregate demand, investment and exports, to be closely correlated with the growth of GDP and productivity. Investment has a dual role, influencing both demand and supply. Strong demand for investment goods initiates a multiplier–accelerator process, increasing capital utilization, employment and labour reallocation, all of which push up the growth rates of output and productivity in the short run. In the longer term, these new investment goods affect aggregate supply and productivity, both by increasing the capital–labour ratio and by introducing improved technologies. In addition, continued high demand for investment goods induces the capital goods industry to incorporate the latest technologies in its product, and may also induce further inventions and innovations, actually speeding up technological progress (Cornwall,1977, Chapter 7). The growth rate of investment is a prime candidate in our list of determinants of productivity growth.

Export growth affects productivity growth primarily by justifying the adoption of best practice technologies and allowing exploitation of dynamic scale economies.[6] Especially in small economies, expanded export markets justify using larger scale production methods, and in small and large economies alike the need to remain competitive in foreign markets encourages use of best available production techniques. In addition, expanding intra-industry trade spurs the adoption of both product and process innovations, particularly in markets governed by non-price competition. The implications for productivity growth are clear, and suggest that the growth rate of exports, particularly of manufactured goods, plays a key role in determining productivity growth rates.

In addition to growing aggregate demand, the core theory stresses the part played by high levels of aggregate demand, implying that high unemployment may be detrimental to growth. Here we look beyond the short-run impacts of changes in unemployment, that is the rapid output and productivity growth typical of the recovery stage of the business cycle. Instead, we focus on the longer term effects of persistent high (or low) levels of economic slack, which influence both the amount and the type of investment undertaken. The reduced macro risk provided by high and growing aggregate demand induces firms to increase capacity by investing in technologies that exploit scale economies and introduce innovations, both of which increase productivity growth. But this 'enterprise investment' carries the problem of indivisibilities, and consequently the possibility of excess capacity. In protracted periods of high unemployment and sluggish demand growth, firms are likely to choose a 'defensive investment' strategy, simply replacing worn out components of existing equipment. This reduces the risk of excess capacity, but postpones the productivity gains that new technology would bring. A second type of risk accompanies new technologies, because they often need modification to suit local conditions and require adjustments by both labour and management. These measures carry unknown costs, presenting a risk that firms avoid in lean times by shifting to a defensive investment strategy. Like the risk of excess capacity, this keeps productivity growth lower than it would be in times of strong demand growth. Both types of risk suggest that unemployment has a negative impact on productivity growth.

There is a second way in which unemployment can influence the type of investment. In the case of cooperative industrial relations systems, labour is less likely to oppose the introduction of new technologies. This not only encourages firms to invest in new technologies, but reduces the likely costs of the adjustments and modifications associated with such change. Cooperative industrial relations are typical of the low unemployment economies. This characteristic strengthens the likelihood that an unemployment variable will be able to capture the effect on productivity growth of this qualitative dimension of investment.

C. ESTIMATION OF THE MODEL FOR 1960–2000

1. The Reduced Form Unemployment Equation

We have conceptualized our model as having two analytical levels. The 'deeper' of these is represented by equation (1), which investigates the roles of institutions and power in establishing aggregate demand policy, and through it the level of unemployment. At the second level, we examine productivity growth in terms of economic variables that stress its dependence on aggregate demand. We tested the model using a sample of sixteen OECD countries, with data for the years 1960–67, 1968–73, 1974–79, 1980–89 and 1990–2000. The intervals approximate the business cycles of the period, and provide two observations for the Golden Age and three for the subsequent episode covered by the data.

The properties of the political preference function and the Phillips curve yield complex non-linearities in the reduced form. Further complications

Table 8.1 Definitions of the variables used in the unemployment equation

U:	average unemployment rate over the period
LV:	left-of-centre votes as a proportion of total votes cast in elections during the period
EMS:	dummy variable for membership in the European monetary system
STR:	logarithm of man days lost to strikes per thousand workers, lagged one period
ED:	reduced form estimate of the weighted average unemployment rate of the other fifteen countries in the sample scaled by the country's own exports to GDP ratio
LINF:	average inflation rate, lagged one period
DLINF:	lagged inflation (LINF) multiplied by a dummy variable for the last three periods covered

Sources: Voting data, Mackie and Rose (1991, 1997); strike data, *ILO Yearbook of Labour Statistics*, various issues; OECD data are used for the remaining variables. The sixteen countries included are: United States, Japan, Germany, France, Italy, United Kingdom, Canada, Australia, Austria, Denmark, Finland, Ireland, The Netherlands, Norway, Sweden and Switzerland. There were five observations for each, for the years 1960–67, 1968–73, 1974–79, 1980–89 and 1990–2000.

arise from the unavoidable imprecision of measures of institutional characteristics. For these reasons we have used a simple linear specification for the unemployment equation. The variables used are defined in Table 8.1, which also lists the countries used in the sample.

The left-of-centre votes variable is used to measure political preferences for inflation and unemployment. Because governments in democratic systems must face election, we assume their *effective* political preferences reflect prevailing voting patterns. A high proportion of left-of-centre votes will push right wing governments toward the centre, or give leftist governments the power base to resist the demands of organized business and financial interests.

Institutions and history also affect political preferences. A prime example is the level of aversion to inflation; among the responses to the experience of rapid inflation is a greater readiness to adopt measures to avoid its repetition. These include increasing the independence of the central bank[7] and a greater readiness to join international monetary agreements, both of which act to distance monetary policy from the political arena and its concern with unemployment. As indicators of social preferences, they help to explain why two equally 'left wing' countries experience different unemployment outcomes. The country that gives more independence to its central bank is expected to have flatter indifference curves, showing its relatively greater willingness to accept higher unemployment to keep down inflation. Membership in the European monetary system has a similar effect. It required coordinated exchange rate policies, which lowered inflation (see, for example, Jenkins, 1996); the decision to join reveals a preference that would reduce the slope of the indifference curves.[8] An earlier study (Cornwall, 1999) used an index of central bank independence constructed by Cukierman *et al*., (1992) and found it to play a significant role in explaining unemployment rates prior to the 1990s. However the index has not been updated to cover the longer time period of this study. Instead we place greater reliance on a dummy variable included to capture the effect of the similarly inflation-averse behaviour of joining the European monetary system.

The extent of conflict in industrial relations is a determinant of the position of the Phillips curve. Its effect is measured here by the strike volume, lagged one period to allow time for changes to demonstrate their longer term effects. Cooperation between labour and management at the firm level reduces conflict in wage negotiations, reducing the likelihood of strikes. Lower strike activity is expected to act via wage bargaining to improve the inflation-unemployment trade-off.

Lagged inflation often appears as a determinant of the position of the Phillips curve. However, here we use the average rate for the previous subperiod. This relatively long time lag makes it a measure of the cumulative effects of past inflation on the position of the Phillips curve. These effects are transmitted through institutional change, such as real and relative wage protection becoming a legitimate objective of wage settlements. Past inflation also affects the Phillips curve via the restrictive policies it elicits, generating hysteretic effects. A country's political preferences are also likely to be influenced by its historical inflation experience, as discussed above. The use of a dummy variable multiplied by lagged inflation is used to test whether the accelerating inflation that began in the late 1960s led to a general change in tolerance for inflation.

Exposure to international demand conditions presents a slight complication for estimation. The dependent variable is the unemployment rate, chosen as a prime indicator of the strength of aggregate demand. Our pooled sample consists of sixteen open economies, and most of their trade is within the OECD, that is with each other. It therefore follows that for each country, a good measure of external demand is the weighted average unemployment rate of the others in the sample. Scaling by the export–GDP ratio would reflect the considerable variability of openness, both among countries and over time. However, because we are using pooled cross section data this introduces simultaneity, leading to biased estimates. To circumvent this, and to ensure that the equation is truly a reduced form, an instrument for the weighted average unemployment rate was constructed by regressing it on a subset of predetermined variables in the model; the resulting variable was then scaled by the exports to GDP ratio as a measure of exposure to international demand conditions.

Table 8.2 reports the regression results for two versions of the reduced form unemployment equation.

With the exception of lagged inflation, tests for changes in the coefficients after 1973 showed the estimates to be very stable. When the post-1973 dummy variable (*DLINF*) is used, its coefficient is positive and statistically significant, while the coefficient of lagged inflation becomes negative. This suggests that the higher inflation of the late 1960s and after triggered a marked change in policy response, reflecting an increase in aversion to inflation. The negative sign of the coefficient for the earlier episode is consistent with greater tolerance of the low inflation prior to the late 1960s and the preoccupation with growth and employment as prime policy targets in the Golden Age. The coefficient of the external demand variable (*ED*) is significant, and its estimated value in the vicinity of one in each specification is consistent with the extent of trade among these countries, and sufficient time in each period for the transmission of changes.

Table 8.2 Regression results for the unemployment equation

Equation		*1a*	*1b*
Left-of-centre votes	*LV*	−4.392	−3.911
		(3.71)	(3.60)
EMS membership	*EMS*	2.468	2.470
		(4.04)	(4.44)
Strikes	*STR*	0.831	0.884
		(6.69)	(7.77)
External demand	*ED*	1.118	0.815
		(4.45)	(3.39)
Lagged Inflation	*LINF*	0.188	−0.359
		(2.61)	(2.38)
Lagged inflation, post-1973	*DLINF*	–	0.469
			(4.02)
Constant		−0.673	0.406
		(0.84)	(0.52)
Adjusted R^2		0.7405	0.7846

Note: The figures in parentheses are the absolute values of the t-statistics.

Of special interest to our model are the estimated coefficients of the institutional and power variables. These are all of the expected sign, and significantly different from zero at the 5 per cent level. The estimates show *EMS* membership is associated with higher unemployment as expected. An increase in left-of-centre votes decreases unemployment, as governments' effective preferences shift; the estimates suggest that a 12-percentage point increase in left-of-centre votes will reduce unemployment by about half a percentage point. Higher strike ratios where adversarial industrial relations prevail act via the Phillips curve to increase unemployment. Lagged inflation affects both the Phillips curve via hysteretic effects as discussed and political preferences when experience induces inflation aversion. In each case, the result is to increase unemployment. The strong partial correlations of these variables with unemployment, and the high overall explanatory power of the estimates support the view that power and institutions play a significant part in determining unemployment rates.

These results can be compared with earlier estimates based on only the first four time periods and a slightly different sample of countries (see Cornwall and Cornwall, 2001, Chapter 5). Other differences include the use of an index of central bank independence and a different measure of external demand. Despite all this, the estimated coefficients of left-wing votes, strikes, EMS membership and post-1973 lagged inflation are of comparable magnitude and lead to the same conclusions in the two samples.

2. The Productivity Growth Equation

The variables used for estimation of the productivity equation are defined in Table 8.3.

To account for stage of development effect we use real per capita income at the start of each period, because changes in sectoral output structure are responses to changing demand caused by rising per capita incomes. Everything else being equal, two economies with different per capita incomes will have different demand structures, and the one with the largest share of its output in the highest productivity growth sector will have the highest average productivity growth rate. Second, to capture the catch-up growth bonus we use the gap between US and domestic per capita income as a proportion of domestic per capita income for each country. This ratio is a proxy for the technology gap, which represents the opportunity for technological borrowing. The rapid narrowing of technology gaps especially early in the sample period demonstrates that this opportunity was exploited.

The contribution of exports to productivity growth is measured using the growth rate of the volume of goods exports. Over the period covered,

Table 8.3 Definitions of the variables for the productivity growth equation

$\dot{q}$	Growth rate of real GDP per person employed; average for each period.
y	per capita GNP in real 1980 international dollars (thousands); value at the beginning of each period.
TGAP	$(y_{us} - y_i)/y_i$, where y_{us} is US per capita income, and y_i is the per capita income for each of the other countries in the sample; value at the beginning of each period.
XG	Growth rate of the volume of goods exports; average for each period.
IG	Growth rate of investment in machinery and equipment; average for each period.
U	Standardized unemployment rate; average for each period.

Notes: All data are from the OECD. Precise sources are available on request.

manufactures dominated exports, so that this variable is expected to explain a substantial part of productivity growth. The growth rate of investment in machinery and equipment is used as the best measure of the introduction and spread of new technologies. This variable is also sensitive to economic slack, but to include economic slack more directly, we use standardized unemployment rates.

Table 8.4 shows the regression results and test statistics for two variants of the productivity equation.

Chow tests indicated that we are justified in pooling the data for the five periods. As well, dummy variables were used to test for intercept and slope differences between the pre- and post-1973 periods, which had very different performance records. No significant differences were found, indicating that the coefficients were stable over these two periods. This allows us to conclude that the effects of structural changes between the two periods were captured by changes in the values of the explanatory variables used for estimation. Among several tests for heteroscedasticity, none showed evidence of its presence.

Equation 2a includes both the stage of development variable (per capita income) and the technology gap variable (*TGAP*). Equation 2b omits per capita income to test our specification. In both equations, the estimated slope coefficients have the expected signs, and are statistically significant at the five per cent level. The positive coefficients of the technology gap, export growth and investment growth show that all increase the average productivity growth

Table 8.4 Regression results for the productivity growth equation[a]

Equation	*2a*		*2b*	
Per capita income (y)	−0.116		–	
	(2.94)			
Catch-up ($TGAP$)	0.742		1.524	
	(2.13)		(6.43)	
Export growth (XG)	0.240		0.253	
	(6.59)		(6.67)	
Investment growth (IG)	0.060		0.053	
	(2.09)		(1.78)	
Economic slack (U)	−0.068		−0.113	
	(2.11)		(4.05)	
Constant	1.946		0.294	
	(3.11)		(1.02)	
Standard error	0.739		0.776	
Adjusted R^2	0.790		0.768	
Test Results	*Test statistic*	*Critical value*	*Test statistic*	*Critical value*
Stability: Break at 73–74				
Chow (6, 68)	2.931	2.137		
Specification: omit per capita income				
Hocking Sp			0.118	2.027

Note: [a] The absolute values of the t-statistics are in parentheses.

rate, while the presence of economic slack reduces it. The negative coefficient for per capita income reflects the expected lower productivity growth rate brought by the increasing service sectors in high income economies.

The omission of per capita income does not cause rejection of the hypothesis of no misspecification according to several tests; Table 8.4 reports the Hocking test statistic. While the omission causes relatively little change in the values of the coefficient estimates (except for per capita income), it should be recalled that these tests value parsimony. For theoretical reasons, our view is that equation 2a is to be preferred. In each estimate, the effect of unemployment upon productivity growth is clear. Depending on which equation is used, a 5-percentage point increase in unemployment is expected to lower productivity growth between 0.3 and 0.6 percentage points directly, with further decreases transmitted through the effect of economic slack on investment behaviour. These results support our theoretical model by demonstrating the link between economic performance (productivity growth) and the institutions and power structure of economies, via their influence on aggregate demand.

Like the unemployment equation, the productivity results can be compared to earlier estimates using data for only the first four periods. In both samples, the estimated coefficients are significant at the 5 per cent level, and have the expected signs and similar values. The current estimates show export growth to be more important, and the remaining variables slightly less so, than the earlier results; these differences notwithstanding, both estimates support our theory.

D. CONCLUSION

The theory presented in Chapter 7 is in large part a response to the failings of recent developments in macroeconomics. Its overarching objective is to provide a theory that can explain the broad historical tendencies of post-World War II macro development, and that can provide a sound basis for deriving useful policy principles. The most outstanding feature of the post-World War II era has been its clear division into the economic success of the Golden Age, followed after the mid-1970s by the generally poor performance that in many countries continues still. Current mainstream, that is New Keynesian, economics has no theory that can account for this medium run, nor does it have a long-run theory that examines the effects on economic performance of the observed changes in institutions and power structures, or in the changing output structure brought about by rising incomes. Yet these are fundamental to the long-run development of capitalist economies. The model tested here considers all of these factors, although the emphasis has been on the roles of power and institutions on the choice of aggregate demand policy and the impact of the policy choice on unemployment and productivity growth. This emphasis presents economic outcomes as dependent upon collective action, a clear break with the free market *laissez-faire* analysis and its individualism.

We have estimated a model that encompasses the long run; it links aggregate demand to the changing institutions and power structure of the economy, and links supply, in the form of productivity growth, to the strength of aggregate demand. Our model is designed to examine the ways institutions impose boundaries on the set of policy options, and how the distribution of power determines which policies are selected from those available. These are policies that influence both aggregate demand and aggregate supply, and account very significantly for the differences in economic performance observable over time and among countries. This is the first level of the analytical model used for our empirical tests. The econometric results strongly support the importance of institutions and the distribution of power on aggregate demand outcomes.

A second class of problems stems from the New Keynesian adoption of neoclassical growth theory, which explains growth in terms of proximate supply-side factors such as the growth of capital and labour. It neglects changes in the composition of output and employment as economies modernize and fails to consider the effect of overall economic performance, that is the state of aggregate demand, on growth rates. Maintaining consistency with the core theory, we avoid these problems by using a model that explains economy-wide productivity growth as the outcome of both demand and supply, and stresses the impact that changes in the distribution of outputs and inputs among different sectors of the economy have on performance. This is the second level of our analysis, and the estimation results support the link between aggregate demand and supply.

NOTES

1. For a variety of perspectives on demand-led growth, see Setterfield (2002).
2. This uses more recent data to update earlier work (Cornwall, 1999).
3. This analysis, like that of Lipsey (1965), permits shifts of the Phillips curve, including shifts induced by policy.
4. Some studies, for example Layard *et al.* (1991), use such frequently changed individual regulations as payroll taxes and wage replacement ratios to measure real wage rigidities, which are then used to explain unemployment. Others find that such rigidities fail to explain differences in unemployment (Freeman, 1995).
5. Clearly, policy targets other than unemployment may well complicate the analysis. Perhaps the most likely is the balance of payments position; this may impose a binding constraint, possibly reached before the preferred (optimal) unemployment rate has been reached. Then actual unemployment will be on a higher indifference curve and not at a point of tangency with the Phillips curve.
6. McCombie and Roberts (2002) stress not only these direct contributions of exports to growth, but also the effect of exports in allowing domestic income and consumption to grow without causing balance of payments problems.
7. Debelle and Fischer (1994), show that inflation-averse countries tend to have more independent central banks.
8. It has been suggested that aversion to inflation is more accurately laid at the door of financial capital, rather than productive capital (especially in a mark-up pricing system), and that these variables are doing this. In this simple model, we have not attempted to make such a distinction. In fact it is not clear that it could be done, given the very close relationships that exist between the two in many European economies and in Japan.

REFERENCES

Alesina, A., N. Roubini and G. Cohen (1997), *Political Cycles and the Macroeconomy*, Cambridge, MA: MIT Press.

Cornwall, J. (1977), *Modern Capitalism: Its Growth and Transformation*, London: Martin Robertson.

Cornwall, J. and W. Cornwall (2005), 'Power and institutions in macroeconomic theory', in B. Laperche and D. Uzunidis (eds), *John Kenneth Galbraith and the Future of Economics*, London: Palgrave Macmillan, Chapter 8.

Cornwall, J. and W. Cornwall (2001), *Capitalist Development in the Twentieth Century*: *An Evolutionary-Keynesian Analysis*, Cambridge: Cambridge University Press.

Cornwall, W. (1999), 'The institutional determinants of unemployment', in M. Setterfield (ed.), *The Political Economy of Growth, Employment and Inflation*, London: Macmillan, Chapter 5.

Cukierman, A., S.B. Webb and B. Neyapti (1992), 'Measuring the independence of central banks and its effect on policy outcomes', *World Bank Economic Review*, **6**, 353–98.

Debelle, G. and S. Fischer (1994), 'How independent should a central bank be?', *Goals, Guidelines, and Constraints Facing Monetary Policymakers*, Federal Reserve Bank of Boston Conference Series, No. 38.

Freeman, R. (1995), 'The limits of wage flexibility to curing unemployment', *Oxford Review of Economic Policy*, **11**(1), 63–72.

Hibbs, D. (1987), 'Political parties and macroeconomic policy', in D. Hibbs, *The Political Economy of Industrial Democracies*, Cambridge, MA: Harvard University Press.

Jenkins, M. (1996), 'Central bank independence and inflation performance: panacea or placebo?', *Banca Nazionale del Lavoro Quarterly Review*, **49**(June), 241–70.

Kalecki, M. (1971), 'Political aspects of full employment', in M. Kalecki (ed.), *Selected Essays on the Dynamics of the Capitalist Economy*, Cambridge: Cambridge University Press.

Layard, R., S. Nickell and R. Jackman (1991), *Unemployment: Macroeconomic Performance and the Labour Market*, Oxford: Oxford University Press.

Lipsey, R.G. (1965), 'Structural and deficient-demand unemployment reconsidered', in A.M. Ross (ed.), *Employment Policy and the Labour Market*, Berkeley, CA: University of California Press.

Leon-Ledesma, M. and A.P. Thirlwall (2002), 'The endogeneity of the natural rate of growth', *Cambridge Journal of Economics*, **26**, 441–59.

Mackie, T. and R. Rose (1991), *The International Almanac of Electoral History*, London: Macmillan.

Mackie, T. and R. Rose (1997), 'A decade of election results: updating the international almanac', *Studies in Public Policy*, Centre for the Study of Public Policy, University of Strathclyde, Glasgow, Scotland.

McCombie, J.S.L. and M. Roberts (2002), 'The role of the balance of payments in economic growth', in M. Setterfield (ed.), *The Economics of Demand-led Growth*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, Chapter 6.

OECD (1997), *Employment Outlook*, Paris: OECD.

Setterfield, M. (ed.) (2002), *The Economics of Demand-led Growth*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

9. The relevance of the Cambridge–Cambridge controversies in capital theory for econometric practice

G.C. Harcourt*

I

There is a well-known tale of the mathematician who used to burst into tears at the sight of the binomial theorem 'because it was so beautiful'. I have had occasion to remark that, for the same reason, economists at least get lumps in their throats at the sight of the Cobb–Douglas production function because it has such beautiful properties: the exponents of K and L measure the respective shares of profits and wages in the national income; the marginal products of K and L measure the return to capital and the wage-rate; the marginal products themselves relate in a very simple way – proportionally, where the factors of proportionality have clear economic meaning – to their respective average products.[1] Moreover, in growth theory the Cobb–Douglas allows simple measures of the contributions to growth in income per head of the respective growth in capital and labour. That is why, apart from Australian chauvinism/patriotism, I prefer Swan (1956) to Solow (1956). Trevor Swan's algebra and diagrams neatly exploit the above properties (and more) to illuminate the processes being analysed and, in particular, show why competitive markets give out stabilising signals which guide Harrod's warranted rate of growth towards equality with his natural rate of growth by affecting the choice of technique (as reflected in capital–output ratios). And when technical progress is included, as Solow (1957) showed, manipulation of the basic equations allows us to get simple measures of the respective contributions of deepening and (disembodied) technical progress to the growth in output per head over time.[2] Finally, the same function (or

* I thank, but in no way implicate, Nigel Allington, Stephanie Blankenburg, Robert Dixon, John McCombie, Tom Michl and the editors of the volume for their comments on a draft of the chapter.

its more sophisticated cousins, such as the constant-elasticity-of-substitution (CES) production function) have provided yeoman service in endogenous growth theory in recent years.

II

In this chapter I want to concentrate on another aspect of the Cobb–Douglas (and cousins) when used in applied, econometric work. It arises from the implicit assumption in much econometric specification that the short period and the long period may be collapsed into one. Then, for certain forms of the aggregate production function, exactly the same values of key parameters (and therefore variables) are involved, whether we are considering greater or lesser utilisation of a given stock of capital goods in the short run, that is, movements up or down what Joan Robinson called the utilisation function, or changing capital–labour and capital–output ratios as the result of differential rates of growth of accumulation and the labour force over time; in the latter process, there may not only be more capital per head and per unit of output but also better capital per head and per unit of output. Such a specification, allied with the assumptions of competitive market structures in the economy concerned and static expectations about the future courses of the prices of products and of the factors of production (so that the simple marginal productivity implications of cost-minimising and profit-maximising may go through) allows the use of actual 'real world' statistics on wages, profits, capital and so on when fitting the specified model. This in turn allows the estimation of key parameters, for example, the exponents of the variables of the function, the elasticity of substitution of capital for labour; and so on.

This methodological point was the basis of the criticism by Joan Robinson (1964) of, for example, Solow's procedures in his 1963 de Vries Lectures (Solow, 1963). There, for much of the book, he used what she called a butter model in the theoretical sections and in the specifications of his empirical work. (The main objective of his lectures was to develop theoretical measures of the Fisherian social rate of return on investment in a number of different scenarios. He treated it as a technocratic measure – the potential return to a bit more saving/investment at full employment. He estimated its values in what was then West Germany, and in the USA. As the resulting values were considerably greater than those of near riskless returns on certain financial assets, the inference was that more investment should be encouraged in both countries.)

In the model, butter was both input (B) and output (B') and the parameters of the model were usually functions of key *ratios* only, B'/L and

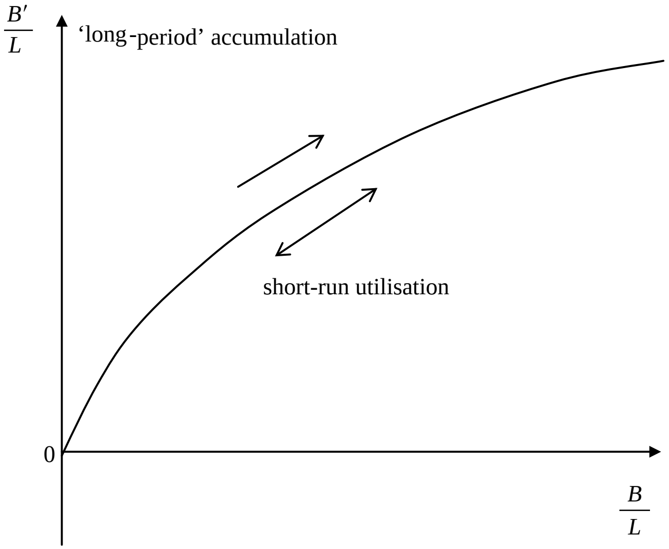


Figure 9.1 Short-period utilisation possibilities doubling up for long-period accumulation possibilities

B/*L*, where *L* was the potential work force. Ignoring technical progress for the moment, it did not matter whether the thought experiment was concerned with running up and down the short-period utilisation function with varying values of *B*′/*L* and *B*/*L*, or whether the changes in the values of *B*′/*L* and *B*/*L* were due to accumulation over 'time' so that *B*′/*L* was taken to be increasing as deepening occurred – 'moving down the production function' as Joan Robinson (1959; 1960) once put it.

As I argued above, the 'real world' observations are, by definition, observed points on the existing utilisation functions of each instant of time, since, though in the long run we are all dead, the living are always to be found in the short run. Nevertheless, they were meant to serve as well as observations of long-period values taken from, in effect, the same production function, see Figure 9.1.

III

We do not have to go into the intricacies of the capital-reversal and reswitching debates and results (see, for example, Cohen and Harcourt, 2003; Harcourt, 1972; 1995) in order to criticise the conceptual basis of this standard procedure.[3] In 1963 Robin Matthews, who was then review editor of the *Economic Journal*, asked me to review Bagicha Minhas's 1963 book, *An International Comparison of Factor Costs and Factor Uses* in which he exploited the properties of the famous CES production function, which came from an article written by Minhas jointly with Arrow, Chenery and Solow (1961) (hereafter referred to as ACMS). (I don't suppose many PhD students have such illustrious research assistants these days.) The previous year I had published a review article – see Harcourt (1962; 1982) – of the late Wilfred Salter's classic, *Productivity and Technical Change* (1960). Salter's book (which grew out of his early 1950s Cambridge PhD dissertation, supervised by Brian Reddaway) was a pioneering account of vintage (putty-clay) models and their application at firm and industry levels. As a result of what I learnt from Salter then (I still learn today), I argued in the review of Minhas – see Harcourt (1964) – that though the data used in Minhas's study came, of necessity, from existing short-run utilisation functions incorporating stocks of existing capital goods of different vintages associated with past accumulation, it was being used to estimate the values of the characteristics of what Salter called the iso-quants of the latest 'best-practice' techniques. These were, of course, the most up-to-date ways known in various industries of producing different levels of output (or output per unit of input if we assume that the *ex ante* production function – Salter's iso-quant – exhibits constant returns to scale, defined by an iso-quant in $Q/I(=l)$ and $Q/L(=i)$ space, see Figure 9.2). As a result of the choice of technique in each short run, the additions through accumulation at the margins of the stock of existing capital goods reflect the then 'optimum' point on the iso-quant.

Minhas and his co-authors were interested in a number of theoretical and empirical possibilities. Paul Samuelson (1948) had shown for the case of two countries which produce the same two commodities, use the same factors of production and have the same production functions in each industry, but different factor endowments, that free trade will equalise their absolute and relative factor prices. He assumed constant returns to scale and that, at any given ratio of factor prices, the chosen ratio in one industry is always greater or less than the corresponding ratio in the other. Minhas *et al.* showed that if the two commodities are produced with two CES production functions that have different elasticities of substitution of capital for labour, there will always be a critical ratio of factor prices at which their factor intensities are equal, and above (or below) are reversed, requiring, for this case, modifications of Samuelson's factor–price equalisation theorem. Minhas was concerned in his book to fit relationships derived from the CES production function to observed data that came from the same industries in different countries. He wanted to estimate the values of the elasticities and to see whether factor reversals occurred within the

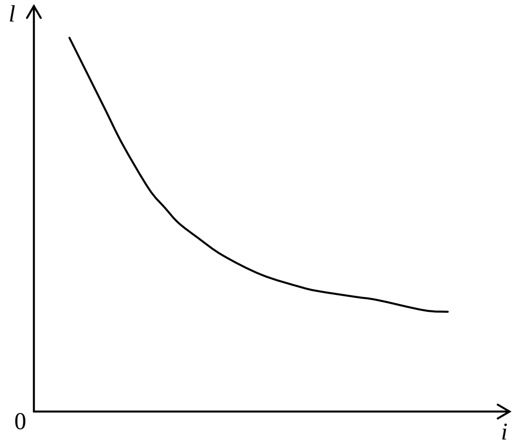


Figure 9.2 Salter's 'best-practice' iso-quant, assuming constant returns to scale

observed range of factor prices. He purported to show: that the CES production functions fitted the data well (if it is assumed that the efficiency of factors used between countries differed neutrally); that the elasticities were usually significantly less than unity (bye-bye, Cobb–Douglas); and that the critical price ratio was within the observed range of factor prices. For our present purposes we note that the 'real world' data were interpreted as points around the 'best-practice' iso-quant in each industry in different countries. The short period and the long period had again been collapsed into one another, where by long period, I mean the choices available at any moment of time for investment in 'best-practice' techniques, that is, the choice is made in the short period, but long-period factors are its dominant determinants.

IV

I followed the review with an article, Harcourt (1966), in which I said in effect: let us grant neoclassical economists every assumption they make in these investigations (I had ACMS and Minhas especially in mind), except that we allow for different vintages of 'best-practice' techniques to have been embodied by past bursts of accumulation into the total stocks of capital goods of the utilisation functions, which directly or indirectly had thrown up the data used by Minhas *et al.* in their estimates of the values of the elasticities of substitution. Will the equations they fitted to such data be 'good' fits, that is, will they provide unbiased estimators of the elasticities of substitution of the 'best-practice' iso-quants, which is their claim? ACMS found a close association between the logarithms of labour productivity (value added per unit of labour used) and money-wage rates in the *same* industries in *different* countries, which was confirmed by the appropriate regressions. If the values added and labour inputs used in their analysis are assumed to be observations from CES production functions, the regression coefficients, say b, in equations of the form:

$$\log.\, q = \log \cdot A + b \log.\, w + \varepsilon,$$

where $q =$ value added per unit of labour, $w =$ money-wage rate and $\varepsilon =$ error term, can be shown to be estimates of the elasticity of substitution of capital for labour (see ACMS, 1961, 228–9; Minhas, 1963; Harcourt, 1972, 51–55). But do the estimates of b provide what is claimed for them? The answer is 'no' as I believe I established in the article, and which I think Solow (1997), in so far as I understand him, accepts. Having argued that all we ever have in the data they used are totals and averages, whereas we are really interested in relationships between marginal quantities, I made up a number of plausible (I hope) stories – Solow has his doubts – and examined how close, qualitatively, the estimates of b would be approached by the use of ACMS's procedures. I then put quantitative orders of magnitude on the biases by using Minhas's data and assuming that some of my stories had generated the data. I found biases both upwards and downwards, of considerable size, relatively to what was known to be their 'true' values.[4]

I shall not discuss the intricacies of the arguments between Anwar Shaikh (1974; 1980) and Solow (1974), because the other contributors to this volume have written extensively on it elsewhere and now here. But I do want to emphasise again the up-frontness of Solow's account of his procedure in his 1957 article. He wrote: 'It merely shows how one goes about interpreting time series if one starts by assuming that they were generated from a production function and that the competitive marginal product relations apply' (Solow, 1974, 121).

So he is not arguing that the world is Cobb–Douglas or CES or . . ., only that if we view our observations *as if* they were observations thrown up by Cobb–Douglas *et al.*, these are the orders of magnitude of the parameters which our econometric procedures allow us to estimate (this is where the Fisher, Filipe, Shaikh, Solow, and McCombie debates begin). Solow does add that if the findings implied that the share of wages was 25 per cent and of profits 75 per cent, he would be less willing to trust his findings.[5]

V

Finally, I want to discuss what in one sense is the heart of the matter, already touched upon in the claim that the short period and the long period have been collapsed into one. In the Cambridge–Cambridge capital theory debates there has been much discussion about the significance and relevance of the results, especially the phenomena of capital-reversal and reswitching, for economic theory and practice. One claim is that the answers to these queries is essentially an empirical one. Two champions of this view are the late Charles Ferguson and Mark Blaug. Here is Ferguson's most (in)famous quote on the Cambridge critique in general and these issues in particular. It comes from his 1969 book.

> [The validity of the Cambridge, England, criticism of neoclassical theory] is unquestionable, but its importance is an empirical or an econometric matter that depends upon the amount of substitutability there is in the system. Until the econometricians have the answer for us, placing reliance upon neoclassical economic theory is a matter of faith. I personally have the faith; but at present the best I can do is to invoke the weight of Samuelson's authority as represented, for example, by the fly-leaf quotation [in Ferguson's book] (Ferguson, 1969, xvii–xviii).[6]

Blaug, as well as defending the faith, has also often asked what is the likelihood of capital-reversal and reswitching occurring in practice? Joan Robinson especially has argued that this is a nonsense question. The phenomena are discussed in theoretical terms, using either *comparisons* of long-period equilibrium stationary states or of steady-state equilibrium growth models. They are therefore concerned with differences – what is the long-period equilibrium stationary state associated with a given value of one of the distributive variables, either *w* or *r*, for a given set of possibilities now? The set may be either a smooth, continuously substitutable, neoclassical production function or an MIT *et al.* book of blueprints, with a different set of techniques for producing each commodity on each page, with the values of the coefficients – inputs per unit of output – in each industry 'changing' discretely from one page to another. (Joan Robinson's use of this apparatus in her 1953–54 article led Solow (1955–56, 106) to quip: 'Everyone who invents linear programming these days seems charmed by it'.) So no process of accumulation in actual historical time is being considered. Moreover, the book of blueprints (or the neoclassical production function, the equivalent of Salter's iso-quant at the level of the economy as a whole) could be expected to change over time due to technical advances, so that time series observations are *at best* taken from a particular point on the *ex ante* production function on a particular page of the newest, latest book of blueprints. There is thus no way this information may be used to test whether capital-reversal or reswitching is contained in any one book of blueprints.

So we are really concerned with a doctrinal debate concerning the coherence of neoclassical intuitions about the characteristics and functions of prices (and their relationship to the underlying scarcity theory of value) at a very abstract level of 'high theory' (no doubt child's play for the heavies accustomed to publishing in *Econometrica*, *QJE*, *Review of Economic Studies*, *JET*, and so on[7]). We are reminded here of Piero Sraffa's account of the different criteria that theory and statistical practice have to meet.

> '[O]ne should emphasise the distinction between two types of measurement . . . the one in which the statisticians were mainly interested. Second . . . measurement in theory. The statisticians' measures were only approximate . . . the theoretical measures required absolute precision . . . If we found contradictions, . . . these pointed to defects in the theory. (Sraffa, 1961, 305).

Just as Ricardo was chasing a Will-o'-the-wisp when he searched for an invariable standard of value which would allow him to precipitate out the effects of changes in distribution *and* technical progress from a measure of the surplus available for accumulation over *time*, so, too, is Blaug's and Ferguson's search for empirical findings to provide answers to their questions a similar chase, they are just not there to be found. Stiff cheddar, but there it is.

VI

Or is it? I once suggested an alternative approach, see Harcourt (1966, 233; 1982, 145), which led one of the ACMS gang of four to query whether I was fooling/kidding. It was the following: an article by Salter (1962), published after his tragically early death, was concerned with an empirical enquiry using a questionnaire to ask businesspeople in different industries how much and what type of investment they would do if they wanted to increase their present capacities by a given amount. My suggestion was to test the ACMS hypothesis by asking this question of businesspeople in the same industries but different economies with different actual (and expected) relative factor prices. The resulting observations, on ACMS's methodology, could be points on Salter iso-quants. If the same techniques were found to be associated with widely different relative factor prices (with other techniques chosen in between), that would be some evidence of the empirical presence of reswitching in the book of blueprints of the current 'best-practice' techniques in a situation in which it would be sensible in principle to test for its presence empirically.[8]

What is to be done? Must we despair, in the light of the implications of the capital theory controversies, concerning the possibility of doing useful work using econometric techniques? I do not believe so, though I do think Joan Robinson's critique bears on the underlying conceptual foundations of much current econometric work. Basically, the world is still viewed in a Marshallian, even Pigovian manner (when it is not being viewed as the outcome of the decisions of Frank Ramsey's benevolent dictator). There is a stable (?) long-period equilibrium position 'out there' which both constrains and guides short-run movements as though it were a powerful magnet holding them in check, drawing the short-run values of prices and quantities toward itself and its own corresponding long-period values (or, at least, making the former fluctuate around the latter). So actual observations may be interpreted as coming from (and approximating to) short-period flow equilibrium values, each corresponding to a station on the way to the long-period equilibrium cross.[9] Now certainly this is the structure of Marshall's *theory*, but he never claimed that it was even an approximate description of the world. He was describing tendencies towards long-period equilibrium, providing that none of the background fundamentals of the initial situation were allowed to change once the process of observing short-period equilibrium flows was started (theoretically, of course). He made it absolutely clear that in actual historical time, the vital components of the initial position could change, especially knowledge of the best ways of doing things.[10] This is where Salter comes into the discussion, arguing that, *analytically*, it is reasonable to suppose that the arrival of new ideas may be treated as if they arrive discretely, so that the accumulation process may embody the current 'best-practice' ideas through investment at the margin of the existing stock of different vintages. This is a far smaller order of abstraction to have to accept than what I take to be implied in cointegration procedures.

But suppose we follow another tack, using the seminal ideas of Richard Goodwin[11] and Michal Kalecki concerning the indissolubility of trend and cycle. Kalecki's succinct (as ever) statement of the approach is the following: 'In fact, the long-run period is only a slowly changing component of a chain of short-period situations; it has no independent entity' (Kalecki, 1968; 1991, 435).[12]

An implication for theory of the indissolubility of trend and cycle is that the separation of the factors responsible for the existence (uniqueness or multiple) of equilibrium from those responsible for *stability* (local and global) is unacceptable, an insight becoming more and more recognised by the mainstream with the examination in recent years of path-dependent models (already signalled by Nicky Kaldor in 1934 and Joan Robinson in 1953, probably earlier). This is, of course, matched by Kalecki's work and

Goodwin's 1967 classic, 'A growth cycle'. So the role for Classical/Marxist centres of gravitation[13] in theory may be adjusted to the conjecture that actual observations may be regarded as near enough to those associated with short-period macroeconomic rest states[14] to allow econometric methods to be used to fit them in, say, time series analysis. Here I leave it for the Andrew Harveys of this world to take over.[15]

APPENDIX 1 THE MAIN ISSUES AND RESULTS OF SOME CAMBRIDGE CONTROVERSIES IN THE THEORY OF CAPITAL

The debates between the two Cambridges (England and Mass.) occurred principally between the 1950s and the 1970s. They started with Joan Robinson's 1953–54 article 'The production function and the theory of capital', and really hotted up with the publication in 1960 of Piero Sraffa's classic, *Production of Commodities by Means of Commodities*. They 'ended' with the publication of Christopher Bliss's 1975 volume, *Capital Theory and the Distribution of Income*, as a result of which Avinash Dixit (1977) pronounced the quasi-rents of previous writings on the issues to be either zero or, in the case of the Cambridge, England, protagonists, negative. That the 'end' may have been prematurely dated is argued by Harcourt (1995) and Cohen and Harcourt (2003).

With hindsight, we may say that the issues related not so much to the *measurement* of capital as to its *meaning*. This carried with it further questions about how the accumulation process in capitalist society may best be envisaged and so modelled. There are two principal competitors: On the one hand, Marx–Keynes–Schumpeterian ruthless, swashbuckling entrepreneurs and capitalists, for whom profit-making and accumulation are ends in themselves, call the tune to which all other classes in society must dance. On the other hand, the consumption and saving behaviour of lifetime utility-maximising agents dominates and all other actors and institutions in the economy, firms, the stock exchange, for example, are but the agents through which they achieve their ends. To both views must be coupled the question: what is the appropriate method with which to analyse the processes of accumulation, distribution and growth?

The first question posed historically was: can we find a technical unit in which to measure capital that is independent of distribution and prices? For, if we are to use a demand and supply approach to explain the origins and sizes of the distributive variables – the rate of profits (r), the wage rate (w) – and distributive shares; if we are to make explicit the intuition of the supply and demand approach that price is an index of scarcity; and if we accept that in a competitive situation there is a tendency to equality of rates of profit in all activities so that we have to explain the origin and size of the overall, economy-wide *r*; *then* we need to know before the analysis starts what we mean by a quantity of capital in order that it may be a determinant of *r* (an exogenous, given variable), and one of the reasons why *r* may be high or low relative to *w* is that we have a 'little' or a 'lot' of capital. If it is not possible to find such a unit (the debates showed that outside one commodity models, it is not), it is not possible to say *r* takes the value it does partly because we have so much 'capital' and because 'its' marginal product has a particular value.

This aspect of the debate was related to a methodological critique associated with the distinction between differences and changes. The results of the debate were mostly drawn from comparisons of long-period positions, which reflect differences in initial conditions. It is argued that they can tell us nothing about processes – changes – in particular, the processes of accumulation. Joan Robinson (1979 [1974]) was to characterise this critique as 'history versus equilibrium'. Its implications are reflected in the discussion in the text about the short period and the long period.

The reaction to the criticism of the aggregate production function and the meaning of 'capital' in its construction was to try to avoid the use of 'capital' and 'its' marginal product and make the social rate of return – Irving Fisher's central concept – a key concept. It provides on the productivity side of the story what the rate of time preference does on the psychological side; see Solow (1963).

Parallel with this development was Paul Samuelson's attempt (1962) to rationalise Solow's use of J.B. Clark–Frank Ramsey–J.R. Hicks models in growth theory and econometric work; see Solow (1956; 1957). Samuelson attempted to show that the rigorously derived results of the simple model were robust, and that they illuminated the behaviour of more complex heterogeneous capital models. Lying behind all this was the conceptual understanding that 'capital' and *r* are related in such a way that the demand curve for 'capital' is well-behaved, that is, downward sloping. This result as well as other neoclassical parables derived from the simple model – the negative associations between *r* and the capital–output ratio and sustainable levels of consumption per head – together with the marginal productivity theory of distribution itself were refuted by the capital-reversing and reswitching results, as Samuelson (1966b) handsomely acknowledged. Capital-reversing (the Ruth Cohen curiosum) is that a *less* productive, *less* capital-intensive technique may be associated with a *lower* value of *r*. The reswitching result is that the same technique, having been the most profitable one for a particular range of values of *r* and *w*, could also be most profitable at another range of values of *r* and *w*, even though other techniques were profitable at values in between. Both refute the agreeable (neoclassical) intuition of the results of the simple models and, Pasinetti (1969; 1970) argued, of Solow's Fisherian approach (Solow, 1963) as well. Solow (1970) did not agree.

APPENDIX 2

The essential methodology of ACMS is as follows: consider the production function

$$Q = F(K, L) \tag{1a}$$

which, because of constant returns to scale, may be written as

$$\frac{Q}{L} = f\left(\frac{K}{L}, 1\right) \tag{1b}$$

that is,

$$q = f(k) \tag{2}$$

where $q = Q/L, k = K/L$

Now

$$\begin{aligned} \frac{\partial Q}{\partial K} &= f'(k) \\ \frac{\partial Q}{\partial L} &= f(k) - f'(k)k \end{aligned} \tag{3}$$

and, assuming perfect competition and static expectations,

$$w = f(k) = f'(k)k \tag{4}$$

Equation (4) has an inverse function that relates k to w and, because $q = f(k)$, it also allows q to relate to w, say

$$q = g^*(w) \tag{5}$$

ACMS turn this procedure around and suppose that the *form* of the relationship between productivity and the wage rate is known. Let it be

$$q = g^*(w) \tag{6}$$

that is, expression (6) is the general form of the regression equation in the text above. Then, with their assumptions,

$$q = g^*(q - f(k)k) \tag{7}$$

which is a differential equation for *f(k)* with a solution

$$q = f(k; \bar{A}) \tag{8}$$

where $\bar{A}$ is a constant of integration. Note that equation (8) is constrained to make $f''(k) > 0$ and $f'(k) < 0$.

We next show that *if* there is this link from the productivity–wage rate relationship to the production function, the regression coefficient, *b,* is not only the elasticity of productivity with respect of the wage rate, $(\mathrm{d}q/\mathrm{d}w)(w/q)$, but also the elasticity of substitution of capital for labour, σ. σ measures the responsiveness of the capital–labour ratio to changes in the ratios of the marginal products of capital and labour, and, therefore, *with perfect competition and static expectations*, to changes in relative factor prices. With constant returns to scale, σ may be defined as follows:

$$\sigma = \frac{(\partial Q/\partial K)(\partial Q/\partial L)}{Q(\partial^2 Q/\partial K \partial L)} \tag{9}$$

We already have expressions for $\partial Q/\partial K$ and $\partial Q/\partial L$: see expression (3) above. $Q(\partial^2 Q/\partial K \partial L)$may be shown to be: $(1/L)(-kf''(k))$, and we know that $Q = Lf(k)$.[16]

Substituting these expressions in expression (9) gives

$$\sigma = -\frac{f'(k)(f(k) - f'(k)k)}{kf(k)f''(k)} \tag{10}$$

Now

$$w = f(k) - f'(k)k \tag{4}$$

so that

$$\begin{aligned} \mathrm{d}w &= f'(k)\mathrm{d}k - f'(k)\mathrm{d}k - kf''(k)\mathrm{d}k \\ &= -kf''(k)\mathrm{d}k \end{aligned} \tag{11}$$

From $q = f(k)$, we obtain

$$\mathrm{d}q = f'(k)\mathrm{d}k$$

and, thus

$$\mathrm{d}k = \frac{1}{f'(k)}\mathrm{d}q \qquad (12)$$

Substituting expression (12) in expression (11), we obtain

$$\mathrm{d}w = -kf''(k)\frac{\mathrm{d}q}{f'(k)}$$

and so

$$\frac{\mathrm{d}q}{\mathrm{d}w} = -\frac{f'(k)}{kf''(k)} \qquad (13)$$

Therefore

$$\frac{\mathrm{d}q}{\mathrm{d}w}\frac{w}{q} = -\frac{f'(k)(f(k) - f'(k)k)}{kf(k)f''(k)} = \sigma \qquad (14)$$

But ($\mathrm{d}q/\mathrm{d}w$) (w/q) is, by definition, b, so that b is an estimate of σ.

NOTES

1. Write $Q = L^{\alpha}K^{\beta}$ (1)

   where $\alpha + \beta = 1$, $Q =$ output (income), $L =$ labour and $K =$ capital.

   Thus $Q = L^{1-\beta}K^{\beta}$ (2)

   and $\frac{\delta Q}{\delta K} = \beta L^{1-\beta}K^{\beta-1}$ (3)

   But $Q/K = L^{1-\beta}K^{\beta-1}$

   so that $\frac{\delta Q}{\delta K} = \beta\frac{Q}{K}$ (4)

   and the share of capital in output (income) is

   $$w_k = \frac{rK}{Q} = \frac{\delta Q/\delta K \cdot K}{Q} \qquad (5)$$

   But $r = \delta Q/\delta K$, so that $w_k = \beta$ (6)

   As β is the ratio of the marginal to the average product of capital, it is the *elasticity* of output with respect to capital. Similarly, it may be shown that $w_1 = a$.
2. Write

   $$Q = A(t) f(K, L) \qquad (1)$$

   Where $A(t)$ is a shift factor reflecting the pull of all forces of technical change. Differentiating (1) with respect to time, we obtain

$$\dot{Q} = A(t)f(K, L) + A(t)\frac{\delta f}{\delta K}\dot{K} + A(t)\frac{\delta f}{\delta L}\dot{L} \tag{2}$$

where dots over variables indicate derivatives with respect to time.
Dividing by Q, we get

$$\frac{\dot{Q}}{Q} = \frac{\dot{A}(t)}{A(t)} + A(t)\frac{\delta f}{\delta K}\frac{\dot{K}}{Q} + A(t)\frac{\delta f}{\delta L}\frac{\dot{L}}{Q} \tag{3}$$

Now

$$A(t)\frac{\delta f}{\delta K} = \frac{\delta Q}{\delta K}$$

and

$$A(t)\frac{\delta f}{\delta L} = \frac{\delta Q}{\delta L}$$

so that if factors are paid their marginal products

$$\frac{\delta Q}{\delta K}\frac{K}{Q} = w_k$$

which is capital's share in product.
So we may write (3) as

$$\frac{\dot{Q}}{Q} = \frac{\dot{A}(t)}{A(t)} + w_k\frac{\dot{K}}{K} + w_l\frac{\dot{L}}{L}$$

which, if $w_k + w_l = 1$, becomes

$$\frac{\dot{Q}}{Q} - \frac{\dot{L}}{L} = \frac{\dot{A}(t)}{A(t)} + w_k\left\{\frac{\dot{K}}{K} - \frac{\dot{L}}{L}\right\} \tag{4}$$

or

$$q^* = a + w_k\, k^*, \tag{5}$$

With Cobb–Douglas, we may write (5) as

$$q^* = a + \beta\, k^* \tag{6}$$

Growth in output per head equals the rate of growth of the shift factor ('technical progress') plus the rate of growth of capital per man ('deepening') with the latter weighted by capital's share:

$$a = q^* - w_k\, k^* \tag{7}$$

and to estimate a we only have to obtain statistics on q^*, w_k and k^*; these either exist or may be constructed.

[3] See Appendix 1 for a brief account of the debates and the main results.

[4] The ability to play God in the computer age that the analysis in this article allowed me to do led me to succumb to temptation again, this time in the company of two Adelaide colleagues, the late Peter Praetz and Al Watson. As a Trinity we made up a number of worlds of our own in which we knew the true values of key parameters. We then generated, by 'Monte Carlo' experiments, observations from these worlds and tested whether the econometric procedures used by our friends, the late Fred Gruen and Alan Powell, to estimate supply responses in Australian agriculture in fact gave unbiased and accurate estimates of the values of the parameters they were interested in (their values/sizes had important implications for policy). Alas, according to us, they did not, though another 'mate', Ray Byron, argued that our procedures were as shonky as those of Powell and Gruen's but for different reasons! It was all good clean fun, of course – see Gruen *et al.* 1967a, 1967b; Powell and Gruen 1966a, 1966b, 1967, 1968, 1970; Watson *et al.*, 1970a, 1970b; Byron 1970a, 1970b.

5. Nevertheless, as Thomas Michl (2002, 53, 29/5/02, letter to G.C. Harcourt) reminded me (in this case, a euphemism for bringing it to my attention for the first time), 'Empirical research guided by the neoclassical growth model has consistently found that the apparent elasticity of output with respect to capital exceeds its predicted value, typically taken to be the share of profit in national income'.
6. Samuelson (1966a, 444–5) wrote: 'Until the laws of thermodynamics are repealed, I shall continue . . . to believe in production functions . . . a many-sectored neoclassical model with heterogeneous capital goods and somewhat limited factor substitutions can fail to have some of the simple properties of the idealised J.B. Clark neoclassical models. Recognising these complications does not justify nihilism'.
7. I think it is at this point that Franklin Fisher and I may still part company, see his comment on this in Fisher (2005).
8. Thomas Michl (29/5/02; Letter to G.C. Harcourt) thought this a useful way forward and suggested that engineers could usefully be asked too. His own conjecture 'is we could find the best-practice firms are already working with the most automated technique'.
9. See, for example, Granger (1993). The link between Granger's article and what is in the text may be more clear to me than to my readers. I heard Granger give the paper at the 1992 RES conference and we discussed the theory underlying it afterwards.
10. Dennis Robertson (1956, 16, emphasis in original) mentions the two concepts of the long period Marshall had in mind: 'one in which it stands realistically for any period in which there is time for *substantial* alterations to be made to the size of the plant, and one in which it stands conceptually for the Never-never land of unrealised tendency'.
11. It is significant that Anwar Shaikh (2005) makes use of Goodwin's classic 1967 prey-predator Marxian model in his analysis.
12. The following statement by Joan Robinson (1962; 1965, 100) is a theoretical counterpart of Kalecki's insight: 'The short period is here and now, with concrete stocks of means of production in existence. Incompatibilities in the situation . . . will determine what happens next. Long-period equilibrium is not at some date in the future; it is an imaginary state of affairs in which there are no incompatibilities in the existing situation, here and now'.
13. At least four different versions of the concept of centres of gravitation (c.g.) may be identified. Three are analogies drawn from physics, the fourth, from meteorology. The first relates to a frictionless pendulum always swinging, but always passing the same minimum point on its path to and fro. The second is a pendulum, the motion of which eventually ends because of friction. It settles at the minimum point of its path, which is also the c.g. of the first version. The third (due to the late David Champernowne) concerns a dog running towards its master, who is riding a bike. The bike is the c.g., always moving. The dog's direction of movement at any point of time may be predicted by the knowledge of where the bike (master) is at that point of time. The fourth is akin to average temperatures (of seasons and places), which are predicted by their relationships to the average values of relevant factors. The averages are good predictors over time but not day by day; see Harcourt (1982, 205–21).
14. Nor is the existence of short-period rest states inconsistent with the view of the world associated with Allyn Young (1928), Kaldor and Myrdal of virtuous and vile cumulative causation processes, as opposed to the stable long-period equilibrium positions of, say, Chicago mainstreamers.
15. Foley and Michl (1999) have introduced the concept of the fossil production function. It allows the econometric application of the Cobb–Douglas function to be salvaged by giving a different meaning to the economic processes producing the data. Taking their lead from Ricardo, Joan Robinson and Kaldor, they suggest that capital-using technical progress occurs over time, with only one best-practice technique being available at any moment of time. Empirical observations, either cross-section or time series, reflect current and past vintages, with average productivity rising over time. Michl (2002) shows that this procedure helps to make sense of evidence taken from OECD countries, and of the findings of discrepancies between the values of capital–output elasticities and profit shares.

16. This can be shown as follows:

$$\begin{aligned}\frac{\partial^2 Q}{\partial K \partial L} &= \frac{\partial}{\partial K}\left(\frac{\partial Q}{\partial L}\right)\\ &= \frac{\partial}{\partial K}(f(k) - f'(k)k)\\ &= \frac{\partial k}{\partial K}\left(\frac{\mathrm{d}}{\mathrm{d}k}(f(k) - f'(k)k)\right)\\ &= \frac{1}{L}(-kf''(k))\end{aligned}$$

REFERENCES

Arrow, K.J., H.B. Chenery, B.S. Minhas and R.M. Solow (1961), 'Capital-labor substitution and economic efficiency', *Review of Economics and Statistics*, **43**, 225–50.

Bliss, C.J. (1975), *Capital Theory and the Distribution of Income*, Amsterdam: North-Holland.

Byron, R.P. (1970a), 'The bias in the Watson–Harcourt–Praetz variant of the C.E.T. production frontier', *Economic Record*, **46**, 567–73.

Byron, R.P. (1970b), 'A reply to the Trinity', *Economic Record*, **46**, 576–77.

Cohen, A.J. and G.C. Harcourt (2003), 'Whatever happened to the Cambridge capital theory controversies?', *Journal of Economic Perspectives*, **17**, 199–214.

Dixit, A. (1977), 'The accumulation of capital theory', *Oxford Economic Papers*, **29**, 1–29.

Ferguson, C.E. (1969), *The Neoclassical Theory of Production and Distribution*, Cambridge: Cambridge University Press.

Fisher, F.M. (2005), 'Aggregate production functions – a pervasive, but unpersuasive, fairytale', *Eastern Economic Journal*, **31**, 489–91.

Foley, D.K. and T.R. Michl (1999), *Growth and Distribution*, Cambridge, MA: Harvard University Press.

Goodwin, R.M. (1967), 'A growth cycle', in C.H. Feinstein (ed.), *Socialism, Capitalism and Economic Growth. Essays presented to Maurice Dobb*, Cambridge: Cambridge University Press, pp. 54–8.

Granger, C.W.J. (1993), 'What are we learning about the long-run?', *Economic Journal*, **103**, 307–17.

Gruen, F.H., L.E. Ward and A.A. Powell (1967a), 'Changes in the supply of agricultural products', in D.B. Williams (ed.), *Agriculture in the Australian Economy*, Sydney: Sydney University Press.

Gruen, F.H. and others (1967b), *Long Term Projections of Agricultural Supply and Demand: Australia 1965 to 1980*, Department of Economics, Monash University.

Harcourt, G.C. (1962), 'Review article of W.E.G. Salter, *Productivity and Technical Change* (1960)', *Economic Record*, **38**, 388–94; reprinted in Harcourt (1982).

Harcourt, G.C. (1964), 'Review of B.S. Minhas, *An International Comparison of Factor Costs and Factor Use* (1963)', *Economic Journal*, **74**, 443–5.

Harcourt, G.C. (1966), 'Biases in empirical estimates of the elasticities of substitution of C.E.S. production functions', *Review of Economic Studies*, **33**, 227–33.

Harcourt, G.C. (1972), *Some Cambridge Controversies in the Theory of Capital*, Cambridge: Cambridge University Press.

Harcourt, G.C. (1982), *The Social Science Imperialists: Selected Essays of G.C. Harcourt*, edited by Prue Kerr, London: Routledge and Kegan Paul.

Harcourt, G.C. (1995), 'The capital theory controversies', in G.C. Harcourt, *Socialism, Capitalism and Post-Keynesianism: Selected Essays of G.C. Harcourt*, Aldershot, UK and Brookfield, US: Edward Elgar, pp. 41–46.

Kaldor, N. (1934), 'A classificatory note on the determinateness of equilibrium', *Review of Economic Studies*, **1**, 122–36.

Kalecki, M. (1968), 'Trend and business cycle reconsidered', *Economic Journal*, **78**, 263–76; reprinted as 'Trend and the business cycle' (1991), in *Collected Works of Michal Kalecki*, Vol. II., edited by Jerzy Osiatyński, Oxford: Clarendon Press, pp. 435–52.

Michl, T.R. (2002), 'The fossil production function in a vintage model', *Australian Economic Papers*, **41**, 53–68.

Minhas, B.S. (1963), *An International Comparison of Factor Costs and Factor Use*, Amsterdam: North Holland.

Pasinetti, L.L. (1969), 'Switches of technique and the "rate of return" in capital theory', *Economic Journal*, **79**, 508–31.

Pasinetti, L.L. (1970), 'Again on capital theory and Solow's "rate of return"', *Economic Journal*, **80**, 428–31.

Powell, A.A. and F.H. Gruen (1966a), 'Problems in aggregate agricultural supply analysis: I. The construction of time series for analysis', *Review of Marketing and Agricultural Economics*, **34**, 112–35.

Powell, A.A. and F.H. Gruen (1966b), 'Problems in aggregate agricultural supply analysis: II. Preliminary results for cereals and wool', *Review of Marketing and Agricultural Economics*, **34**, 186–201.

Powell, A.A. and F.H. Gruen (1967), 'The estimation of production frontiers: the Australian livestock/cereals complex', *Australian Journal of Agricultural Economics*, **2**, 63–81.

Powell, A.A. and F.H. Gruen (1968), 'The constant elasticity of transformation production frontier and linear supply system', *International Economic Review*, **9**, 315–28.

Powell, A.A. and F.H. Gruen (1970), 'Biases in the estimation of transformation elasticities: a rebuttal', *Economic Record*, **46**, 564–66.

Robertson, D.H. (1956), *Economic Commentaries*, London: Staples Press.

Robinson, Joan (1953–54), 'The production function and the theory of capital', *Review of Economic Studies*, **21**, 81–106.

Robinson, Joan (1959), 'Accumulation and the production function', *Economic Journal*, **69**, 433–42.

Robinson, Joan (1960), *Collected Economic Papers*, **2**, Oxford: Basil Blackwell.

Robinson, Joan (1962), 'Review of H.G. Johnson, *Money, Trade and Economic Growth* (1962)', *Economic Journal*, **72**, 690–92.

Robinson, Joan (1964), 'Solow on the rate of return', *Economic Journal*, **74**, 410–17.

Robinson, Joan (1965), *Collected Economic Papers*, **3**, Oxford: Basil Blackwell.

Robinson, Joan (1979 [1974]), 'History versus equilibrium', Thames Papers in Political Economy, reprinted in *Collected Economic Papers*, **5**, Oxford: Basil Blackwell, pp. 48–58.

Salter, W.E.G. (1960), *Productivity and Technical Change*, Cambridge: Cambridge University Press, 2nd edition, 1966.

Salter, W.E.G. (1962), 'Marginal labour and investment coefficients of Australian manufacturing industry', *Economic Record*, **38**, 137–56.

Samuelson, P.A. (1948), 'International trade and the equalisation of factor prices', *Economic Journal*, **58**, 163–84.

Samuelson, P.A. (1962), 'Parable and realism in capital theory: the surrogate production function', *Review of Economic Studies*, **29**, 193–206.

Samuelson, P.A. (1966a), 'Rejoinder: agreements, disagreements, doubts, and the case of induced Harrod – neutral technical change', *Review of Economics and Statistics*, **98**, 444–8.

Samuelson, P.A. (1966b), 'A summing up', *Quarterly Journal of Economics*, **80**, 568–83.

Shaikh, A. (1974), 'Laws of production and laws of algebra: the Humbug production function: a comment', *Review of Economics and Statistics*, **56**, 115–20.

Shaikh, A. (1980), 'Laws of production and laws of algebra: Humbug II', in Edward J. Nell (ed.), *Growth, Profits and Property. Essays in the Revival of Political Economy*, Cambridge: Cambridge University Press, pp. 80–95.

Shaikh, A. (2005), 'Non-linear dynamics and pseudo-production functions', *Eastern Economic Journal*, **31**, 447–66.

Solow, R.M. (1955–56), 'The production function and the theory of capital', *Review of Economic Studies*, **23**, 101–8.

Solow, R.M. (1956), 'A contribution to the theory of economic growth', *Quarterly Journal of Economics*, **70**, 65–94.

Solow, R.M. (1957), 'Technical change and the aggregate production function', *Review of Economics and Statistics*, **39**, 312–20.

Solow, R.M. (1963), *Capital Theory and the Rate of Return*, Amsterdam: North-Holland.

Solow, R.M. (1970), 'On the rate of return: reply to Pasinetti', *Economic Journal*, **80**, 423–8.

Solow, R.M. (1974), 'Laws of production and laws of algebra: Humbug production function: a comment', *Review of Economics and Statistics*, **56**, 121.

Solow, R.M. (1997), 'Thoughts inspired by reading an atypical paper by Harcourt', in Philip Arestis, Gabriel Palma and Malcolm Sawyer (eds), *Capital Controversy, Post-Keynesian Economics and the History of Economic Thought: Essays in Honour of Geoff Harcourt, Volume One*, London and New York: Routledge, pp. 419–24.

Sraffa, P. (1960), *Production of Commodities by Means of Commodities. Prelude to a Critique of Economic Theory*, Cambridge: Cambridge University Press.

Sraffa, P. (1961), 'Comment', in F.A. Lutz and D.C. Hague (eds), *The Theory of Capital*, London: Macmillan, pp. 305–6.

Swan, T.W. (1956), 'Economic growth and capital accumulation', *Economic Record*, **32**, 334–61.

Watson, A.S., G.C. Harcourt and P.D. Praetz (1970a), 'The C.E.T. Production Frontier and estimates of supply response in Australian agriculture', *Economic Record*, **46**, 553–63.

Watson, A.S., G.C. Harcourt and P.D. Praetz (1970b), 'Reply to Powell and Gruen, and Byron', *Economic Record*, **46**, 574–75.

Young, A. (1928), 'Increasing returns and economic progress', *Economic Journal*, **38**, 527–42.

10. Foreign direct investment and productivity spillovers: a sceptical analysis of some OECD economies*

Carlos Rodríguez, Carmen Gomez and Jesus Ferreiro

1. INTRODUCTION

One of the main features of the current process of globalisation is the liberalization of foreign direct investement (FDI) flows. Many countries, mainly developing and transition economies, but also developed countries, have implemented active policies to attract as many FDI inflows as possible, for instance, by privatising previously state-owned firms. Closely related to an export-led growth strategy, FDI flows, that is, the presence in the local economy of subsidiaries of multinational enterprises (MNEs), have been considered as a key tool to accelerate economic growth.

One of the channels through which inward FDI can promote the economic growth in host economies is the existence and absorption of productivity spillovers. Our paper is an attempt to evaluate the existence of the size and direction of these externalities – a question, as we will see later, subject to a deep controversy. Although there exists in the literature a high number of papers related to this issue, the novelty of this paper is that we use a database not used in previous papers: the OECD database 'Measuring Globalisation: the Role of Multinationals in OECD Economies'. This database gives data about employment and value added for 21 OECD economies and, what is more relevant for our paper, data about the share of foreign firms in those variables. These data are available for the whole economy and also for some industries. We have used these data to calculate the productivity of foreign and local firms in the economies/industries for which data are available, and, therefore, we have estimated whether a relationship exists between the

* This work is within the remit of the research project financed by the University of the Basque Country (Project code: 1/UPV 00032.321-H-15295/2003).

evolution of the productivity gap between foreign and local firms (a proxy for the existence and direction of productivity spillovers) and the presence (and change) of foreign firms in the local economy/industry.

The paper is structured as follows: after this brief introduction, in the second part we will review the theoretical and empirical principal findings on productivity spillovers; in the third part we specify the econometric model and the data to analyse spillovers; in the fourth part we comment on the results, and then we conclude.

2. INWARD FDI AND ECONOMIC GROWTH: THE ROLE OF PRODUCTIVITY SPILLOVERS

International organizations have strongly recommended reliance on FDI as an engine for growth, mainly in the case of the host economies for these flows. The argument is that FDI is superior to other types of capital flows. It is the 'good cholesterol' that offers a number of advantages to the host economy: an additional supply of new funds for investment that are less volatile than other financial-portfolio capital resources, a direct increase in the productivity of capital stock, and, mainly, a bundle of important assets for growth in the form of access to modern technology and know-how. In this sense, specific attention to FDI has been devoted over and above other effects to the question of whether inward FDI does involve superior technology (that is, MNEs, both parent firms and their foreign subsidiaries, are more efficient in competitiveness and have higher productivity than local firms in foreign markets) and, if it does, whether it 'spills over' to domestically owned firms rather then being retained entirely by the foreign-owned firms.

In the optimistic view about the impact of inward FDI-MNEs on host economies, FDI inflows contribute to higher economic growth via: the increased supply of financial funds thanks to net FDI inflows, increased productive capacity thanks to greenfield investments, the higher and superior efficiency-productivity of foreign subsidiaries, and, finally, the technological-productivity spillovers generated by foreign subsidiaries that, once absorbed by local firms, increase the productivity of the latter firms.

Regarding the first question, we can say that it is a matter of fact that MNEs have a superior technology vis-à-vis local firms. In the FDI literature, technology is considered in a rather broad sense, encompassing the way firms organize the production processes, the marketing and the management functions. All the FDI theories, running from Hymer′s monopolistic approach to the eclectic paradigm of Dunning and the so-called new FDI theory from Markusen and Venables, accept that a firm, in order to do FDI and become a MNE, needs 'ownership advantages' (Ietto-Gilles, 2005).

The other question about spillovers is especially significant for the supporters of FDI and endogenous growth theory. FDI productivity spillovers, resulting from technological externalities,[1] occur when the entry, presence and operations of MNEs increase the productivity of domestic firms in a host country and MNEs do not fully internalise the value of these benefits. But there is no guarantee (either theoretical or empirical) that they should always be positive. They could be also non-existent or even negative. Actually, despite the conventional wisdom of the international organizations that there are positive externalities coming from FDI, the empirical evidence runs against this wisdom.

The existence of spillover effects is not a trivial issue. The presence of spillovers is a key requirement to consider FDI inflows as a tool to promote economic growth. We must keep in mind that if a MNE is more efficient-productive-competitive than a local firm, the presence of a foreign subsidiary can crowd out the less efficient local firms, and consequently the initial net positive effect on growth can be diminished. Therefore, the positive effects of FDI will be the highest when improvements in the productivity-efficiency of local firms are directly induced by the presence of foreign subsidiaries.[2]

3. A THEORETICAL SURVEY OF THE SPILLOVERS LITERATURE

Endogenous growth theories stress the importance of human capital accumulation and technological externalities in the growth process. These theories emphasize that FDI can increase the existing stock of knowledge in the recipient country through labour training, skill acquisition and diffusion, and the introduction of alternative management practices and organizational arrangements. In sum, FDI is expected to be a potential source of productivity gains via spillovers to local firms. In this sense, inward FDI raises economic growth by generating technological diffusion from the developed world to host countries. FDI offers know-how and technology (the most advanced production and organization methods) which are proprietary to the investors (the MNEs). Therefore, MNEs are seen as a natural and powerful vehicle of technology transfer to less developed economies (UNCTAD, 1992).

As De Mello (1997) stated in his survey about FDI and growth in developing countries, ‘whether FDI can be deemed to be a catalyst for output growth, capital accumulation and technological progress is a less controversial hypothesis in theory than in practice’. And nowadays, this conclusion seems to be truer. Navaretti and Venables (2004), in a recent book, alert us in the beginning of a brief survey that findings of cross-country growth regressions embedded in endogenous growth models, the usual way to check for spillovers, are quite mixed and rarely conclusive. Really, the empirical macroeconometric literature utilising cross-section, time series or panel data, provides opposite results, not only about the existence of a significant link between FDI and growth and the sign of this relationship, but also about the causal relationship between both variables. As Nunnenkamp and Spatz stated, 'these results are based on FDI flows which are not corrected for potential endogenity biases (i.e., higher economic growth causing higher FDI inflows)' (Nunnenkamp and Spatz, 2003, p. 4). Carkovic and Levine′s paper faces this problem (Carkovic and Levine, 2002). They find that there is no reliable cross-country empirical evidence supporting the claim that FDI per se (exogenously) accelerates economic growth, even when the above elements are controlled.

Despite the above stated, a question remains: are we overlooking positive impacts that occur at a micro level (industry or firm) that we can not see with macromodels using aggregated FDI data? There is also a lot of micro empirical work trying to prove the existence of spillovers, the term that encompasses all the external effects that MNEs may exert upon local firms in the host country.

In order to organize the analysis of spillovers, it is practical to separate them into those that may pass horizontally to local competitors (at an intraindustry level) or vertically to local suppliers or customers (interindustry level, up and down).

In the case of positive intraindustry spillovers, there are several channels through which these spillovers can be generated:

- By copying, imitation and demonstration effect (Wang and Blomström, 1992; Blomström and Kokko, 1996, Görg and Strobl, 2001). Demonstration effect involves exposure to the superior technology of MNC, possibly leading local firms to update their own production methods (Saggi, 2000).
- By worker mobility (Fosfuri *et al.*, 1998; Glass and Saggi, 1999).
- By access to foreign markets (Aitken *et al.*, 1997).
- By increased competition from foreign enterprises, improvement in resource allocation and reduction of X inefficiencies (Wang and Blomström, 1992)

However, negative or non-existent horizontal spillovers can also be generated:

- MNEs have an incentive to protect and avoid such spillovers (their ownership assets or advantages) to competitors by formal protection of their intellectual property through patents, trade secrecy, and so on (Smarzynska and Spatareanu, 2005).

- By efficiency wages: by paying higher wages than local firms, foreign subsidiaries can avoid spillovers making workers' mobility from foreign to local firms more difficult, and, besides, they can hire the best workers, meaning that local firms must hire the less productive workers (Globerman *et al.*, 1994).
- By competing with local firms, specially if the foreign affiliate is geared to domestic market, MNEs may reduce the market share for local firms (OECD, 2002).
- To take advantage of spillovers, there must be a relative (low technology gap between foreign and local firms) or absolute (enough education level or human capital) absorption capacity in the host country (Graham, 2000).

Potential positive spillovers can also be vertical ones: through backward and forward linkages with local suppliers and customers, foreign affiliates may generate productivity spillovers (Rodríguez, 1996). These spillovers more easily occur when local firms do not compete with foreign firms, and it may be in the interest of both parties to collaborate. The channels through which these positive vertical spillovers take place are the following ones:

- Suppliers may increase production scale, thanks to the demand from the affiliate firm.
- Suppliers may be induced by foreign affiliates to a technology upgrading by training, quality requirements, technical specifications, and so on.
- Other local enterprises may benefit from better and cheaper inputs supplied by local and/or foreign firms.
- In the case of final products, customers and distributors may benefit from marketing and other knowledge of the MNE. In the case of intermediate products, customers may benefit from better and cheaper intermediate products.

In sum, vertical spillovers depend on the number, kind and quality of the linkages with local firms, and this, in turn, depends on the local content, which is determined by the size of the local market, the intensity of intermediate consumption of the MNEs' production processes,[3] the technical capabilities of local enterprises, the MNEs' sourcing policies, the local content regulation, and the time horizon (OECD, 2002; UNCTAD, 2001).[4]

4. AN EMPIRICAL SURVEY OF THE SPILLOVERS LITERATURE

Since the mid-1990s, many studies about productivity spillovers have appeared. The most frequent way to check for these spillovers has been to compare whether the productivity of local firms is higher with the greater presence of foreign firms, controlling for all other possible factors affecting productivity. The best econometric way to do this is by using panel datasets and not cross-section data, which are likely to give biased results. The surveys on this matter conclude that the negative results from panel data studies are more reliable than those for cross-sections, and, therefore, that there is little evidence of positive spillovers from FDI. Most of the studies with panel data for developed and developing countries have a negative sign for spillovers, or are statistically insignificant in the extensive list they present (Görg and Greenaway, 2001; Görg and Strobl, 2001; Navaretti and Venables, 2004).

However, a number of studies of productivity spillovers based on panel data have appeared in recent years. These studies find some more evidence for positive spillovers than earlier ones, but there are still some mixed results. Haskel *et al.* (2002), using a panel of UK manufacturing plants between 1973–1992, find a positive and robust spillover effect of inward FDI on productivity in local plants. In the same vein, Keller and Yeaple (2003) find positive spillovers in the USA between 1987 and 1996, and Griffith *et al.* (2003) also find positive results for the UK with panel data for the period 1980–1992. On the contrary, Harris and Robinson (2003), also for UK manufacturing firms, could not find positive spillovers with panel data for 1974–1995, and, actually, they get statistically insignificant results. This outcome is shared by the study from Girma and Wakelin (2001), also for the UK with panel data at firm level for the period 1980–1992. In sum, studies for different countries, periods, econometric techniques and variables specifications reveal a sceptical picture about such spillovers.

5. A NEW EMPIRICAL ANALYSIS

Our paper is not an analysis of the consequences that FDI generate on the productivity of local firms (that is, of the existence of productivity spillovers), but, more precisely, of the consequences of the presence of FDI or multinational enterprises on the productivity gap between local and foreign firms (foreign affiliates and subsidiaries). In this sense, we focus our paper on the size of the productivity spillovers. The bigger the size of these spillovers (assuming that they exist and that they are positive), the higher the growth of productivity of local firms, and, therefore, the smaller the productivity gap. In sum, the spillovers generated by multinational enterprises would lead to a process of convergence of productivity between foreign and local firms.

To test this hypothesis of productivity convergence we use the following model proposed by Blomstrom and Wolf (1994):

$$PC_i = \beta_1 + \beta_2 GAP_i + \beta_3 FS_i + \beta_4 FSGr_i$$

PC_i shows the percentage change of the productivity ratio between foreign and local firms in the unit *i* analysed (industry or economy). *GAP* measures the ratio of productivity between foreign and local firms at the beginning of the period analysed. *FS* is the share of foreign firms in the value added. This share is measured as the average share for the period analysed. Finally, *FSGr* represents the growth during the period of the share of foreign firms in value added.

If we take for granted that foreign firms have higher productivity than local firms, the productivity ratio between foreign and local firms is higher than 1. Therefore, a positive sign of PC_i means a divergence of productivity, that is, productivity in foreign firms grows faster than in local firms. On the contrary, a negative sign of PC_i means a convergence of productivity, that is, productivity in local firms grows faster than in foreign firms.

Productivity of foreign and local firms diverge if the sign of the coefficients β_2, β_3 and β_4 is positive. Conversely, productivity of foreign and local firms converge if the sign of the coefficients β_2, β_3 and β_4 is negative. The economic interpretation of the coefficients is the following one. If β_2 is positive, it means that the greater the productivity gap at the beginning of the period, the greater that gap will be in the future. The higher efficiency-productivity of foreign firms leads to a process of cumulative causation and to a reinforcement of their competitive advantages. That behaviour is associated with the difficulties of local firms in absorbing the technological and knowledge advantages enjoyed by multinational corporations and their foreign affiliates. A positive sign of β_2 would mean the incapacity of local firms to absorb the potential spillovers from foreign affiliates. On the contrary, a negative sign of β_2 means a convergence in productivity between foreign and local firms. The greater the productivity gap, the higher the incentives for local firms to fully absorb the spillovers generated by the presence and working of foreign affiliates, and the faster the pace of rising productivity of local firms.

In the cases of the variables *FS* and *FSGr*, if β_3 and β_4 are positive, it means that the greater the presence of multinational enterprises, the greater the productivity gap (divergence) will be between foreign and local firms.

This divergence is interpreted in the sense that foreign subsidiaries are crowding out local firms by making the latter less competitive, since their productivity will be lower. In other words, foreign subsidiaries generate negative spillovers.

We must keep in mind that we are not actually measuring spillovers, that is, whether FDI increases the productivity of local firms or not. When we talk of 'positive' ('negative'), we mean that productivity growth of local firms is higher (lower) than that of foreign affiliates. We identified productivity with competitiveness, and what we mean is that the presence of foreign firms can improve the competitiveness of local firms by making them more similar to the former (positive spillovers) or that, on the contrary, the presence of foreign firms can worsen the competitiveness of local firms by increasing the productivity gap with the former (negative spillovers).

In sum, the model proposes that the evolution of the productivity gap between foreign and local firms is related to the size of the productivity gap and to the size of foreign firms in the local economy. In this sense, the sign and size of spillover effects are determined by the above elements.

To test this hypothesis and model, we use the data from the OECD database 'Measuring Globalisation: The Role of Multinationals in OECD Economies'. With the data included in the database, we have calculated the share of foreign firms in total value added and the productivity (value added for employee) of local and foreign firms. These data have been calculated at a national level, and within a country at an industry level (classification ISIC Rev.3). Though the database includes data for 21 countries, our analysis is made on the basis of only 12 countries due to lack of data on the rest. Besides, we have focused our analysis on the manufacturing sector because of its higher industry disaggregation.

Table 10.1 shows the basic outcomes of the analysis of the productivity of local and foreign firms. The data corroborate one of the basic assumptions of the theory of FDI and international production: in all cases, with the sole exception of Finland, the productivity of foreign subsidiaries is higher than that of local firms and, therefore, it could be stated that foreign firms are more competitive than local firms. However, the size of the productivity gap varies among countries. Countries like Finland, France and Sweden have low productivity gaps, indicating a high level of productivity for local firms. On the contrary, countries like Hungary, Ireland, Portugal or Turkey have very high gaps: in these countries the productivity gap is well above 2. The evolution of the productivity gap is not homogenous: the productivity gap rises in eight countries but it falls in four economies (France, the Netherlands, Sweden and the United Kingdom). In all cases, productivity rises in both foreign and local firms, and, therefore, the

Table 10.1 *Productivity of foreign and local firms in the manufacturing sector (millions of local currency units) and participation of foreign firms in total value added (%)*

		Foreign firms (1)	Local firms (2)	Foreign Share (%)	Productivity Gap (1/2)
Czech Rep.					
	1997	0.467	0.2774	16.8	1.685
	1999	0.533	0.3008	25.5	1.770
Finland					
	1995	0.319	0.3193	9.7	1.000
	1999	0.371	0.3682	16.0	1.008
Netherlands					
	1995	0.167	0.1117	27.3	1.493
	1998	0.201	0.1397	28.7	1.436
Ireland					
	1991	0.076	0.0277	68.4	2.744
	1998	0.188	0.0376	81.9	5.011
Portugal					
	1996	0.032	0.0172	13.6	1.831
	1999	0.034	0.0154	31.5	2.195
Turkey					
	1992	584	251	10.7	2.327
	1998	21,555	8,471	12.9	2.545
France					
	1993	0.357	0.2844	28.7	1.254
	1998	0.395	0.3349	31.2	1.178
Hungary					
	1993	1.034	0.667	40.6	1.550
	1999	4.203	1.536	70.4	2.737
Japan					
	1994	9.473	7.5629	1.0	1.253
	1996	12.176	8.0848	1.2	1.506
Norway					
	1991	0.370	0.3305	8.1	1.120
	1998	0.667	0.4059	25.7	1.642
Sweden					
	1990	0.341	0.3087	14.8	1.105
	1998	0.571	0.5261	21.1	1.086
UK					
	1995	0.059	0.0323	25.7	1.811
	1998	0.048	0.0369	32.9	1.306

Source: Our calculations.

evolution of the productivity gap is explained by the different paces of growth of productivity in foreign and local firms. The exceptions to this rule are the UK for foreign firms, and Portugal for local firms. In the British case, the sharp fall in the productivity gap is mainly related to a fall in the productivity of foreign firms, but in the Portuguese case, the increase in the productivity gap is mainly explained by the fall of productivity of local firms.

In relation to the behaviour of the foreign share in total value added, this share has increased in all the countries. However, this common trend conceals a big disparity. In countries like Japan, Turkey and Finland, the share of foreign firms in the manufacturing value added is below 20 per cent. By contrast, countries like Hungary or Ireland have shares above 70 per cent.

As mentioned, our objective is to test the hypothesis that the evolution of the productivity gap between foreign and local firms is explained by the spillover effects from FDI, where the size and sign of these spillovers are related to the initial size of the productivity gap and the size and evolution of the presence in the local economy of foreign subsidiaries. With this aim, we have developed a cross-country regression at the whole manufacturing sector level, using an OLS method. With this analysis, therefore, we are testing the existence of interindustry spillovers.

Table 10.2 shows the outcome of this regression. In opposition to the 'optimistic' view, all the variables have the opposite sign than would be expected in the spillovers mainstream literature. *GAP* and *FS* seem to play no role in the evolution of the productivity gap. However, both variables have a positive sign. Therefore, the greater the productivity gap and the greater the presence of foreign firms at the beginning of the analysed

Table 10.2 Determinants of dependent variable convergence for the manufacturing sector

Constant	−0.285
	(−1.114)[a]
GAP	0.093
	(0.521)
FS	0.004
	(0.812)
FSGr	0.021*
	(2.040)
R^2 adjusted	0.476
F	4.324**
N	12

Notes: [a] t-values in parenthesis. * significant at 10%, ** significant at 5%.

period, the greater will be the productivity gap between foreign and local firms.

The only significant variable is the growth of the foreign share in total manufacturing value added (*FSGr*). The coefficient has a positive sign: the higher the growth of foreign firms' share, the greater the productivity divergence between both types of firms. As we saw in Table 10.1, foreign firms have increased their share in manufacturing value added in all the countries, the (relative) productivity-competitiveness of local firms has fallen in all the analysed economies. In sum, FDI, in general, increases the productivity gaps irrespective of the initial productivity gap. Therefore, FDI presence is not generating inter-industry spillovers big enough to reduce the productivity gaps.

However, this conclussion suffers from the high level of aggregation of our analysis, since we are focusing on the whole manufacturing sector. On the one hand, foreign firms may be investing in high productivity activities (industries or phases of the value added chain) whilst local firms would be investing in low productivity activities. On the other hand, since we focus on whole manufacturing sectors, the outcomes could be different with a higher sectorial disagreggation. To check this possibility, we have proceeded with the same model but with a higher disaggregation, analysing the evolution of the productivity gap for nine manufacturing industries. We have applied an OLS cross-country analysis for manufacturing industries, thus checking the existence of intraindustry spillovers. The results for the nine industries are shown in Table 10.3.

As Table 10.3 shows, the model is only relevant for four industries: food, beverages and tobacco; wood and paper; basic and fabricated metal products; and total machinery and equipment. Focusing on the determinant variables of the evolution of the productivity gap, the share of foreign firms (*FS*) in the industry value added is significant in only one industry (total machinery and equipment), and with a positive sign (the presence of foreign subsidiaries involves a higher divergence in productivity between foreign and local firms). In the case of the initial productivity gap (*GAP*), it is only significant in one industry (basic and fabricated metal products), and with a negative sign (the initial productivity gap is associated with higher productivity gaps at the end of the period). The only relevant variable would be the growth of the foreign share in value added. This variable is significant in four of the nine industries (food, beverages and tobacco, wood and paper, basic and fabricated metal products, and total machinery and equipment). In three out of these four industries, the sign of the coefficient is positive (food, beverages and tobacco, basic and fabricated metal products, and total machinery and equipment), and, therefore, the increase in the share of foreign firms in value added leads to a higher divergence of productivity between foreign and local firms. This outcome

Table 10.3 Determinants of dependent variable convergence for manufacturing industries

INDUSTRY	Constant	*GAP*	*FS*	*FSGr*	R^2 adjusted	*F*	*N*
Food, beverages and tobacco	−0.015	0.016	0.003	0.161*	0.765	8.597**	8
Textiles	0.258	−0.125	0.001	−0.003	0.014	0.023	9
Wood and paper	−0.081	0.0156	−0.006	−0.127***	0.478	3.444***	9
Chemical products	−205.047	199.567	−2.666	95.155	0.265	1.722	7
Non-metalic mineral products	0.252***	−0.022	−0.004	−0.161	0.265	1.960	9
Basic & fabricated metal products	0.079	−0.167***	0.005	0.465***	0.668	6.364**	9
Total machinery & equipment	−0.241**	0.001	0.005***	0.384*	0.985	112.635*	6
Medical, precission & optical instruments	3.095	−4.715	0.073	2.666	−0.099	0.790	8
Transport equipment	0.812	−0.448	0.011	−0.118	−0.046	0.911	7

Notes: * significant at 1%, ** significant at 5%, *** significant at 10%.

is associated with the existence of negative spillovers. Only in the case of wood and paper is the sign of the coefficient negative, and, therefore, the productivity gap falls (positive spillovers).

From the industry perspective, in two industries (food, beverages and tobacco, and total machinery and equipment) FDI is associated with higher productivity gaps between foreign and local firms. In one case (wood and paper), by contrast, FDI is associated with lower productivity gaps between foreign and local firms. And in one case (basic and fabricated metal products) the result is indeterminate, since the initial productivity gap leads to lower future productivity gaps, but the growth of the foreign share in value added leads to a higher productivity gap. In the other five industries, the outcomes are not significant.

In sum, it seems clear that the analysis seems to conclude that there does not exist any generalized outcome about the existence and sign of the spillover effect. If anything, it could be concluded that the outcomes would be negative and that FDI reinforces the differences in productivity between foreign and local firms. However, there exist some caveats in this preliminary conclusion. The first one is the existence of a low number of observations due to poor data availability for both countries and industries. The second caveat is related to a specification problem of our model. The model does not focus on productivity levels, as proxied by labour productivity, but, actually, on productivity convergence, that is, on the relative growth of productivity levels for foreign and local firms.

6. CONCLUSIONS

The paper contributes to reinforce the sceptical view about the existence of productivity spillovers arising from the superior technologies of foreign subsidiaries. In any case, the outcomes show that there exists a greater likelihood of negative spillovers, leading to a wider productivity gap between foreign and local firms, although this outcome depends on the industry analysed.

In this sense, the contribution of inward FDI to economic growth for enhanced productivity of local firms seems to be, at best, small, and even nil if we take into account the possibility of the crowding out of local firms generated by the presence in the local economy of more competitive foreign subsidiaries. Furthermore, FDI can contribute to generate a dual economy with a group of MNEs operating in the local economy as an 'enclave', that is, a group of high productivity foreign firms with poor linkages with the lower productivty local firms. Therefore, the positive contribution of FDI to the economic growth of host economies would only come from the higher efficiency and productivity of foreign subsidiaries.

We must stress the fact that our analysis has focused on OECD economies, that is, mostly developed economies where the presence of positive spillovers should be higher and where the capacity of local firms to absorb these spillovers should also be higher. However, the outcomes pose serious doubts about the existence of these spillovers and about the absorption capacity of host economies. Therefore, the existence of productivity spillovers and the absorption capacity in developing economies might be even lower.

NOTES

[1] There are other types of spillovers related to export or wages, but we will centre the analysis around productivity, because this is the principal vector for growth.

[2] From a FDI policy, we must also consider that the existence of these spillovers may justify the existence of public subsidies to the location of foreign subsidiaries in the local economy.

[3] This is related to the phase of the value chain in which foreign subsidiaries operate.

[4] In the sense that local linkages increase with time, that is, with the life of the foreign subsidiary.

REFERENCES

Aitken, B., H. Hanson and A. Harrison (1997), 'Spillovers, foreign investment, and export behavior', *Journal of International Economics*, **43** (1–2), 103–32.

Blomström, M. and A. Kokko (1996), 'The impact of foreign investment on host countries: a review of the empirical evidence', World Bank Policy Research Paper, **1745**.

Blomström, M. and E. Wolf (1994), 'Multinational corporations and productivity convergence in México', in W. Baumol, R.R. Nelson and E.N. Wolf (eds), *Convergence of Productivity: Cross-National Studies and Historical Evidence*, Oxford: Oxford University Press, pp. 263–84.

Carkovic, M. and R. Levine (2002), 'Does foreign direct investment accelerate economic growth?', mimeo, University of Minnesota, Minneapolis.

De Mello, L.R. (1997), 'Foreign direct investment in developing countries and growth: a selective survey', *Journal of Development Studies*, **34** (1), 1–34.

Fosfuri, A., M. Motta and T. Ronde (1998), 'FDI and spillovers through workers mobility', Universidad Pompeu Fabra, *Working Paper* **258**, Barcelona, Spain.

Girma, S. and K. Wakelin (2001), 'Regional underdevelopment: is FDI the solution? A semiparametric analysis', *GEP Research Paper*, **11**.

Glass, A.J. and K. Saggi (1999), 'Multinational firms and technology transfer', *World Bank Policy Research Working Paper*, **2067**.

Globerman, S., J. Ries, and I. Vertinsky (1994), 'The economic performance of foreign affiliates in Canada', *Canadian Journal of Economics*, **27** (1), 143–56.

Görg, H. and D. Greenaway (2001), 'Foreign direct investment and intraindustry spillovers: a review of the literature', *GEP Research Paper* **37**, Nottingham, Leverhulme Center for Research on Globalization and Economic Policy.

Görg, H. and E. Strobl (2001), 'Multinational companies and productivity spillovers: a meta-analysis', *Economic Journal*, **111** (475), 723–39.

Graham, E. (2000), *Fighting the Wrong Enemy: Antiglobal Activists and the Multinational Enterprise*, Washington, DC: Institute for International Economics.

Griffith, R., S.J. Redding and H. Simpson (2003), 'Productivity convergence and foreign ownership at the establishment level', *CEPR Discussion Paper*, **3765**.

Harris, R. and C. Robinson (2003), 'Foreign ownership and productivity in the United Kingdom. Estimates of UK manufacturing using the ARD', *Review of Industrial Organization*, **22** (3), 207–23.

Haskel, J., S. Pereira and M. Slaughter (2002), 'Does inward FDI boost the productivity of domestic firms?', *NBER Working Paper*, **8724**.

Ietto-Gilles, G. (2005), *Transnational Corporation and International Production: Concepts, Theories and Effects*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Keller, W. and S.R. Yeaple (2003), 'Multinational enterprises, international trade and productivity growth: firm level evidence from the United States', *NBER Working Paper*, **9504**.

Navaretti, G.B. and A.J. Venables (2004), *Multinational Firms in the World Economy*, Princeton, NJ: Princeton University Press.

Nunnenkamp, P. and J. Spatz (2003), 'Foreign direct investment and economic growth in developing countries: how relevant are host-country and industry characteristics?', *Kiel Working Paper*, **1176**.

OECD (2002), *FDI for Development: Maximizing Benefits, Minimizing Costs*, Paris: OECD.

Rodríguez, A. (1996), 'Multinationals, linkages and economic development', *American Economic Review*, **86** (4), 852–73.

Saggi, K. (2000), 'Trade foreign direct investment and international technology transfer: a survey', *World Bank Working Papers Series in International Economics*, **2349**.

Smarzynska, B. and M. Spatareanu (2005), 'Disentangling FDI spillovers effects: what do firm perceptions tell us?', in T.H. Moran, E.M. Graham and M. Blomström (eds), *Does FDI Promote Development*, Washington, DC: Institute of International Economics, pp. 45–71.

UNCTAD (1992), *World Investment Report: TNC as Engines of Growth*, New York and Geneva: United Nations.

UNCTAD (2001), *World Investment Report: Promoting Linkages*, New York: United Nations.

Wang, J.Y. and M. Blomström (1992), 'Foreign investment and technology transfer: a simple model', *European Economic Review*, **36** (1), 137–55.

11. Increasing returns and the distribution of manufacturing productivity in the EU regions

Bernard Fingleton and Enrique López-Bazo

1 INTRODUCTION

Despite the diffusion of new technologies to many rural or low productivity regions, differences in manufacturing productivity levels and growth rates persist across the EU. In this paper we set up a model to account for these growth rate variations, and use the model to estimate the impact on long-run equilibrium productivity levels of alternative assumptions about returns to scale. Our model is influenced by the new wave of theory in urban and geographical economics, which for the first time allows a general equilibrium solution within the context of increasing returns to scale. Our analysis shows that the increasing returns hypothesis stands up to empirical scrutiny, and therefore adds to the growing body of evidence (see for example Hanson, 1997; Brülhart, 1998; Ottaviano and Puga, 1998; Fingleton, 2001a,b, 2003; Henderson and Thisse, 2003; Brakman *et al.*, 2004) supporting the new theory, although, almost irrespective of their theoretical provenance, many regional economists and economic geographers would agree that increasing returns are a fundamental prerequisite for a proper understanding of regional disparities.

However, if we restricted ourselves entirely to our chosen theoretical context, we would be in deep trouble analytically. Geographical economics theory, at least in its most elemental form, is difficult to turn into empirical models without sacrificing formal elegance, because it is apparent that pecuniary externalities are by themselves insufficient for an unbiased econometric specification. To help us go down this road in this paper, while drawing on the coherent framework of the new theory, our starting point is the interrelated urban economics model rather than geographical economics per se, since this provides a clearer route to econometric analysis.

2 A PRODUCTIVITY GROWTH MODEL – THEORY

The new urban and geographical economics theory allows increasing returns to scale, while at the same time the decision problem for each actor is explicitly stated as one of profit or utility maximization, with market structure assumptions based on the Dixit–Stiglitz theory of monopolistic competition. In the case of geographical economics (Fujita *et al.*, 1999), monopolistic competition normally applies to 'industry', while 'agriculture' is competitive. In urban economics (Rivera-Batiz, 1988; Abdel-Rahman and Fujita 1990; Quigley, 1998) it is 'industry' that is competitive, while non-traded producer 'services' are under monopolistic competition, and it is these latter assumptions which underlie our empirical model[1] linking manufacturing productivity growth to the growth of manufacturing output.

We derive the reduced form at the core of the empirical model by substituting the level of composite services I into the Cobb–Douglas production function in equation (1):

$$Q = M^{\beta} I^{1-\beta} \tag{1}$$

The level of composite services is determined by the CES production function, given by equation (2):

$$I = [\int_{d=1}^{D} i(d)^{1/\mu} \partial d]^{\mu} \tag{2}$$

In equation (2), the exogenous parameter[2] μ reflects the level of monopoly power of producer service firms and *i(d)* is the 'typical' output of a service variety, and there are D varieties. Since we assume a very large number of varieties we approximate the continuous integral by the discrete summation. At equilibrium *i(d)* is a constant across all varieties (see Appendix A), and therefore we can reduce the summation to a product as in equation (3):

$$I = [\sum_{d=1}^{D} i(d)^{1/\mu}]^{\mu} = [Di(d)^{1/\mu}]^{\mu} = D^{\mu} i(d) \tag{3}$$

Given this simplified form, we substitute for I in equation (1) and use the equilibrium values for the number of varieties D and M[3] to obtain the relationship between Q and N, hence:

$$Q = M^{\beta} I^{1-\beta}$$

$$I = D^{\mu} i(d)$$

$$Q = M^{\beta}(D^{\mu} i(d))^{1-\beta}$$

$$Q = M^{\beta} D^{\mu-\mu\beta} i(d)^{1-\beta}$$

$$D = \frac{(1-\beta)N}{ai(d)+s}$$

$$M = \beta N$$

$$Q = N^{\beta+\mu-\mu\beta} \beta^{\beta} (ai(d)+s)^{\mu(\beta-1)} i(d)^{1-\beta} (1-\beta)^{-\mu(\beta-1)}$$

$$Q = N^{\beta+\mu-\mu\beta} \phi'$$

$$Q = \phi' N^{1+(1-\beta)(\mu-1)} \tag{4}$$

In equations (4), s is the fixed labour requirement and a is the marginal labour requirement of producer services.

The theory outlined up till now therefore reduces to equation (5):

$$Q = \phi' N^{\gamma'} \tag{5}$$

with returns to scale equal to γ' and

$$\gamma' = [1 + (1-\beta)(\mu-1)] \tag{6}$$

The model described up to this point excludes technological externalities,[4] but as argued earlier these are necessary for unbiased estimation. The 'two' externalities we are concerned with are the effects of congestion and the impact of knowledge spillovers on productivity growth.

With regard to congestion, there are of course many sources, but in general congestion arises when firms use common, but unpriced, inputs[5] in short supply, such as road space or other communications networks or unpolluted air or water. It occurs when firms 'get in each others' way' or 'step on each other's toes' (Cameron, 1996). Congestion therefore comes from various sources that make production more difficult in a restricted space. As Gordon and McCann (2000) argue, we can really only observe the net realized effects of diverse simultaneous externalities, rather than individual sources. We therefore simply represent 'congestion' as a single parameter, following Ciccone and Hall (1996). To show this, let us introduce land (L) explicitly, hence:

$$Q = [M^{\beta} I^{1-\beta}]^{\alpha} L^{1-\alpha} \tag{7}$$

Up to now we have implicitly assumed that $\alpha = 1$, hence the amount of land makes no difference to production. By now introducing $\alpha < 1$, land becomes a factor, and by assuming $L = 1$, Q per unit of land is affected by congestion on that unit. Since $1^{1-\alpha} = 1$, we have

$$Q = (M^{\beta} I^{1-\beta})^{\alpha} = (\phi' N^{\gamma'})^{\alpha} = \phi N^{\gamma}$$
$$\gamma = \alpha[1 + (1-\beta)(\mu - 1)] \tag{8}$$

Our empirical analysis below tests whether we have evidence of constant ($\gamma = 1$), increasing ($\gamma > 1$) or diminishing ($\gamma < 1$) returns to scale, linearizing and rearranging to give the level of manufacturing output per labour unit as a function of the level of manufacturing output, thus:

$$\ln(Q/N) = \frac{\ln(\phi)}{\gamma} + \left[\frac{\gamma - 1}{\gamma}\right]\ln(Q) \tag{9}$$

And, assuming $M/N = \beta$, this becomes

$$\ln(Q/M) = \frac{\ln(\phi)}{\gamma} + \left[\frac{\gamma - 1}{\gamma}\right]\ln(Q) - \ln(\beta) \tag{10}$$

Equation (10) has the advantage of not requiring knowledge of the service sector per se, whereas information on manufacturing production is more accessible and reliable.

Knowledge spillovers within and between regions and cities are increasingly recognized (Fujita and Thisse 1996; Quigley 1998) as an important concentrator of economic activity, reflecting the earlier work of Jacobs (1969) among others.[6] In what follows, we focus on across-region variations in the rate of technical progress as a function of regionally differentiated knowledge spillovers.

To lead into our technical progress rate submodel, consider labour efficiency units:

$$M_t = E_t A_t = E_t A_0 e^{\lambda t} \tag{11}$$

In equation (11), E_t is the level of manufacturing employment and A_t is the efficiency level at time t based on initial level A_0 and technical progress rate λ. Re-expressing equation (10) gives our empirical growth model:

$$\ln(Q/E)_t = \frac{\ln(\phi)}{\gamma} + \left[\frac{\gamma - 1}{\gamma}\right]\ln(Q)_t - \ln(\beta) + \ln(A_0) + \lambda t \tag{12}$$

We assume that the technical progress rate λ in equation (12) depends on the level of human capital (*H*), the initial level of technology gap (*G*) and the trans-regional spillover of knowledge (*S*), plus an autonomous rate (ε)reflecting 'learning by doing' irrespective of the other factors. Hence

$$\begin{aligned} \lambda &= \nu H + \pi G + \rho S + \varepsilon \\ G &= 1 - \frac{P}{P*} = 1 - R \\ S &= Wp \\ W_{ij}^{*} &= Q_j^{\eta} d_{ij}^{-\delta} \\ W_{ij} &= \frac{W_{ij}^{*}}{\sum_j W_{ij}^{*}} \end{aligned} \qquad (13)$$

Hence what matters for growth is the level of human capital, as in Nelson and Phelps (1966),[7] who assume that it determines technology diffusion rates. In our model, this is one reason why low productivity regions with high human capital grow faster and catch up. We also assume that the human capital stock affects innovation rates, so that low productivity regions move up the productivity ladder. We hypothesize that convergence is determined by the level of technology gap variable *G*. In contrast, Benhabib and Spiegel (1994) and others assume that the human capital level interacts with the level of technology gap.[8] More precisely, according to equation (13), regions with larger human capital stocks are expected to make faster technical progress ($\nu > 0$) due to higher domestic innovation rates based on an enhanced level of local or domestic research and development, and due to the larger spillover of knowledge both from local and remote sources.

One problem is the scarcity of accurate measures of human capital stocks. Our data come from a labour force survey giving the share of the population aged 25–59 with higher educational attainment levels by EU region, and we nullify possible errors in variables via the use of instrumental variables and two-stage least squares.

The catching up-convergence mechanism assumes that the technical progress rate is a function of the initial technology level. Regions with a low level of technology see faster technical progress because innovations diffusing from the technological leadership are more beneficial (see Veblen, 1915 and 'the advantage in backwardness'), meaning that once regions are no longer backward, their advantage disappears. This is catching up in the true sense, whereas catching up due to enhanced human capital alone evidently allows overtaking.

The standard method for testing the technological catch up hypothesis (see Fingleton and McCombie, 1998) is to use a proxy for initial technology level that is some function of the initial level of GDP per capita. The functions employed have included the level of GDP per capita or per employee, its reciprocal, and the ratio of the level of GDP per capita to the technology leader. Since here we are concerned with manufacturing productivity growth, it seems reasonable to use the initial manufacturing productivity level gap *G*. The hypothesized positive relationship ($\pi > 0$) between the technical progress rate and *G* is the fundamental determinant of convergence.

The diffusion process underpinning catching up implies no distance decay, with freely circulating knowledge available from sources such as journals, books and the worldwide communications media. On the other hand, with spatially impeded information flows, regions with fast technical progress occurring in 'neighbouring' regions see faster than otherwise technical progress, the corollary being that slow technical progress nearby restricts technical progress locally. 'Who your neighbours are' matters, because of the impedance to knowledge flow across space. Working with the NUTS 2 regional system of the EU means that we are likely to see significant flows across region boundaries because they are largely formal or administrative units with little functional coherence. In addition, national barriers have been progressively reduced within the EU, thus helping spillover between EU countries.

Spillover may come from sharing of the same labour pool in a common local labour market area straddling regional boundaries, the thesis being that the rate of productivity growth occurring in one region will be transmitted to other nearby regions as workers embodying knowledge switch jobs without changing residence. It might be argued that this would be entirely a pecuniary externality operating via the labour market. Firms may spend money on training labour but not appropriate all the benefits as nearby firms gain by not having to pay the full cost of that training. However, some knowledge transfer is not mediated by market transactions. For example some aspects of knowledge will be of a learning-by-doing variety accumulated on the job, and carried to other firms as an incidental attribute (quality) of labour, without price or market, but which nevertheless contributes to technical progress. For example a programmer may incidentally learn from others how to install and service hardware, and this accumulated knowledge may in turn be transferred to colleagues, as well as to family and friends in a process of collective learning.

The second mechanism entails demonstration[9] of the efficacy of knowledge for productivity. This involves inter-firm interaction across region boundaries. Because of proximity, firms locally and in nearby regions may be competitors for the same local markets (a 'creative destruction effect'), or collaborate as part of a localized production chain. In either case, fast technical progress in neighbouring regions will tend to induce technical progress and thus fast productivity growth locally.

These considerations entail the presence of Wp in equation (13), with cell i of vector Wp equal to the weighted average of labour productivity in regions 'surrounding' region i. Size, to a degree, offsets remoteness, because of the extensive trade and labour market that a large diverse local economy naturally generates, and this is apparent in the definition in equation (13) of the absolute (conditional) level of interaction between regions i and j, namely W^*, in which Q_j is the economy size proxy, the 1975 level of output in region j. Given the size of region i, the interaction with region j is likely to be stronger if region j possesses a larger economy. Proximity is represented by the great circle distance (d_{ij}) between the centres of regions i and j. Given a negative parameter (δ), increasing distance reduces the absolute conditional interaction between i and j. It is assumed that $\delta = 2$ and $\eta = 1$ as a result of trials of different values reported in Fingleton (2001a).

We summarize the determinants of the technical progress rate by equation (14):

$$\lambda = \pi G + \upsilon H + \rho Wp + \varepsilon \tag{14}$$

Finally we arrive at our growth model, which is obtained by inserting (19) into (17), differentiating with respect to time, and adding a well-behaved perturbance, which gives:

$$p = \varepsilon + \frac{\gamma - 1}{\gamma} q + \pi G + \upsilon H + \rho Wp + \xi$$
$$\xi \sim N(0, \sigma^2 I) \tag{15}$$

where p is the logarithmic growth of manufacturing labour productivity (Q/E), and q is the logarithmic rate of growth of manufacturing output (Q).

3 ESTIMATION

In this section we examine issues relating to the estimation of equation (15), commencing with the question of endogeneity, which we expect to be a significant property of our specification that needs to be considered in our endeavour to produce consistent estimates. However, we find only marginal empirical evidence of endogeneity. Using the Hausman test, testing q, H and Wp jointly gives a p-value equal to 0.146 in the $F_{3,170}$ distribution, which is sufficiently large to suggest that they are exogenous. The Hausman test of the exogeneity of each single variable in the presence of other endogenous variables[10] gives test statistics for Wp, q and H equal to 2.128, 3.959 and 0.437 respectively, with p-values equal to 0.1446, 0.0466 and 0.5085 in the appropriate χ^2 distribution with one degree of freedom. Hence there is some very marginal evidence to suggest that q is endogenous, but the other two variables appear to be exogenous. With regard to Wp, which is by definition endogenous, it appears that the tests are insufficiently sensitive. However, in order to guard against any potential inconsistency not detected by these tests, it is safer to use instrumental variables.

Table 11.1 summarizes both (inconsistent) OLS estimates and 2sls estimates based on the exogenous technology gap variable G and the spatial lag WG. We also use two other instruments, q_1 and its lag Wq_1, in the first stage.

Table 11.1 OLS and two-stage least squares groupwise heteroscedasticity estimates

Variable	2sls			ols		
	Estimate	Standard error	t-ratio	Estimate	Standard error	t-ratio
Constant	−0.0240	0.0084	−2.84	−0.0218	0.0058	−3.74
q	0.3265	0.0820	3.98	0.4900	0.0639	7.67
G	0.0476	0.0071	6.66	0.0379	0.0072	5.30
H	0.0716	0.0274	2.61	0.0541	0.0145	3.74
Wp	0.4327	0.1640	2.64	0.5120	0.1340	3.82
σ^2				0.0001		
σ^2 (peri.)	0.00024					
σ^2 (core)	0.00011					
R^2				0.38 (0.49)		
Sq. corrn	0.4635			0.3793 (0.4915)		
Residual[a] *autocorrelation*				z = 1.545		
Exog. variables	q_1, Wq_1, G, WG			q, G, H, Wp		

Note: [a] z is the standardized value of Moran's I statistic for regression residuals, using randomization moments. Since $p = q - e$, where e is the growth of employment, some spurious correlation induced by the presence of q on both sides of the regression. In order to obtain a corrected R^2, as is given in Table 11.1, we fit the OLS model with e as the dependent variable (mathematically this is an identical model, with estimated signs exactly equal to the negative of those in Table 11.1, apart from the coefficient on q which is equal to one minus the value given in the table). The values in brackets are the result of fitting the model with p as the dependent variable, and allow us to gauge the effect of spurious correlation induced by having left hand side p and right hand side q. The square correlation is the square of the correlation between observed and fitted values.

The instrument q_1 is obtained using 3 groups (see Leser, 1966; Koutsoyiannis, 1977; Kennedy, 1992; Johnston 1984) with values 1, 0 or –1 according to whether q is in the top, middle or bottom third of its ranking, which ranged from 1 to 178. The assumption is that this mapping eliminates any correlation between q_1 and ξ induced by simultaneity. With four instruments, and given three presumed right hand side 'endogenous' variables, the necessary order condition for identification is satisfied, hence the equation is identified exactly.

There are additional or alternative additional instruments that could be introduced leading to over-identification, and some of the results of using different instruments or estimation techniques are given in Fingleton (2000, 2001a, b). As an illustration, and to provide supporting evidence of increasing returns and the significance of our variables independently of our instrument q_1, we give in Table 11.2 some new results based on a different set of exogenous variables, together with maximum likelihood

Table 11.2 ML and two-stage least squares estimates

Variable	ML			2sls		
	estimate	Standard error	t-ratio	estimate	Standard error	t-ratio
Constant	−0.0204	0.0052	−3.94	−0.0282	0.0069	−4.08
q	0.4957	0.0622	7.97	0.8645	0.1545	5.60
G	0.0386	0.0070	5.54	0.0301	0.0083	3.61
H	0.0543	0.0143	3.81	0.0781	0.0191	4.09
Wp	0.4520	0.1103	4.10	0.4626	0.1627	2.84
σ^2	0.00018			0.00023		
Sq. corrn	0.4914			0.4604		
Log likelihood	512.9					
Residual autocorrelation	3.081 (LM test)			z = 0.9623		
Exog. variables	*Q, G, H*			*G, luxdij, u, pop, pd*; first lags; country dummies		

Note: The exogenous variables are defined as follows: *luxdij* is the great-circle distance from Luxembourg, *u* is a dummy indicating whether the region is urban or rural (population density < 500 per sq km), *pop* is the population level in 1975, *pd* is the 1975 population density. The first lags are *WG*, *Wluxdij* etc. There are 11 country dummies (ie excluding Ireland and Luxembourg). Here z is the standardized value of the test statistic suggested by Anselin and Kelejian (1997) for residuals from IV estimation. The LM test statistic is distributed as χ^2_1under the null hypothesis of no error dependence (Anselin, 1988).

estimates based on the assumption that the only endogenous variable is the spatial lag. However our preferred set of parameter estimates remain the groupwise heteroscedasticity estimates, which control for residual heterogeneity,[11] and which provide the most conservative estimate of increasing returns to scale, which nonetheless remains highly significant.

It is evident from Table 11.1 that all the variables are significant, and controlling for the spillover of knowledge across region boundaries, the effect of human capital and catching up, we see that there is a significant link between q and p pointing to the presence of increasing returns. From our preferred estimates, $\gamma = \alpha[1 + (1 - \beta)(\mu - 1)] \approx 1.5$, so that congestion ($\alpha$) has not eliminated increasing returns to scale. Nonetheless, as we show below, the positive technology gap coefficient leads to regional convergence, not divergence.

4 SIMULATING THE LONG-RUN DISTRIBUTION OF MANUFACTURING PRODUCTIVITY IN THE EU REGIONS

A pivotal feature of the dynamics is the coefficient on G, which determines whether the regions converge to a stable equilibrium or not. Define equilibrium as the state in which the expected rate of productivity growth p is exactly the same across regions. Critical to our analysis is the assumption that

$$P_{t+1} = P_t \exp(E(p_t)) \tag{16}$$

so that at equilibrium the vector of values giving the level of technology gap[12] $G = 1 - P/P^* = 1 - R$ is constant.

In order to illustrate this, we use a simple iterative technique (see Fingleton, 2000, for an analytical solution). Define X as an n by k matrix containing as columns the unit vector and the k–1 variables q, G and H and b denotes the corresponding preferred parameter estimates from Table 11.1. Equations (17) give one round of iteration; to obtain the dynamics, this sequence of calculations is repeated for $t = 1 \ldots T$ and we observe how R evolves as t increases. In the dynamics, the vector of productivity growth rates at time t, $\mathrm{E}(p_t)$, depends on the matrix X as at time t, but X evolves because one column of the matrix, denoted by $X_{t+1,G}$ contains variable G_{t+1} which depends on the level of productivity at time t, P_{t-1}, and the latter is determined by the rate of growth of productivity $\mathrm{E}(p_t)$:

$$\begin{aligned} E(p_t) &= (I - \rho W)^{-1} X_t b \\ P_{t+1} &= P_t \exp(E(p_t)) \\ P^*_{t+1} &= P^*_t \exp(E(p^*_t)) \\ G_{t+1} &= 1 - \frac{P_{t+1}}{P^*_{t+1}} \\ X_{t+1,G} &= G_{t+1} \end{aligned} \tag{17}$$

The results of this section provide the input into our conditioning analysis. That is, using equations (17) we obtain the long-run distribution from the model estimates, and obtain virtual distributions as a result of changing the model coefficient values so as to accommodate alternative assumptions about returns to scale.

5 DISTRIBUTION DYNAMICS 1975–95

In order to visualize the evolution of productivity levels, we employ the density function and the stochastic kernel (see Appendix B), first describing how manufacturing Gross Value Added per worker for 178 EU regions evolved over the period 1975–95. Throughout we use the ratio $R_{it} = P_{it}/P_t^*$ in which P_{it} is the level of labour productivity in region i ($i = 1, \ldots, 178$) at time t ($t = 1975, 1980, 1985, 1990, 1995$), and P_t^* is the productivity level of the leading region at time t. Since R's upper bound is 1 (the leading region) and its lower bound approaches 0, the densities were estimated using a truncated Gaussian kernel (see Silverman, 1986).

Figure 11.1 shows that the distribution in 1975 is centred on a ratio of approximately 0.5, with the two tails containing an important share of the overall distribution and a cluster of regions characterised by very low productivity (low R). Figure 11.1 shows the development of a persistent 'hump' at a ratio of about 0.2, and over time the distribution shows signs of polarisation.

The stochastic kernel (Figure 11.2) reveals fluid internal dynamics at the high and moderate values of R (high and medium productivity levels) but relative homogeneity in the low productivity group, as indicated by the sharp peak on the vertical axis for low R. There is minimal 'churning' among the lowest regions, which remain persistently nearer the bottom of the productivity ladder, and the very low productivity regions have become more homogenous, as illustrated by the clockwise turn in the kernel. Regions in the range of roughly 0 to 0.4 become more concentrated in the range 0.1 to 0.3.

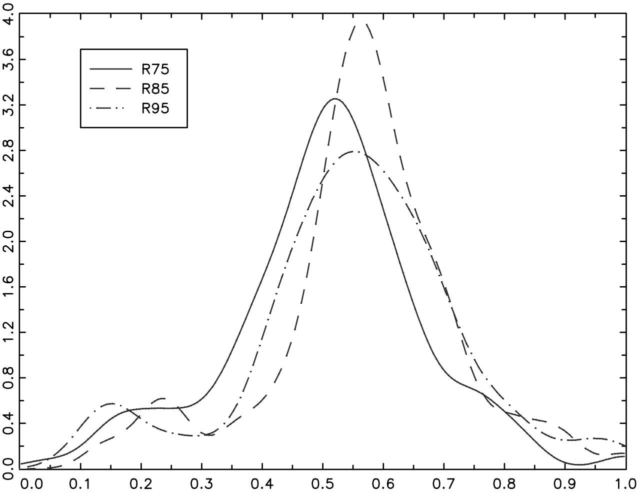


Figure 11.1 Estimated density function for the productivity level ratios (R)

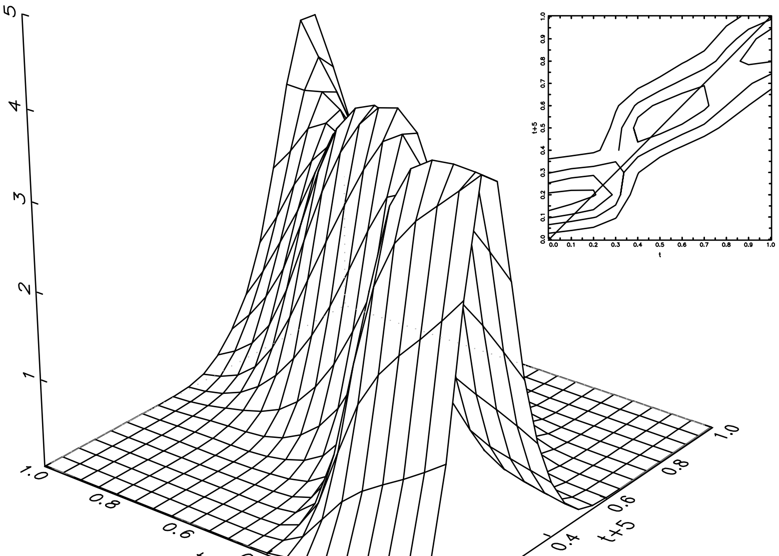


Figure 11.2 Stochastic kernel for 5-yearly transitions for the productivity level ratios

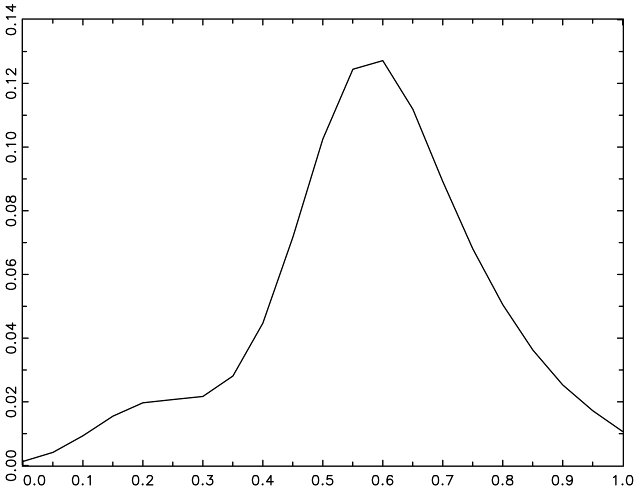


Figure 11.3 Ergodic distribution from the stochastic kernel

The ergodic distribution (Figure 11.3) is the long-run equilibrium derived from the stochastic kernel based on the 5-yearly transitions described above. Aspects of the historical data (Figure 11.1) are maintained in the long-run equilibrium, with the mode at roughly the same position, round about 0.4–0.5 and the cluster of low productivities (*R* around 0.2) persists as a long-run phenomenon.

6 EFFECTS OF INCREASING RETURNS ON THE LONG-RUN DISTRIBUTION OF MANUFACTURING PRODUCTIVITY

In this section, we compare the observed 1995 productivity distribution with a long-run steady-state R^e distribution assuming that the effects of factors influencing the equilibrium (returns to scale, human capital, rate of catch-up) are maintained in the long run. First, we assess how this differs from the observed distribution, and then compare it with an alternative R^e distribution based on an assumption that increasing returns are eliminated.

6.1 Equilibrium Distribution: The Long-run Equilibrium Implied by Our Preferred Estimates

Figure 11.4 compares the distribution of *R* in 1995 to that of R^e. The stochastic kernel twists clockwise through the whole range of *R*, though the

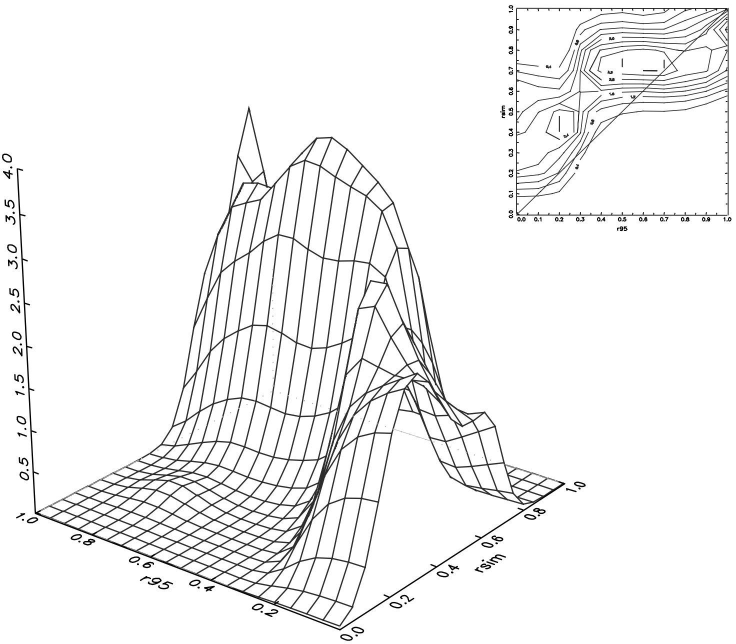


Figure 11.4 Stochastic kernel for the distribution in 1995 and the steady-state distribution from model estimates

twist is more intense for the medium and high values. Below 0.2, it is evident that there is a high probability that there will be an increased *R*. Nevertheless, the equilibrium distribution is characterized by considerable dispersion (regions with a productivity level ratio in 1995 of around 0.2 can end up with similar values in the equilibrium distribution or with values around 0.3–0.5). For regions with medium and large *R*, convergence at equilibrium is to a level that is only slightly higher on average than for low-*R* regions, but the dispersion is much greater and there is a much higher probability of attaining a value close to the leading region.

6.2 Equilibrium Under Different Assumptions About Returns to Scale

Increasing returns to scale make the spatial concentration of manufacturing activities more attractive, which translates into increasing inequality in manufacturing productivity across regions. Therefore, increasing returns should be responsible for some of the dispersion in the equilibrium distribution and are likely to affect (to some extent) the fortunes of the less productive regions. In this spirit, we simulate the equilibrium

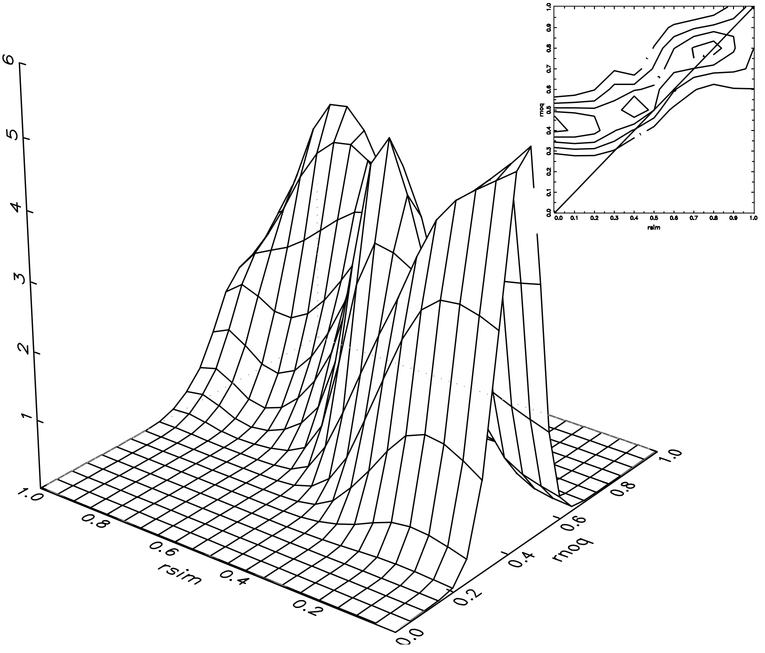


Figure 11.5 Stochastic kernel: IRTS conditioning

R distribution based on the assumption that γ (the coefficient that captures the extent of returns to scale in our model) is equal to one, which in turn means that the coefficient for *q* in equation (15) equals zero. This amounts to an assumption that manufacturing output growth rates are equal across regions, and therefore do not play a part in determining the rate of growth of manufacturing productivity. The determinants of manufacturing productivity growth under this scenario are therefore human capital, cross-region spillovers and catch-up, using the observed values and coefficient estimates from Table 11.1. Figure 11.5 summarizes the effect of nullifying increasing returns. The stochastic kernel shifts above the diagonal, indicating that the conditional distribution shifts to the right of the unconditional one. That is, R^e for most of the EU regions would be higher with constant rather than increasing returns to scale. Increasing returns contribute to the relatively poor situation of the less advanced regions. The clockwise turn in the kernel for low R^e values indicates clearly that were it not for the existence of increasing returns, the low productivity regions would have a much higher R^e. In contrast, the lack of increasing returns has relatively little impact on high R^e regions' productivity relative to the leading region.

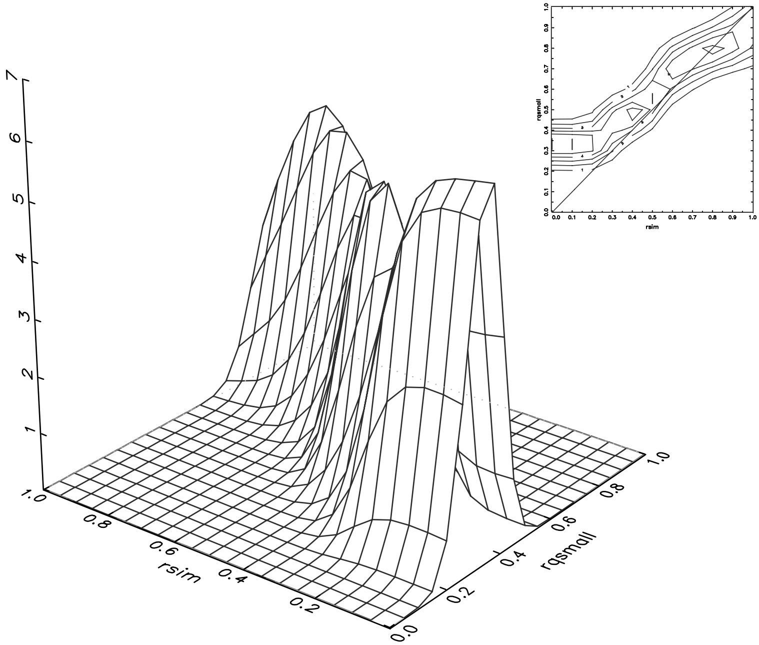


Figure 11.6 Stochastic kernel: small RTS conditioning

Next we adjust rather than eliminate the amount of increasing returns to scale. Figure 11.6 illustrates the outcome with the coefficient on *q* set to equal 75 per cent of the estimated value. Figure 11.7 shows the result of assuming a coefficient equal to 125 per cent of the estimated value. All other coefficients are set equal to their estimated values.

With large returns to scale, regions tend to a lower R^e as the distribution becomes stretched, and those with faster manufacturing output growth rates experience the greatest productivity boost and thus the greatest impact in terms of R^e. However, this impact is confined to the top of the R^e range; the majority of regions are relatively worse off vis-à-vis the technological leadership. In contrast, Figure 11.6 shows that lower than estimated returns to scale have the effect of pulling the lowest regions upward, with the upward shift the greatest for the very lowest regions. We might imagine that returns to scale would fall in the real world if congestion worsened significantly, as it might do in some regions (such as the South-East of England) unless there is a radical change of policy.

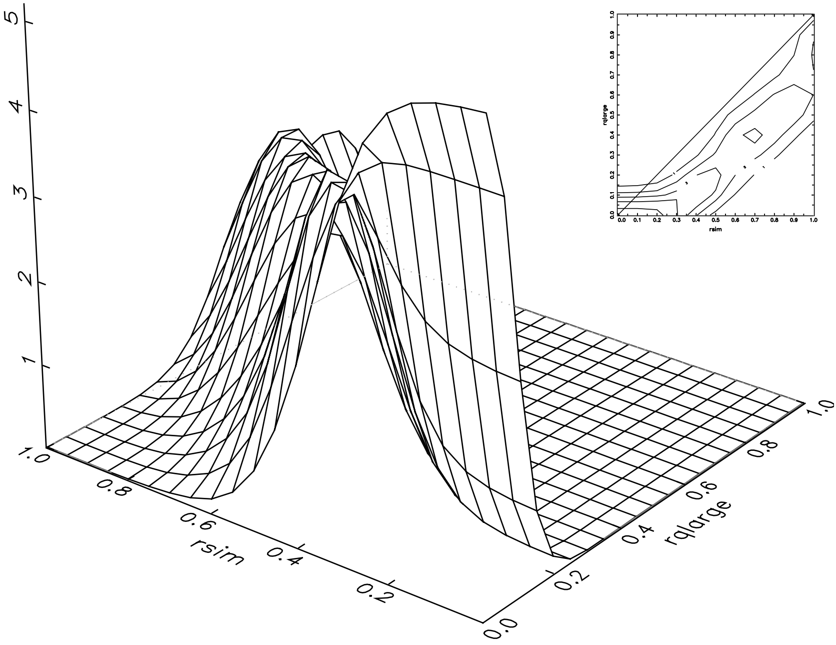


Figure 11.7 Stochastic kernel: large RTS conditioning

7 CONCLUSIONS

This paper combines recent developments in urban and geographical economics theory with an empirical spatial econometric model to provide support for one of the fundamental tenets of the new theory, that regional development occurs within the context of increasing returns to scale. However, we also show the influence of other factors, namely each region's stock of human capital and its initial technology gap. There are undoubtedly other variables as well that we have not considered, but our belief is that at our scale of resolution, these are averaged out of existence across regions, assuming the appearance of random shocks which are captured by the error term. The main variables act in different ways, for while we might expect regions' productivity growth rates to remain differentiated over time, and therefore for interregional manufacturing productivity levels to diverge as a result of persistent differences in human capital and output growth, we also envisage a catching up process counteracting the divergence mechanism and ultimately causing, in the long run, regions' growth rates to equalize as the advantages of backwardness are lost, with the effect that regions reach a steady state of differentiated productivity levels. This mechanism hinges on an assumption that differentiated manufacturing productivity growth feeds back to the technology gaps affecting the technical progress rate and thus the productivity growth rate. Therefore while increasing returns on their own would lead to ever-growing divergence of levels because it implies permanent differences in productivity growth rates, there are counteracting forces operating. This does not mean, however, that increasing returns to scale are irrelevant, for although our system converges to equilibrium, the amount of returns to scale affects the relative productivity levels to which regions converge.

To visualize outcomes associated with different returns to scale, we use stochastic kernels, which show that if conditions in the future remain as they are 'now', then the long-run manufacturing productivity level distribution is a more concentrated one than that which 'currently' exists (Figure 11.4), although it remains the case that in the long run equilibrium will be characterized by productivity differences. There is evidently no absolute converge on the distant horizon.

The effect of assuming constant rather than increasing returns to scale is to enhance the comparative status of the poorer regions, although the impact is a subtly differentiated one as is apparent from Figure 11.5. In Figures 11.6 and 11.7 we assume different amounts of increasing returns to scale. A reduction in returns to scale could occur as a result of increased congestion, as reflected by equation (8). An increase in returns to scale might ensue from structural changes in the economy, so that services under monopolistic competition become a relatively more important input to competitive manufacturing. Alternatively, it could be the result of a strengthening of monopoly power in the service sector, either due to legislative or technological reasons: for instance, if patenting law was strengthened to protect innovators, or technological prerequisites meant that entry costs became higher. Another reason could be transport infrastructure improvements, or changes in working practices such as a greater degree of working from home, which weaken the offsetting impact of congestion. Given the scope for changes in the fundamental determinants of returns to scale, it is important to examine the likely consequences of such changes, which is a contribution we make in this paper. We have not, of course, considered here the likely effects of changes in the regional distribution of human capital, due for instance to government-inspired educational policy. It appears however that this will not be constant. Therefore there is a great deal of uncertainty surrounding the likely direction of manufacturing productivity growth. Our simulations are merely suggestive rather than forecasts, and meant simply to highlight the possible consequence of changes in just one of a number of factors.

APPENDIX A: THE URBAN ECONOMICS MODEL

Endogenous Variables

1. Manufacturing labour (workers):

$$M = N\beta \tag{A1}$$

Manufacturing labour (workers) (M) equals total labour (N) times β, which is the equilibrium allocation of labour to manufacturing under competitive conditions.

2. Manufacturing output:

$$Q = M^{\beta} I^{1-\beta} \tag{A2}$$

This is a Cobb–Douglas production function. Output (Q) equals workers (M) raised to the power β, multiplied by the level of composite services (I) to the power ($1-\beta$).

3. Composite services:

$$I = [\int_{d=1}^{d=D} i(d)^{1/\mu} dt]^{\mu} \tag{A3}$$

$$I = [Di(d)^{1/\mu}]^{\mu} = D^{\mu} i(d) \tag{A4}$$

This is the CES (constant elasticity of substitution) (sub) production function for I, which is a function of the output of the typical services firm ($i(d)$), the number of services firms(D) and the elasticity of substitution, which diminishes with increasing μ. As μ approaches 1, then the services level approaches the number of firms times their output, as $\mu \gg 1$ it is more than this due to the effect of the number of varieties (D), so that increasing firms results in a proportionately larger I.

4. Equilibrium output level of typical service firm:

$$i(d) = \frac{s}{a(\mu - 1)} \tag{A5}$$

When firms are at equilibrium, so that (marginal) costs equal (marginal) revenues and profits are driven to zero, the output per firm can be shown to equal the fixed labour requirement (s) divided by the marginal labour requirement (a) times μ–1.

5. Cost:

$$c = w(ai(d) + s) \tag{A6}$$

Cost of production equals wage rate (w) times amount of labour ($ai(d) + s$)

6. Marginal cost equals wage rate (w) times marginal labour requirement (a):

$$mc = wa \tag{A7}$$

Revenue equals wage rate (w) times marginal labour requirement (a) times markup on costs (μ) [$wa\mu = p =$ price] times equilibrium output ($i(d)$):

$$r = wa\mu i(d) \tag{A8}$$

7. Marginal revenue equals price ($p = wa\mu$) times ($1–1/E$) where E is the constant (subjective) price elasticity of demand [which can be shown to equal $1/(1–1/\mu)$], thus $(1–1/E) = 1/\mu$:

$$mr = \frac{wa\mu}{\mu} = wa \tag{A9}$$

Hence $mr = p$ times $1/\mu = p/\mu$. Note, here we are talking about imperfect competition so that price is unequal to marginal revenue. In fact price (p) = wage rate (w) times marginal labour requirement (a) times markup (μ)

$$p = wa\mu \tag{A10}$$

If $\mu = 1$ we have perfect competition so then $mr = p$

The number of service firms (varieties):

$$D = \frac{(1 - \beta)N}{ai(d) + s} \tag{A11}$$

The number of firms (D) equals the total services labour force $(1 - \beta)N$ divided by the labour force per firm ($L = ai(d) + s$) at equilibrium.

8. Labour requirement:

$$L = s + ai(d) \tag{A12}$$

The labour requirement equal to fixed labour requirement (s) plus marginal labour requirement (a) times firm's output ($i(d)$)

Exogenous Variables

Marginal labour requirement (a): this is the exogenously determined increase in labour needed by the firm per unit increment of output (note that since output can be measured in any units, this can be left as 1).

Fixed labour requirement ($s>0$): this is the fixed cost in terms of service labour that must be incurred to produce any variety. It implies that increasing returns to scale exist in the service sector.

Monopoly power/elasticity of substitution (μ): as μ increases, the elasticity of substitution diminishes, as μ approaches 1, the services approach being perfect substitutes and variety diminishes in importance as a determinant of I.

Note that the elasticity of substitution is

$$\frac{\mu}{\mu-1} \tag{A13}$$

Total labour force (N): note how total manufacturing output (Q) is a nonlinear function of N, showing increasing returns with city size. However the latter is not modeled here and we treat N as exogenously determined.
The relative importance of workers versus services (β).

Equilibrium

Occurs when the level of output is such that marginal revenue (mr) equals marginal cost (mc), firms have entered shifting the demand curve to the left, driving down profits to zero, at which point entry stops. This is the equilibrium, when total revenue equals total costs and there are zero profits. This determines the equilibrium output level $i(d)$.

Hence, at equilibrium, profits are zero and costs equal revenues:

$$c = w(ai(d) + s) = r = wa\mu i(d) \tag{A14}$$

Hence

$$ai(d) + s = a\mu i(d)$$

$$i(d) = \frac{s}{a(\mu-1)} \tag{A15}$$

We can choose units of output to be anything we want, which means we can choose them so that the marginal labour requirement $a=1$. This gives the simplified version

$$i(d) = \frac{s}{\mu-1} \tag{A16}$$

APPENDIX B: DENSITY FUNCTIONS AND STOCHASTIC KERNELS

The stochastic kernel was initially introduced as a reaction to the rigid dynamics linked to neoclassical growth regressions (Barro, 1991; Mankiw *et al.*, 1992; Barro and Sala-i-Martin, 1992; Durlauf and Quah, 1999). Commencing with the Markov chain approach,[13] Quah (1993, 1996a, b) argued that this perspective enlightened our view of the dynamical properties of real economies. Using the stochastic kernel, attention is focused both on the external shape of the distribution, and on movements within the distribution, so as to reveal the complexity and detail of real dynamics.

In order to estimate the external shape of the interregional productivity level distribution, we carry out non-parametric estimation of the density function via the kernel method. It is useful to think of this kernel density estimator as a smooth version of a histogram, but with the 'bars' replaced by smooth 'bumps' (Silverman, 1986). Smoothing is accomplished by differentially weighting observations according to distance. More technically, the kernel density estimate of a series X at a point x is estimated by

$$f(x) = \frac{1}{Nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{B1}$$

where N is the number of observations, h is the bandwidth (or smoothing parameter) and $K(\ )$ is a kernel function that integrates to one. The kernel function is a weighting function that determines the shape of the bumps. We have used the Gaussian kernel in our estimates:

$$\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2) \tag{B2}$$

where u is the argument of the kernel function. The bandwidth h controls the smoothness of the density estimate. In this paper we have used the data-based automatic bandwidth suggested by Silverman (1986, equation 3.31):

$$h = 0.9N^{-\frac{1}{5}}\min\{s,\ Q^R/134\} \tag{B3}$$

where s is the standard deviation, and Q^R is the interquartile range of the series.

Our analysis of intra-distribution dynamics is via the estimation of a stochastic kernel (Stokey and Lucas, 1989) for the productivity level distribution over the period under analysis. Thus, the stochastic kernel provides the likelihood of transiting from one place in the range of values to the others.[14]

Following Johnson (2000), let R (the ratio of productivity in each economy to the leading economy) be the variable under analysis, and $f_t(R = x)$ and $f_{t+k}(R=x)$ the probability density of $R=x$ in period t and $t+k$ respectively. Assuming a first-order time-invariant process for the evolution of the distribution of R, and existence of marginal and conditional density functions for the R distribution, the relationship between both distributions can be summarized by:

$$f_{t+k}(R=y) = \int_0^t g_k(R=y|R=x)f_t(R=x) \tag{B4}$$

where $g_k(R=y|R=x)$ is the density of $R=y$ in period $t+k$ conditional on $R=x$, k periods before. Then, $g_k(R=y|R=x)$ summarizes information on movements within the distribution over time. It is computed by first estimating the joint density for the distributions at t and $t+k$ by the kernel method[15] and then dividing it by the marginal density of R at t, obtained by integrating the joint density over R at $t+k$.

Figure 11.2 is a typical three-dimensional graph representing a stochastic kernel, in this case the conditional density estimated from our data on R. It shows, conditional on the R density at t, the probability density 5 periods ahead. The vertical or z-axis of the three-dimensional plot measures the conditional density of each pair of points in the x-y space containing the R values at t and $t+5$. The lines that run parallel to the $t+5$ axis measure the probability of transiting from a specific R on the t-axis to any other R at 5 periods ahead. The two-dimensional graph in the top right corner contours the three-dimensional plot.

When the mass of probability concentrates on the positive diagonal, this indicates that there is strong persistence so that a region's R at t is very much the same as it was at t. On the other hand, the kernel can shift above or below the diagonal, indicating increases or decreases in values of R at time $t+5$. It can also twist clockwise or counterclockwise. A clockwise twist would indicate convergence, with the low productivity region having a high probability of moving to higher levels while a high productivity region would tend to move down. In the limit, when the mass of probability is parallel to the t-axis, all economies end up at similar values in $t+5$ regardless of their position at time t.

The ergodic density of R is the long-run shape of the distribution. Given the dynamics summarized by $g_k(R=y|R=x)$ it is the solution to:

$$f_\infty(R=y) = \int_0^1 g_k(R=y|R=x)f_\infty(R=x) \tag{B5}$$

Finally, note that the stochastic kernel can also be used to describe movements between two distributions, rather than being restricted to the analysis of the same variable at different points in time (Quah, 1996a).

ACKNOWLEDGEMENT

The authors wish to thank participants at the CEPR workshop[16] for their most helpful comments.

NOTES

1. The reduced form deriving from the model is identical to Verdoorn's law, which is commonly used to represent increasing returns in manufacturing in the post-Keynesian literature (Fingleton, 2001b).
2. This is the substitution parameter of the CES production function, which determines the elasticity of substitution, the price elasticity of demand and the internal returns to scale given by the average cost to marginal cost ratio for producer services in equilibrium.
3. This assumes that the economy is in a competitive equilibrium and workers are paid the value of their marginal products.
4. Krugman (1991) argues that while technological externalities may be relevant, they leave no 'paper trail' and prohibit formal analysis. Pecuniary externalities, on the other hand, are embodied within the theory.
5. Since they are unpriced, the market has failed and therefore there is no market; we treat them as technological as opposed to pecuniary externalities. They are concerned with the technical relationship between inputs and outputs, with the structure of the production function rather than with prices in a market. If the congestion effect was due to outputs so that the firm's costs rise as the output of other firms rises, then we would have a pecuniary externality. However since congestion in a region is caused by the number of firms producing there, an input measure, it is a technological externality.
6. We can distinguish between urbanization externalities or Jacobs externalities (after Jane Jacobs) and localization externalities or Marshall–Arrow–Romer (MAR) externalities. While both involve knowledge spillovers between firms, MAR externalities restrict the spillover to firms in the same industry whereas Jacobs externalities refer to spillovers across different industries.
7. This has been questioned by Lucas (1988), who maintains that human capital is an input like any other, so that the level of output should depend on the level of human capital, and consequently the growth of output should depend on the rate of human capital accumulation.
8. Our (unreported) estimates of alternative model specifications including the interaction of G and H indicate that the interaction term is not a significant variable.
9. Or as Cameron (1996) puts it, a 'standing on shoulders effect' due to knowledge leaks and imperfect patenting, in addition to the movement of skilled labour to other firms reducing rivals' costs.
10. See Maddala, 2001, for details.
11. There are separate estimates of error variance for the two groups ($i=1,2$) of regions defined as core (within 500 km of Luxembourg) and peripheral regions.
12. P^* is the maximum value of P in 1975.
13. For related work, see Bianchi, 1997; Fingleton, 1997; Magrini, 1999; López-Bazo *et al.*, 1999; Johnson, 2000; Lamo, 2000.

[14] See Durlauf and Quah (1999) for a formal definition and some properties of stochastic kernels in the study of distribution dynamics

[15] A Gaussian kernel and bandwidth as described in Silverman (1986, section 4.3.2) was applied to estimate the joint densities. The Gauss procedure used to estimate the joint densities is that created by G. Suettrim.

[16] 'A Dialogue between Economists and Geographers', 24–26 October 2003, supported by the European Science Foundation Exploratory and organized by the Centre for Economic Performance and the Centre for Economic Policy Research, London.

REFERENCES

Abdel-Rahman, H. and M. Fujita (1990), 'Product variety, Marshallian externalities, and city sizes', *Journal of Regional Science*, **30**, 165–83.

Anselin, L. (1988), *Spatial Econometrics: Methods and Models*, Dordrecht: Kluwer.

Anselin, L. and H.H. Kelejian (1997), 'Testing for spatial error autocorrelation in the presence of endogenous regressors', *International Regional Science Review*, **20**, 153–82.

Barro, R. (1991), 'Economic growth in a cross section of countries', *Quarterly Journal of Economics*, **106**, 407–43.

Barro, R. and X. Sala-i-Martin (1992), 'Convergence', *Journal of Political Economy*, **100**, 223–51.

Benhabib, J. and M. Spiegel (1994), 'The role of human capital in economic development: evidence from aggregate cross-country data', *Journal of Monetary Economics*, **34**, 143–73.

Bianchi, M. (1997), 'Testing for convergence: evidence from non-parametric multimodality tests', *Journal of Applied Econometrics*, **12**, 393–409.

Brakman, S., H. Garretsen and M. Schramm (2004), 'The spatial distribution of wages and employment: estimating the Helpman-Hanson model for Germany', *Journal of Regional Science*, **44**, 437–66.

Brülhart, M. (1998), 'Economic geography, industry location and trade: the evidence', *World Economy*, **21**, 775–801.

Cameron, G. (1996), *Innovation and Economic Growth*, DPhil thesis, Ch.2, University of Oxford.

Ciccone, A. and R.E. Hall (1996), 'Productivity and the density of economic activity', *American Economic Review*, **86**, 54–70.

Cheshire, P.C. and D.G. Hay (1989), *Urban problems in Western Europe: an Economic Analysis*, London: Unwin Hyman.

Durlauf, S.N. and D. Quah (1999), 'The new empirics of economic growth', in J. Taylor and M. Woodford (eds), *Handbook of Macroeconomics*, North-Holland, New York and Oxford: Elsevier Science, pp. 235–308.

Fingleton, B. (1997), 'Specification and testing of Markov chain models: an application to convergence in the European Union', *Oxford Bulletin of Economics and Statistics*, **59**, 385–403.

Fingleton, B. (2000), 'Spatial econometrics, economic geography, dynamics and equilibrium: a third way?', *Environment & Planning A*, **32**, 1481–98.

Fingleton, B. (2001a), 'Theoretical economic geography and spatial econometrics: dynamic perspectives', *Journal of Economic Geography*, **1**, 201–25.

Fingleton, B. (2001b), 'Equilibrium and economic growth: spatial econometric models and simulations', *Journal of Regional Science*, **41**, 117–48.

Fingleton, B. (2003), 'Increasing returns: evidence from local wage rates in Great Britain', *Oxford Economic Papers*, **55**, 716–39.

Fingleton, B. and J. McCombie (1998), 'Increasing returns and economic growth: some evidence for manufacturing from the European Union regions', *Oxford Economic Papers*, **50**, 89–105.

Fujita, M., P. Krugman and A.J. Venables (1999), *The Spatial Economy: Cities, Regions, and International Trade*, Cambridge and London: MIT Press.

Fujita, M. and J.-F. Thisse (1996), 'Economics of agglomeration', *Journal of the Japanese and International Economies*, **10**, 339–78.

Glaeser, E.L. (1999), 'Learning in cities', *Journal of Urban Economics*, **46**, 254–77.

Gordon, I. and P. McCann (2000), 'Industrial clusters: complexes, agglomeration and/or social networks?', *Urban Studies*, **37**, 513–32.

Hanson, G.H. (1997), 'Increasing returns, trade, and the regional structure of wages', *Economic Journal*, **107**, 113–33.

Henderson, J.V. and J.-F. Thisse (eds) (2003), *Handbook of Urban and Regional Economics*.

Jacobs, J. (1969), *The Economy of Cities*, New York: Random House.

Johnson, P.A. (2000), 'A nonparametric analysis of income convergence across the US States', *Economics Letters*, **69**, 219–23.

Johnston, J. (1984), *Econometric Methods*, New York: McGraw Hill.

Kennedy, P. (1992), *A Guide to Econometrics*, Oxford: Blackwell.

Koutsoyiannis, A. (1977), *Theory of Econometrics*, London: Macmillan.

Krugman, P. (1991), *Geography and Trade*, Leuven: Leuven University Press.

Lamo, A. (2000), 'On convergence empirics: some evidence for Spanish regions', *Investigaciones Economicas*, **24**, 681–707.

Leser, C. (1966), *Econometric Techniques and Problems*, London: Griffin.

Lucas, R. (1988), 'On the mechanics of economic development', *Journal of Monetary Economics*, **22**, 3–42.

López-Bazo, E., E. Vayá, A.J. Mora and J. Suriñach (1999), 'Regional economic dynamics and convergence in the European Union', *The Annals of Regional Science*, **33**, 343–70.

Maddala, G.S. (2001), *An Introduction to Econometrics*, New York: Wiley.

Magrini, S. (1999), 'The evolution of income disparities among the regions of the European Union', *Regional Science and Urban Economics*, **29**, 257–81.

Mankiw, N.G., D. Romer and D.N. Weil (1992), 'A contribution to the empirics of economic growth', *Quarterly Journal of Economics*, **107**, 407–37.

Nelson R. and E. Phelps (1966), 'Investment in humans, technological diffusion, and economic growth', *American Economic Review: Papers and Proceedings*, **51**, 69–75.

Ottaviano G.I.P. and D. Puga (1998), 'Agglomeration in the global economy: a survey of the new economic geography', *World Economy*, **21**, 707–31.

Quah, D. (1993), 'Empirical cross-section dynamics in economic growth', *European Economic Review*, **37**, 426–34.

Quah, D. (1996a), 'Convergence empirics across economies with (some) capital mobility', *Journal of Economic Growth*, **1**, 95–124.

Quah, D. (1996b), 'Regional convergence clusters across Europe', *European Economic Review*, **40**, 951–58.

Quigley, J. (1998), 'Urban diversity and economic growth', *Journal of Economic Perspectives*, **12**, 127–38.

Rivera-Batiz, F. (1988), 'Increasing returns, monopolistic competition, and agglomeration economies in consumption and production', *Regional Science and Urban Economics*, **18**, 125–53.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, New York: Chapman and Hall.

Stokey, N. and R.E. Lucas Jr (1989), *Recursive Methods in Economic Dynamics*, Cambridge, MA: Harvard University Press.

Veblen, T. (1915), *Imperial Germany and the Industrial Revolution*, London: Macmillan.

12. The role of wage-setting in a growth strategy for Europe[1]

Andrew Watt

INTRODUCTION

Economic performance in Europe has been disappointing since 2000, when EU heads of state and government agreed, at the Lisbon European Council, to make the European Union 'the most competitive and dynamic knowledge-based economy in the world, capable of sustainable economic growth with more and better jobs and greater social cohesion'. In the Euro Area, economic growth averaged just 1.4% from 2001 to 2004, compared with 2.7% from 1996 to 2000. Employment growth in the same periods also halved and the rate of unemployment rose by a full percentage point from an already unacceptably high level.

The explanation of this dismal performance offered by mainstream economists, and accepted by parties across much of the political spectrum, is simple: labour market and other institutional imperfections – and thus a 'lack of structural reform' – are making it impossible for Europe to compete against more flexible advanced capitalist countries (like the US) and the rising low-wage economies of Asia, while such institutional rigidities raise unemployment by maintaining wages above 'equilibrium' levels.

Ultimately it matters little that the facts (not least Europe's rather good economic performance in the late 1990s and the growth and employment patterns *within* Europe) do not support this simplistic view. European governments, spurred on by the Lisbon Strategy, have embarked on a series of liberalising reforms, particularly in the area of labour market and welfare state institutions. To date these reforms appear, if anything, to have worsened an already difficult situation. At the time of writing, Germany, which accounts for one third of EMU output, appears set to embark on a further round of cutbacks and tax increases; meanwhile the European Central Bank (ECB) has begun to raise interest rates. It is hard to see how, under such conditions, the fragile upturn that had just begun to manifest itself after five years of virtual economic stagnation can be sustained.

From both the academic and political margins, there has been a sustained critique of this mainstream view: Europe's economic woes are due not to its labour market institutions or other 'market rigidities', but, largely, to inappropriate macroeconomic policy. Both the monetary policy of the ECB and national fiscal policy, as constrained by the Stability and Growth Pact, have come in for criticism, and a range of reform proposals made.

This contribution is located broadly within that second, Keynesian-inspired tradition. However, it deals less with macroeconomic – monetary and fiscal – policy itself, focusing rather on the contribution that wage-setting can play, *in conjunction with* a more expansionary macroeconomic policy stance, in addressing Europe's economic and employment growth problems. The aim is to set out the role that wage-setters along with the monetary and fiscal authorities would need to play to achieve higher economic growth over an extended period, bringing Europe close to what might be considered 'full employment'. Attention will be paid to the coordination requirements of such a strategy.

The proposals made here need to be seen in the light of 'political economy' constraints. As I have argued in previous work, the Maastricht architecture is effectively 'set in stone' (Watt, 2005). Fundamental changes to that 'regime' are unlikely, except, possibly, by way of a major economic crisis, whose political outcomes might well prove less rather than more favourable. Consequently, an attempt is made here to build on existing institutions and to argue, where possible, in ways that can be related to elements of mainstream thinking and policymaking.

The paper is structured as follows. Section 1 briefly reviews the treatment of wage-setting and wages (or incomes) 'policies' in post-Keynesian thought, showing how the focus has moved away from incomes policies since the end of 'full employment capitalism' in the 1980s. It also discusses the role of the non-accelerating inflation rate of unemployment (NAIRU) in mainstream thinking, and how it might be conceived in a Keynesian-inspired policy perspective. Section 2 presents a simple model of the interaction between 'wage policy' and 'macroeconomic demand policy' in the context of a monetary union. An optimal trajectory for nominal wage growth and nominal demand is set out. Section 3 considers the extent to which such an optimal constellation can be realised in a real-world situation such as EMU and what reforms this would require. The paper concludes with some brief reflections on policy and theorising in the Keynesian tradition.

1. WAGE-SETTING IN POST-KEYNESIAN THOUGHT AND THE ROLE OF THE NAIRU

In 1979, just before the shift in economic policy paradigm that followed the elections of Thatcher and Reagan, Alfred Eichner edited *A Guide to Post-Keynesian Economics*, one of the first overviews of post-Keynesian thinking. In his Introduction Eichner describes as 'one point on which economists with a post-Keynesian perspective are likely to agree':

> The conventional policy instruments . . . do not moderate, except most imperfectly, the income claims against available output so that the growth of nominal income over time will be equal to the growth of real income, without the need for rising prices to bring the two into balance. It is for this reason that post-Keynesian economists, instead of asking whether an incomes policy is necessary, have generally moved on to the question of how an incomes policy can be made to work effectively and equitably (Eichner, 1979: 17)

Most of the other contributors, whether discussing income distribution, pricing, or the labour market, also refer to the need for (and in some cases the problems of) incomes policies. While expressing some scepticism about controlling wages, Basil Moore closes with three basic policy alternatives:

> Continuing and possibly increasing wage inflation . . . . A slump and a massive rise in unemployment to keep money wage increases low. Or some sort of incomes policy. These three alternatives exhaust the set. There are no other games in town. (Eichner, 1979: 138)

Moreover, closing the book with 'A look ahead', Eichner states:

> The preceding essays have been like a chorus in arguing that inflation cannot be brought under control – except at too great a cost in terms of reduced output and higher unemployment – unless the conventional policy instruments for regulating the economy are supplemented by an incomes policy (Eichner, 1979: 174).

The Elgar Companion to Post-Keynesian Economics, edited by J.E. King can be seen as a modern equivalent to Eichner's book, providing an overview of post-Keynesian thought at the start of the 21st century (and with contributions from a number of the authors in the earlier book). Of the more than seventy-five entries, there is no entry devoted to 'incomes policy' itself. There is a discussion of Weintraub's proposal for a 'tax-based incomes policy'. But even this is couched in terms of historical interest: Weintraub's work may prove useful 'should stagflation return' (Seidman in King, 2003: 336). Apart from that there are only fleeting references to wages policy, again often in an historical context. Discussing 'Economic policy', in the same book, Malcolm Sawyer does mention incomes policy, as being supported by 'some' post-Keynesians, but the emphasis is placed clearly on policies to ensure sufficient investment to prevent inflationary bottlenecks.

Numerous other examples could be given of the preoccupation of earlier post-Keynesian economists.[2] The interesting question in the present context, though, is what explains the virtual abandonment or at least sidelining of this line of enquiry by most contemporary post-Keynesians, illustrated by the *Elgar Companion*.[3]

One obvious point is the decline in inflation. Incomes or wages policies were seen primarily as a means to reduce inflation, which in the 1970s was unacceptably high. Now that inflation has been conquered – albeit using hugely costly deflationary macroeconomic policies – Keynesian-oriented economists have lost interest in incomes policies as a means to reduce it.

Another possible reason lies in changes in collective bargaining systems. In many countries, and especially in the UK and the US, collective negotiation of wages, which is a prerequisite for getting a 'handle' on nominal wage developments, has been eroded in favour of more 'decentralised' forms of wage bargaining; see for example in Traxler *et al.* (2001) and Schulten (2004). This has been accompanied by declines in the union density and collective bargaining coverage indicators compiled by the OECD. Thus it can be argued that, even if a wages policy can in theory be a useful weapon in policymakers' armoury, it is not a practical alternative because the institutional basis is lacking.

A third reason is a concern that a wages policy is one-sidedly directed at labour, and will thus tend to promote a shift (or exacerbate an existing trend) in the functional distribution of income from labour to capital. To the extent that such a redistribution, via its effects on demand, is also held to be detrimental to growth and employment, such a policy is claimed to be both unjust and, ultimately, ineffective.

Yet arguments can be adduced against all three positions. To the first, I argue presently that a policy of ensuring appropriate net wage-setting is, at heart, a policy for growth and employment. I discuss the other two arguments more fully in Section 3. Suffice it to say at this stage that developments during the 1990s show that the trend to bargaining decentralisation is neither pervasive nor irreversible, and one of the main aims of focusing on a guideline for nominal wage growth is precisely to prevent a further deterioration in the functional distribution of income.

Unemployment and inflation can be seen as two sides of the same coin in a capitalist economy. This is evident from the post-Keynesian quotes above (and is, in principle, not a matter of dispute between different schools of economic thought). Thus in principle, an incomes policy approach that is suitable to reduce inflation without causing additional unemployment –

the issue in the 1980s – is also suitable to reduce unemployment without re-igniting inflation, the issue of today. Indeed, given the current institutional and political environment, *any* feasible proposal for faster growth and employment *must* address the issue of how inflation can be kept in check (see also Allsopp, 2006).

This puts centre stage the concept of the non-accelerating inflation rate of unemployment (NAIRU) to which many post-Keynesians are highly averse. Without entering into the extensive debate on this concept (amongst others, see Layard *et al.*, 1991; Galbraith, 1997; Sawyer, 2001; Hein, 2004; Stockhammer, 2004) a number of points are relevant to the analysis here. The NAIRU concept is used by mainstream economists to justify a focus on dismantling supposed labour market rigidities as the solution to unemployment. This requires (at least) that for any economy the NAIRU is: (a) a reflection of those rigidities, and (b) its position is known. Stockhammer (2004: 56 ff) terms this the 'NAIRU story'. All Keynesians reject this approach and the concomitant policy conclusions.

However, as Stockhammer and other post-Keynesians recognise, the NAIRU concept itself is very close to Keynesian ideas of inflation resulting from social conflict over incomes (rather than being caused by changes in the money supply). The point here is that, in conjunction with a central bank with a sole mandate to control inflation (and, in my view, the power to do so under most conditions), the conflict theory of inflation becomes a conflict theory of unemployment. The NAIRU does play a role in the 'story' told below, but a very different one from the mainstream narrative.

2. A MODEL OF MONETARY-WAGE POLICY COORDINATION IN AN EMU-TYPE CONTEXT

This section[4] describes a simple model of how actors could behave to ensure a consistent policy mix that maximises growth and employment opportunities while ensuring price stability. It begins by abstracting from national differences, considering a single economy with a single monetary, fiscal and wage policy. In a second step we move closer towards the reality of EMU, with a single monetary policy, but national fiscal and wage policies.

a. A Single-Country Model

In this model it is assumed that, together, the public authorities can determine, in a medium-run perspective, the rate of growth of nominal demand in the economy. Simplifying further, the central bank is assumed to set short-run nominal interest rates in such a way that aggregate nominal demand (*M*) expands at a given rate (*m* – throughout rates of change are indicated by lower case letters). This aggregate nominal demand has a quantity and a price component,[5] so that changes in nominal demand are the sum of changes in real output (*y*) and prices (*p*).

$$m = y + p \text{ or } y = m - p.$$

Output is also defined as labour input (employment, *E*) multiplied by the productivity of labour (*Y*/*E*),

$$Y = E * Y/E$$

Thus changes in output (economic growth) are also equal to the sum of the change in employment and in productivity (π), or

$$y = e + \pi$$

Combining the two equations for y and rearranging we obtain our basic equation:

$$e = m - p - \pi$$

that is, the rate of employment growth is equal to the growth of nominal demand less inflation, less the rate of productivity growth. We can now consider, in this simple model, 'optimal' behaviour by the different actors.

The task of an inflation-targeting central bank, as in the ECB mandate, is to ensure that inflation stays, in the medium term, close to a target (*p**).

The question is, what determines how nominal demand is transposed into increases in *y* and *p*. In standard models it is the level of domestic aggregate demand with respect to the existing productive potential that does this.[6] Assume that, at the outset, the rate of inflation is constant at the central bank's target. At this point the stock of existing capital is at its 'normal' capacity utilisation and the level of unemployment is such that trade unions are sufficiently weakened to prevent them pushing through inflationary wage increases, firms cannot raise prices, and the economy is considered to be in equilibrium: inflation will be constant at the target rate (*p**), output will be equal to productive potential (*Y**) and unemployment will be at the NAIRU. In the standard model, this is nirvana: macroeconomic policymakers can do no better than this. If the authorities expand nominal demand beyond this point inflation will result (*m* will no longer raise *y* but merely *p*). It is the lack of pricing power of workers and firms (wage- and price-setters) resulting from ‘sufficiently high’ unemployment that ensures price stability.

Clearly this standard model assumes – among other things – that wages are determined in a simple way such that, when unemployment is above the estimated NAIRU, the growth of nominal wages is higher than the sum of productivity and the current rate of inflation, and below that sum when unemployment is below it. In a market-driven wage-setting environment (provided the NAIRU estimate is ‘right’ and everyone, the ‘representative’ wage-setter and the central bank believes in it) this may be true.

However, in reality nominal wages are set in complex institutional structures. Particularly in highly organised, centralised bargaining environments, which remain typical in much of Europe (Schulten, 2004) the ‘social partners’ reach agreement on rates of nominal wage growth for thousands or even millions of workers at a time. Suppose that they can set this rate at will. Specifically, assume that wage-setters agree on (and are able to enforce) a formula whereby, whatever the current rate of inflation and level of demand, nominal wages increase at a rate equal to the rate of medium-run labour productivity growth plus the target inflation rate of the central bank. Subject to the further condition that the scope for price-setters to raise prices is tied to the rate of wage increases – in other words that in the medium term the capital and wage shares of national income and thus the mark-up of prices over costs are constant – in such an environment, when M increases Y will increase by the rate of change of M, *whatever that rate is*, minus the target inflation rate of the central bank. Thus we can write:

$$y = m - p^*$$

Under these conditions, the labour costs of producing a unit of output are the decisive variable in determining inflation and thus the extent to which rising nominal aggregate demand is ‘lost’ to price increases rather than raising output and employment.

The two variables determining nominal unit labour costs (ULCs) are the growth of overall labour costs and of productivity. In the short to medium run, the growth of productivity is relatively insensitive to policy influence.[7] Thus nominal wage growth becomes the decisive variable determining the distribution of M between Y and P. Formally we arrive at the simple equation that, given the above assumptions,

$$e = m - w$$

In other words, employment growth is equal to the rate of nominal demand growth minus the rate of wage growth (see Koll, 2005: 189). Mathematically this result is obtained by inserting the assumed $w = \pi + p^*$ into our basic equation $e = m - p - \pi$.

This, in turn, puts the institutional mechanisms of (nominal) wage determination and nominal demand creation centre stage. Provided wage-setters (can) set nominal wage growth at the rate of productivity growth plus the target inflation rate of the central bank, the central bank is able to set interest rates at the level that expands nominal demand at the rate required to hit a target rate of growth for the economy. For a given productivity trend, this also determines the rate of employment growth.

An important implication of this model is that, subject to its conditions, the NAIRU, as traditionally understood, loses its role as a guideline for monetary policy (see Hein, 2004). So-called 'structural reforms' (lowering unemployment benefits, weakening trade unions and so on) whose aim, in different ways, is to reduce the NAIRU become superfluous, if not harmful. This is the result, of course, of the assumed ability of wage-setters to set the pace of nominal wage growth autonomously. For this model and the policy prescriptions associated with it to be considered relevant, we cannot duck the question of how long a process of nominal demand expansion *cum* stability-oriented wage development can continue. Otherwise, it would appear that there is no limit to the increase in output and employment. We return to this question in Section 3a. Before doing so, we must consider some implications of the fact that, in EMU, wages and fiscal policies are largely set at national level, while monetary policy is set at the level of the currency area as a whole.

b. A Multi-Level Model

Moving one step closer to reality, any consideration of the case of a single monetary policy with multiple wage-setting and fiscal 'authorities' gives rise to some complications. They are not those that might appear at first sight, however. It is frequently argued, for instance, that both productivity levels and trends and collective bargaining institutions in Europe are too diverse to permit wage coordination. This argument rests on a misunderstanding, however. Such diversity does not pose problems in itself: all that is required is for aggregate wage trends at the national level to conform to the national-productivity-plus-target-inflation-rate rule. This is easily shown.

Consider two countries in a monetary union. Let the rate of productivity growth in the first country be 2%, and in the second 3%. The (common) target inflation rate is 2%. Then a sufficient condition for medium-run price stability (and also for an unchanged competitive position between the two countries) is a nominal wage increase of 4% and 5% respectively.

For a currency area (*CA*) of *n* countries (*a*, *b* . . . *n*) we can write:

$$w_a = \pi_a + p^*, w_b = \pi_b + p^* \dots\dots\dots w_n = \pi_n + p^* \Rightarrow w_{CA} = \pi_{CA} + p^*$$

Clearly, the result for prices in the currency area is independent of the relative size of the countries, as unit labour costs in all countries will grow at the same rate, namely the target rate of the central bank. Moreover, the institutional arrangement that generates this outcome in each country is, in principle, irrelevant.[8]

Matters are more complex, however. So far we have talked only in terms of average inflation rates. While this is the key concern of the central bank, this overall figure consists of the weighted average inflation rates in the member states. For various reasons these are likely to differ, and, moreover, the patterns of such differences will also change over time. This national rate is likely to be of greater interest to national wage-setters and fiscal policymakers. It can be argued, as in the above equation, that if all actors, in their respective national contexts, stick to the overall guideline based on the common inflation target, these inflation differentials will disappear. This is logically correct. However, it assumes that inflation differentials are 'a bad thing', that they should be eliminated. Or, to put it another way, that the initial competitive position of countries (their real exchange rates) in the currency area is in equilibrium and also remains that way. This is unlikely to be the case, however. Differential inflation rates remain, even within a developed monetary union, and certainly within an 'immature' union, an important adjustment mechanism for national economies.

Two cases can be considered.[9] The first is where countries enter the monetary union at an incorrect real exchange rate. Countries that enter at too high (low) a rate will have to undergo a period of below-average (above-average) inflation if they are to regain competitive equilibrium. If this adjustment is blocked by adhering to the above wage norms, the former countries will suffer higher (lower) unemployment, with knock-on and probably pro-cyclical effects on fiscal policy. The second case is where, even though countries enter at the right rate, subsequent developments necessitate an adjustment of the real exchange rate. Again, two main possibilities come to mind. One is the case of an asymmetric shock, a shift in commodity prices or a shift in demand for certain products, that disproportionately affect certain countries of the currency union. The other is the need to allow for what might be called 'historical' adjustment mechanisms. The obvious example here is the Belassa–Samuelson effect. Countries undergoing a catch-up phase tend to have a lower domestic price level at the exchange rate that ensures external balance. As their productivity in the traded sector rises, this pulls up wage levels also in the non-traded-goods sector, and the price level rises. This is a normal and welcome adjustment process, and the wage norm should not seek to counteract it (and the inflation target of the central bank should be high enough to allow it).

On the other hand, it is clearly not the case that national wage norms should focus on the current national inflation rate. This would perpetuate inflationary (or disinflationary) processes that result from imported inflation, overheating and the like, and destroy the inflation-containing properties of the model.

From this we can draw the provisional conclusion that the national price component in the wage norm should normally lie *between* the central bank target rate and the national inflation rate. What is required is that the (weighted) average of the price components in national wage norms is consistent with the overall price target, that is, that countries in which wage and price inflation is above average are offset by those in which it is below, and that this reflects necessary adjustments in competitive position (and thus that they come to an end or are reversed as circumstances change).

Thus the wage-policy condition for a currency area with countries a to n needs to be rewritten as follows:

$$w_{\mathrm{a}} = \pi_{\mathrm{a}} + p_{\mathrm{a}}{}^{*},\ w_{\mathrm{b}} = \pi_{\mathrm{b}} + p_{\mathrm{b}}{}^{*} \ldots\ldots\ldots w_{\mathrm{n}} = \pi_{\mathrm{n}} + p_{\mathrm{n}}{}^{*} \Rightarrow w_{\mathrm{CA}} = \pi_{\mathrm{CA}} + p^{*}_{\mathrm{CA}}$$

where $p_{\mathrm{n}}{}^{*}$ represents the country-specific target inflation rate for country *n*. In each case this rate will lie between the current national inflation rate and the target rate for the currency area as a whole, in other words either $p_{\mathrm{a}} < p_{\mathrm{a}}{}^{*} < p^{*}_{\mathrm{CA}}$, or $p_{\mathrm{b}} > p_{\mathrm{b}}{}^{*} > p^{*}_{\mathrm{CA}}$ for below and above-average inflation countries respectively. In addition the weighted average of the national price components must be equal to the overall price target: $\alpha p_{\mathrm{a}}{}^{*} + \beta p_{\mathrm{b}}{}^{*} + \ldots \theta p_{\mathrm{n}}{}^{*} = p^{*}_{\mathrm{CA}}$ where α, β, . . ., θ represent the relative weights of the countries in the inflation 'basket' of the central bank (and sum to 1).

c. A Consistent Trajectory of Nominal Wages and Nominal Demand for the Euro Area in such a Model

The model is clearly a gross simplification of reality. Nevertheless, before turning to consider ways in which it might be implemented, at least partially, in the real-world situation of EMU, it is useful to plug some numbers into the model, to see the orders of magnitude involved. Given that since 2000 the economic and employment policies of the EU and its member states are supposed to be occurring under the umbrella of the Lisbon Strategy, the parameters of that strategy are taken – arbitrarily – as the normative point of reference.

The Lisbon Strategy – running from 2000 to 2010 – was predicated on achieving economic growth of 3% and employment growth of 1% per annum (p.a.), implying productivity growth of 2%.[10] Price stability is defined as a medium-run ceiling of 2% p.a. increase in the price index (*HICP*). The above equation ($e = m–w$) indicates that, if productivity remains unchanged, there is only one consistent trajectory for the other variables: the rate of nominal demand growth should be around 5% p.a. on average, while nominal wages for the Euro Area as a whole should increase at around 4%. As indicated above, nominal wage growth in individual countries should be somewhat higher or lower to permit necessary adjustment processes.

Regarding the appropriate trajectories at national level, it is difficult to be more precise about the 'correct' price component for the wage settlement in each country. Determining the degree of intra-area adjustment required is an empirical matter. For instance, while some authors have expressed scepticism concerning the quantitative importance of the Belassa–Samuelson effect in the Euro Area (such as DIW, 2005), differences in national price levels remain significant. Eurostat purchasing power parity estimates indicate that between 1999 and 2004, the price level fell in Germany by around five percentage points (p.p.) with respect to the EU15 average (from roughly 110% to 105%) whereas it rose in Spain by about the same amount (from roughly 80% to 85%). On this basis, and assuming a slow but consistent trend towards a more equal price level within a monetary union, an extended period of above-average unit labour cost and price increases in Spain and lower-than-average in Germany would be expected and justified.

On the one hand, the need to allow for competitive adjustment makes it harder to decide on the appropriate quantitative guideline for national wage policy in any given circumstances. The lack of clarity about whether prevailing inflation differentials are justified or need to be counteracted by wage policy will exacerbate the already difficult task of coordinating wage bargaining. On the other hand, such a guideline will be easier to follow in the sense that it reduces the extent to which social partners or trade unions need to impose settlements on their members, because the distance between the pay norm and the rate that market pressures will be pushing towards will be less – in both directions – if the price component of the wage increase is closer to the current country-specific inflation rate.[11]

We have so far considered the monetary union as a closed economy. However, prices depend not only on domestic costs but also on changes in the prices of imports. This need not be modelled explicitly. In fact, domestic actors should retain their (medium-run) orientation irrespective of changes in import prices. Consider the case of rising oil prices. In the past this has led either to an attempt to raise nominal wage demands with an initially accommodative but subsequently all the more restrictive demand-side policy (a typical reaction pattern in the 1970s and 1980s) or a non-reaction by nominal wages but nevertheless a significant tightening of monetary policy, as occurred in 2000 in the Euro Area.[12]

In such a case, if aggregate nominal wage growth continues to be oriented towards trend productivity growth plus the target inflation rate, higher import prices will not be passed through into wages and domestic prices. Headline inflation will initially rise, but any increase will be contained by the 'anchor' function that nominal unit labour costs have on medium-run inflation.[13] There will, in short, be no second-round effects, and no need to tighten macropolicy. It is important that this neutrality by wage and aggregate demand policy is applied symmetrically, that is also in the case of a transitory deflationary external shock.

3. REAL-WORLD RELEVANCE OF THE MODEL

The model described above has certain properties derived from the identities used and, in particular, the assumptions made. This section considers the extent to which these assumptions either hold in the real world or can be made to do so by means of appropriate reforms and behavioural changes.

The relevance of the model for policy purposes can be called into question along five main lines:

a. the public authorities (or the central bank) cannot control the path of nominal aggregate demand
b. the social partners (or trade unions) cannot control the path of nominal wages
c. the link between wages and prices is too unreliable (control of wages does not ensure control of prices)
d. productivity is endogenous and cannot be assumed constant
e. the model requires coordination mechanisms between the actors that do not exist and are unlikely to be developed.

a. Controlling Nominal Demand

Post-Keynesian economists are united in, amongst other things, their rejection of theories of money and monetary policy centred on the exogenous control of monetary aggregates by the central bank, and their rejection of the neutrality of money. Beyond that there are considerable differences of opinion on issues such as the effectiveness of monetary policy in demand management. The proposal advocated here is consistent with what is consensual in the post-Keynesian view.[14] On the effectiveness of monetary policy, it sides with those that believe that, at least in a large advanced economy, like the Euro Area, the central bank can steer nominal demand reasonably effectively in a medium-term perspective. It is clearly a difficult matter to assess the quantitative impact of monetary policy, because of the difficulty of isolating it from other influences, the self-fulfilling role of expectations in determining that influence, and so on. The ECB assumes an impact on real GDP of about 0.6 p.p. after three years from a 0.5 p.p. rise in base rates (ECB, 2002: 56). Whatever the precise estimate the key point is that the central bank can change rates costlessly, by any amount (subject to a lower bound) and at any time. This gives it operational advantages over fiscal policy. So even if the effect is weak, it merely means that rates must be shifted more often or more substantially. Certainly the ECB would find it easier to manage demand (and to cope with intra-area adjustment needs) if the inflation target were somewhat higher.[15] Of course the bank cannot reduce nominal rates below zero. But even here, it can purchase a whole range of assets from banks for central bank money, injecting nominal demand into the economy.

Clearly, monetary policy is not omnipotent. Its impact is always clouded by some uncertainty and is subject to lags. Moreover, while 'dear money' can be counted on to arrest a boom, if the 'animal spirits' of investors are sufficiently depressed, no amount of 'cheap money' will turn the economy around. In such situations there is a clear case for direct deficit-financed spending by government.

In principle, the model could operate with fiscal policymakers setting nominal demand growth. However, I tend to the view that, in the medium run and in 'normal' circumstances, because of its greater flexibility monetary policy should normally take the lead, at the aggregate level, in managing nominal demand. In a monetary union such as EMU, especially, this would also facilitate the policy coordination process and require less of a change compared with the current 'regime'. There is thus no disagreement of principle with those post-Keynesians who are more sceptical about the capacity of monetary policy to generate sufficient aggregate demand.[16] The choice of monetary versus fiscal policy to manage demand can be seen as a balancing act, the outcome of which will depend on a number of specific features of the economy concerned at any given time.

What, then, is the role of fiscal policy, in such a model, and in the specific context of EMU? In a unified national context, there is, in my view, a strong case for a 'golden rule' for fiscal policy in 'normal times'.[17] This permits deficit-financing of 'capital spending'[18] while allowing the automatic stabilisers to help smooth the cycle. Such an approach also reduces the risk of conflicts between monetary and fiscal policy concerning responsibility for demand management.

However, fiscal policy certainly needs to play a much more active role in the context of a monetary union. Given a common interest and exchange rate, it represents – along with the real exchange rate set, primarily, by wage-setters – the one main instrument left to national policymakers to offset shocks and promote necessary intra-area adjustments. It is mainly by constraining governments' ability to use this instrument that the Stability and Growth Pact has done so much damage. National fiscal policy can play a potentially important role in the model suggested here by using the national Phillips curve to help wages adjust to the desired national trajectory (for a more developed wage-fiscal strategy at national level see Hancke and Soskice, 2003). Indeed for countries lacking appropriate collective bargaining institutions this is the only way to steer nominal wage growth, and the policy recommendations linked to the 'NAIRU story', with all their negative social implications, become difficult to escape.

b. Controlling Nominal Wages

In the discussion so far, we have heroically assumed that nominal wages are autonomously set by wage-setters, implying a monopsonistic trade union, or at least a highly centralised and cooperative collective bargaining system able to prevent individual wage bargains that contravene the agreed norm. Wage-setters can thus make a credible commitment to other actors to ensure a given rate of nominal wage increases. Although there is no space here to discuss collective bargaining structures in detail (Schulten, 2004; Traxler *et al.*, 2001; Janssen and Mermet, 2003), both common sense and historical experience in national economies with social pacts, social contracts and the like suggest that, while a degree of control can be exerted by organised collective bargainers, that control is limited. Even if formal bargaining coverage is high, actual wages differ in practice from collectively agreed pay rates (wage drift). The 'devil is in the detail' and it is very difficult to determine the exact value of pay settlements in such a way that it can be compared with the wage norm. More basically, it is undisputed that an expansion of demand and falling unemployment will, at some point, lead to a breakdown of nominal pay discipline and inflationary pressure.

Is this a valid argument against the policy approach advocated here? It certainly means that the model will never work in reality in the 'perfect' way illustrated above. But in terms of real-world policy-making, the argument, while it certainly poses challenges, is not a fatal one. For it merely means that it is not possible in reality to bring the NAIRU to zero (to render it entirely indeterminate). Yet this is not necessary either. The litmus test on which the strategy, as a ‘policy recommendation’, must be judged is whether, by its use, *unemployment can be sustainably brought down below that prevailing under the existing, non-cooperative regime*.

Moreover, despite some recent trends towards decentralisation, most European workers continue to be covered by multi-employer collective agreements, typically at the sectoral level (Schulten, 2004). Contrary to what media reports might lead one to believe, not only does collective bargaining coverage remain high, but the 1990s saw a resurgence of ‘social pacts’, a new form of national corporatism that has led to a centralisation of wage negotiations in a number of European countries (Fajertag and Pochet, 2000). Indeed, there is an extensive literature suggesting that coordinated, centralised wage bargaining is associated with better macroeconomic outcomes (Traxler *et al.*, 2001; OECD, 1997; IMF, 2003; Howell, 2004). This is not least because a coordinated wage policy avoids the economic fluctuations that arise from using the national Phillips curve, in both directions, to bring the economy back to a sustainable path.

Meanwhile European trade unions are engaging in various activities to coordinate their wage demands and ensure that wage developments are consistent with non-inflationary growth, while ensuring workers a balanced share of rising national income. Space constraints preclude an extended description (see Schulten, 2004; Janssen and Mermet, 2003). Experience with these wage norms so far has been mixed. The coordination mechanisms rely on similar forces (benchmarking, peer pressure) as the EU’s ‘open method of coordination’ and suffer from the same limitations: they are fine in good times, but when under pressure, they exert little binding power over trade unions concerned primarily with national priorities and constraints. It is not yet the case that national union federations see such forms of coordination as being in their vital interest.

Ultimately, though, the extent to which nominal wage trends can be controlled cannot be known in advance. It must, in practice, be the subject of an iterative social experiment in which confidence is built between the actors (social partners and monetary authorities) and demand is expanded slowly to the point where wage pressures start to occur. This iterative process must be managed: that is why coordination mechanisms are needed to underpin the process (see sub-section e).

c. Price–Wage Link

I will address the issue of the wage–price link, and the extensive literature concerning the distribution between wages and profits, merely by way of reference to the empirical evidence. There is a very close empirical relationship between changes in nominal unit labour costs (ULCs) and changes in prices (Watt, 2006). This is very much in line with the Keynesian/post-Keynesian view that, as a class, workers cannot raise their real wage by raising nominal wages because the latter are the prime determinant of price developments (mark-up pricing).[19] It is true that the wage share has declined slowly but steadily in Europe during the last 20 years or so (ULCs below inflation). But this has reflected high unemployment in most countries: the wage share has stabilised or more recently risen in low unemployment countries such as the US and UK. Thus to the extent that growth is boosted and unemployment brought down within the strategy advocated, the link between ULCs and prices should become closer.

d. Productivity Impact

Productivity is assumed to be exogenous in the model purely for simplicity. Mainstream economists tend to assume that raising employment rates is associated with declining rates of productivity growth, because the additional labour brought into the production process will tend to be of below-average skill level. However, there are many reasons why faster demand and output growth might lead to faster productivity growth. These include higher capacity utilisation, economies of scale, faster growth of the capital stock and thus the incorporation of new technical knowledge, greater incentives to train workers and to undergo training, and so on (McCombie and Thirlwall in this volume, and Watt and Janssen, 2005).[20]

In any case, for the purposes of this exercise it is not necessary to take an *ex ante* view. Because of the difficulties of distinguishing between cyclical and structural changes in productivity, nominal wage growth should be based on medium-term productivity developments. This norm should be invariant to short-run changes, but be potentially adaptable once firm evidence emerges of a structural shift in productivity in either direction.

e. Need for Coordination

I have dealt with the issue of policy coordination at greater length elsewhere (Watt, 2006; Watt and Hallwirth, 2003). To put it most succinctly: a potentially appropriate coordination instrument – the Macroeconomic Dialogue (MED) – does already exist, and although it is currently not effective, it is politically feasible to make it so. Just as is the case with control over aggregate nominal demand and nominal wages, coordination mechanisms do not have to be perfect in order to generate growth and employment outcomes that are markedly superior to the prevailing situation.

The Macroeconomic Dialogue was established in 1999, just after the start of EMU.[21] Its aim is to contribute, via an improved macroeconomic policy, to a ‘sustainable reduction of unemployment’. The specific contribution of the MED is to institute a dialogue between the actors responsible for the policy mix – monetary, fiscal and ‘wage’ policies – to promote positive interaction between the actors.

The MED takes place twice a year at political level, in each case prepared by a meeting at technical level. The discussions are confidential and there is currently no provision for issuing formal statements or reports as an institution: ‘The substantive core of the MED is an exchange of information and ideas’ (Koll, 2005: 183). Participants discuss their analysis of the economic situation and prospects, formulate their own intended responses to the unfolding situation with a view to the goals of higher employment and non-inflationary growth and, lastly, state their expectations of how other actors should respond.

Thus, on the one hand the MED is clearly located – in the terms of our model – at the key nexus for improving growth and employment outcomes at the European level. If Europe suffers from coordination failures, especially those linking monetary and wage policies, then the MED is ‘in the right place’ to resolve them. On the other hand, it is extremely weakly institutionalised, characterised by a very loose form of ‘soft’ coordination. Thus its ‘purchase’ on actor behaviour is also extremely limited, even at the European level. On top of this comes the problem of the inadequate links between the European and the national level, where fiscal and wage policy decisions are very largely taken.

Reform proposals follow on from this:

- accelerate the rhythm of these meetings[22]
- shift their focus away from discussions about ‘the facts’ and their interpretation, to a more policy-oriented debate focused on consistent, quantitative development scenarios for the European economy and the mutually compatible actions by participants that are required to achieve them
- establish a permanent secretariat to manage coordination and oversee the accumulation of technical knowledge
- establish a parallel structure of national MEDs building on national social partnership traditions and structures but feeding into the EU-level MED.

The key concern must be to develop the institutional interactions and the expertise and knowledge at technical level on which an expanded Dialogue at political level would conduct an ongoing, intensified dialogue on macroeconomic issues. This would make the process more open, transparent and politically legitimate as its importance in policymaking increases. Alongside more regular and structured discussions, ways should be developed to enable informal coordination in response to sudden developments.

CONCLUDING REMARKS

Economic growth and employment outcomes in Europe, and especially in the Euro Area, have been disappointing, contributing not least to widespread disaffection with European integration. Economists in the Keynesian tradition are convinced that this reflects, above all else, failures of macroeconomic policymaking.

Numerous recommendations for fundamental reforms of the economic policymaking architecture have been made, in particular changes to the 'rules of the game' for fiscal and monetary policy. However welcome they would be in strictly economic terms, however, they face the huge political problem that altering rules and institutions that have been established by intergovernmental Treaty between 15 or more member governments requires unanimity (Watt, 2005: 238 ff.).

Drawing partially on a tradition that used to belong to the core of post-Keynesian economics, this contribution has sought to point out a path towards higher growth and employment without requiring Treaty changes, developing that tradition under the circumstances of contemporary EMU. Undeniably, the prerequisites of this strategy are also considerable. Wage-setting plays a central role, and Europe's trade unions are making efforts to establish the information, reporting and coordination procedures necessary. Some limited progress has been made. A decisive breakthrough, though, will depend on macroeconomic policymakers, and especially the central bank, manifesting an interest in such a cooperative approach to policymaking. Currently central bankers have – not unjustified – doubts about the current ability of trade unions at European and national levels to commit memberships at lower levels. Policy is therefore locked in a low-confidence trap that is harming growth and employment. Trade unions, under the pressure of high unemployment and attacks on collective bargaining and welfare states, have started down the road to a more cooperative strategy that would permit faster growth. Meanwhile political opposition to and pressure on the ECB are mounting, even in quarters (the financial media, finance ministers) that were until recently fiercely loyal to the idea of an independent, conservative central bank. It is to be hoped that such pressure will force the bank to also embark on this path.

As regards theory, the model presented and the approach advocated here are compatible with other strategies that focus, for instance, more on the use of fiscal policy and the conditions necessary to expand the capital stock. It does however suggest that Keynesian economists could usefully pay greater attention to wage determination in their theoretical work, where they can draw on a rich tradition.

NOTES

1. This article has benefited substantially from comments received during presentations in Cambridge, Brussels, Berlin and Bremen. The usual disclaimer applies.
2. Not least Joan Robinson: see essays 21 and 23 in Robinson 1979. In fact the tradition goes back to Kalecki's famous 1943 article 'Political aspects of full employment', in which Kalecki referred to full employment not only leading to the workers 'getting out of hand', but also to 'the price increase in the upswing'.
3. There are some exceptions in researchers with a direct link to the labour movement (for example Hein et al. (2005)). John Grieve Smith (2001: 114) also makes a brief reference to collective wage bargaining as part of his 'New economic agenda'.
4. This section develops previous work by the author (especially Watt 2006). See also Koll 2005.
5. The reader is reminded that this is a mathematical identity. The central bank sets interest rates and then the supply of money is determined endogenously by demanders of credit. The question of the ability of macroeconomic policy to control nominal demand is dealt with in Section 3.
6. In the model we abstract from external influences (exchange rates, import prices) on the domestic price level, but return to this important point in Section 3.
7. But see Section 3d below.
8. The actual value of w_{CA} and π_{CA} will depend, though, on the relative weights of the countries (see also below).
9. Further work is necessary to address this complex discussion in detail. The aim here is to set out some basic principles in the context of the policymaking approach advocated. Problems of competitiveness within the Euro Area and the appropriate adjustment mechanisms are beginning to tax the minds of economists and policymakers. For further discussion see Allsopp and Artis, 2003; DIW, 2005.
10. Although in the light of poor performance since 2001 these figures seem ambitious, they merely imply a continuation of what the EU achieved between 1997 and 2000. Employment growth of 1% p.a. is derived from the official goal of raising the employment rate to 70% over a ten-year period.
11. For instance, Spanish unions would not have to convince their members to base pay settlements on productivity plus 2% when price inflation in Spain is running at around 3.5%. That would imply real wage growth 1.5 p.p. below the rate of productivity growth, and that in a situation of a booming economy. Instead, allowing for adjustment effects, the wage guideline would be based on price inflation of, say, 3%. This would be offset by a lower target in, for instance, Germany, where unions find it very difficult to achieve pay increases as high as productivity plus 2%.
12. This helped bring to an untimely end the expansion of 1999–2000. Policy tightening occurred because of uncertainty on the part of the central bank about the future course of wage policy. This instructive episode is reviewed more fully in Watt and Hallwirth 2003.
13. In addition, unit labour costs will themselves be lower than if demand and output are curtailed by monetary tightening, as it is well established that in the short-to-medium run productivity is pro-cyclical. Cyclical falls in productivity caused by sudden drops in output are one reason why, in an *ex post* analysis, wages sometimes appear to overshoot in the year an economy begins to decline, leading to claims that wage policy has 'killed jobs'.
14. For instance: 'At the heart of Post Keynesian monetary policy, therefore, is not so much a body of technical analysis which cuts it off from the mainstream (or at least from its

more realistic practitioners) but a desire to rid the practice of policy from its deflationary biases, to reassert the value of discretion in responding to monetary shocks and to restore accountability in the conduct of monetary policy. At the heart of Post Keynesian policy is lower interest rates' (P. Howells in King, 2003: 260).

15. Note that this can be decided autonomously by the ECB and does not require treaty changes.
16. See for instance, Malcolm Sawyer's concept of the 'constant inflation level of output' (CILO) (Sawyer, 2005). The 'case for fiscal policy' is made by Arestis and Sawyer (2003). The former paper does, though, refer to the practical difficulties of using fiscal policy mentioned earlier.
17. As indicated above, if expectations are sufficiently depressed, only direct government spending is likely to get an economy out of recession.
18. Therefore, within this framework the level of public investment can be steadily expanded if considered desirable.
19. This view has just received confirmation by the ECB. In surveys covering almost the entire Euro Area 54% of surveyed firms reported setting prices as a margin over costs and 27% that they follow the lead set by competitors (ECB, 2005).
20. The US experience of the 1990s can be interpreted in these terms. Far from Greenspan 'seeing' higher productivity growth ahead of everyone else and then expanding demand, it seems more plausible to argue that his policy of low interest rates – for example, out of concern about the Asian crisis – was a proximate cause of the productivity increase in the 'roaring nineties'.
21. In theory, the MED covers the entire European Union. In practice, however, and also in this analysis, the focus is very much on EMU and its common monetary policy. See *The Presidency Conclusions of the Cologne European Council* (http://europa.eu.int/council/off/conclu/june99/june99_en.htm). For a more detailed description of the MED see in particular Koll, 2005.
22. An unpublished survey of ETUC affiliates conducted by the author revealed that, prior to EMU (and in some cases since), national union federations typically met with government and the central bank to discuss policy-mix issues on a monthly basis.

REFERENCES

Allsopp, C. and M.J. Artis (2003), 'The assessment: EMU, four years on', *Oxford Review of Economic Policy*, **19**, 1–29.

Allsopp, C. (2006), 'Macroeconomic policy in and out of EMU: a view from outside', in A. Watt and R. Janssen (eds), *Delivering the Lisbon Goals: The Role of Macro Economic Policy*, Brussels: ETUI.

Arestis, P. and M. Sawyer (2003), 'Aggregate demand, conflict and capacity in the inflationary process', *Levy Economics Institute Working Paper*, No. 391.

Arestis, P. and M. Sawyer (2003a), 'The case for fiscal policy', *Levy Economics Institute Working Paper*, No. 382.

Calmfors, L. and J. Driffill (1988), 'Centralisation of wage bargaining', *Economic Policy*, **6**, 13–61.

DIW (2005), 'Auswirkungen von länderspezifischen Differenzen in der Lohn-, Preisniveau und Produktivitätsentwicklung auf Wachstum und Beschäftigung in den Ländern des Euroraums', report for German Minister of Economy and Labour.

ECB (2002), 'Monetary transmission in the euro area: where do we stand?', *ECB Working Paper no. 114*.

ECB (2005), *Monthly Bulletin*, November, 70.

Eichner, A.S. (ed.) (1979), *A Guide to Post-Keynesian Economics*, London: Macmillan.

European Trade Union Institute (1979), *Keynes Plus – A Participatory Economy*, Brussels: ETUI.

Fajertag, G. and P. Pochet (2000), *Social Pacts in Europe – New Dynamics*, Brussels: ETUI.

Galbraith, J.K. (1997), 'Time to ditch the NAIRU', *Journal of Economic Perspectives*, **11**(1), 93–108.

Hancke, B. and D. Soskice (2003), 'Wage-setting and inflation targets in EMU', *Oxford Review of Economic Policy*, **19**(1), 149–60.

Hein, Eckhard (2004), 'Die NAIRU – eine post-keynesianische Interpretation', *Intervention. Zeitschrift für Ökonomie*, **1**(1), 43–66.

Hein, Eckhard, T. Niechoj, T. Schulten and A. Truger (eds) (2005), *Macroeconomic Policy Coordination in Europe and the Role of the Trade Unions*, Brussels: ETUI and WSI.

Howell David (ed.) (2004), *Fighting Unemployment: The Limits of Free Market Orthodoxy*, Oxford: Oxford University Press.

IMF (2003), 'Unemployment and labor market institutions: why reforms pay off?', Chapter IV, *World Economic Outlook*, April.

Janssen, R. and E. Mermet (2003), 'Wage policy under EMU', *Transfer*, **9**(4), Winter, Brussels: ETUI.

Kalecki, M. (1943), 'Political aspects of full employment', *Political Quarterly*, **14**(4).

King, J.E. (ed.) (2003), *The Elgar Companion to Post-Keynesian Economics*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Koll, W. (2005), 'Macroeconomic dialogue development and intentions', in E. Hein, T. Niechoj, Niechoj, T. Schulten and A. Truger (eds), *Macroeconomic Policy Coordination in Europe and the Role of the Trade Unions*, Brussels: ETUI and WSI, pp. 175–212.

Layard, R., S. Nickell and R. Jackman (1991), *Unemployment, Macroeconomic Performance and the Labour Market*, Oxford: Oxford University Press.

OECD (1997), 'Economic performance and the structure of collective bargaining', *Economic Outlook*, July.

Robinson, J. (1979), *Contributions to Modern Economics*, Oxford: Basil Blackwell.

Sawyer, M. (2001), 'The NAIRU: a critical appraisal', in P. Arestis and M. Sawyer (eds), *Money, Finance and Capitalist Development*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, pp. 220–54.

Sawyer, M. (2005), 'Towards a simple macroeconomic model incorporating the key heterodox positions', paper presented at the conference Macroeconomics and Macroeconomic Policies – Alternatives to the Orthodoxy, 28–29 October, Berlin.

Schulten, Thorsten (2004), *Solidarische Lohnpolitik in Europa. Zur politischen Ökonomie der Gewerkschaften*, Hamburg: VSA Verlag.

Smith, J.G. (2001), *There is a Better Way. A New Economic Agenda*, London: Anthem Press.

Stockhammer, E. (2004), *The Rise of Unemployment in Europe: A Keynesian Approach*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Traxler, F., S. Blaschke and B. Kittel (2001), *National Labour Relations in Internationalised Markets. A Comparative Study of Institutions, Change and Performance*, Oxford: Oxford University Press.

Watt, A. (2005), 'Can reform of the macroeconomic dialogue improve macroeconomic policymaking in Europe?', in E. Hein, T. Niechoj, T. Schulten and A.

Truger (eds), *Macroeconomic Policy Coordination in Europe and the Role of the Trade Unions*, Brussels: ETUI and WSI pp. 237–59.

Watt, A. (2006), 'The coordination of economic policy in EMU. What contribution can Macroeconomic Dialogue make to higher growth and employment?', in A. Watt and R. Janssen (eds), *Delivering the Lisbon Goals: The Role of Macro Economic Policy*, Brussels: ETUI.

Watt, A. and V. Hallwirth (2003), 'The policy mix and policy coordination in EMU – how can it contribute to higher growth and employment?', *Transfer*, **9** (4), 610–32.

Watt, A. and R. Janssen (2005), 'The high growth and innovation agenda of Lisbon: the role of aggregate demand policies', *European Economic and Employment Policy Brief, 1/2005*, Brussels: ETUI, http://etui.etuc.org/etui/publications/EEEPB/2005/EEEPB_1-05.pdf.

13. Economic growth and beta-convergence in the East European Transition Economies

Nigel F.B. Allington and John S.L. McCombie

INTRODUCTION

The European Union (EU15)[1] admitted ten new members in May 2004. These included eight former communist states, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Slovenia, and Slovakia (the eight Transition Economies – TE8), together with Cyprus and Malta (EU25). Bulgaria and Romania are to join in 2007 or more likely in 2008, following ratification of their accession treaties by the EU15. The new members have no euro derogation as no opt-out is available to them (unlike the UK and Denmark) and several have signalled that they would like to join the Euro Area (EU12)[2] as soon as possible. This would take a minimum of two years if they joined the Exchange Rate Mechanism II (ERMII) at their accession and subsequently met the other Maastricht nominal convergence criteria over that period. Membership of ERMII is regarded as important for anchoring their exchange rates and expectations about inflation as well as imposing fiscal discipline. Estonia, Lithuania, and Slovenia joined ERMII in June 2004 and Slovenia is expected to become the first to join the euro on 1 January 2007.

The purpose of this chapter is to examine the question as to whether the Transition Economies have exhibited any recent evidence of catching up with the EU15 countries in terms of productivity over the period 1994 to 2002. This is accomplished by estimating a number of specifications of the neoclassical beta-convergence growth model. An alternative measure of convergence, sigma-convergence, which we do not report here, measures whether or not the cross-country variation of group per capita income shrinks over time. Finding β-convergence is a necessary, although not a sufficient, condition for σ-convergence to occur.

We commence with what has been termed the classical β-convergence model, which uses growth rates calculated over the full period. To increase the degrees of freedom, we extend this analysis by using quarterly data and panel estimation methods. Finally, we test whether or not the panel data exhibit unit roots. The implication of the latter is that if we cannot reject the hypothesis of a unit root in the data, then there is no statistically significant convergence.

ECONOMIC GROWTH AND CONVERGENCE

The seminal contributions to neoclassical growth theory are those of Solow (1956) and Swan (1956). These, and derivative models, hypothesise that once corrections have been made for differences in steady-state productivity levels across countries, countries with lower initial output per worker will grow at a faster rate. This is referred to as the convergence property of the model or conditional β-convergence. The conditioning variables include $\ln(I/Y)$ and $\ln(n+g+\delta)$ where I/Y is the average gross investment to output ratio, n is population (employment) growth, g is the rate of exogenous technical change, and δ is the rate of depreciation.

Absolute convergence is estimated when there are no conditioning variables. In what is now regarded as a classic empirical investigation utilising cross-sectional relationships, Barro and Sala-i-Martin (1999) estimated absolute β-convergence to be 2 to 3 per cent per annum across US states, Japanese prefectures and 90 European regions. These rates of β-convergence imply that it would take between 25 and 35 years for half the difference between the productivity of poor and rich regions to be eliminated, the so-called half-life of convergence.[3] The impetus for these types of model came from Baumol (1986), but see DeLong's (1988) comment.

Mankiw, Romer, and Weil (1992) estimated conditional convergence, but used the Summers and Heston data set (1988) and added human capital to the capital component of the production function. They found lower rates of β-convergence and thus longer half-lives of 35 to 50 years. Mankiw *et al.* had assumed that the conditioning variables for the steady state were uncorrelated with the country-specific technology term. The country-specific technology term 'reflects not just technology but resource endowments, climate, institutions, and so on; it may therefore differ across countries' (p. 6). However, Islam (1995, p. 1134) conjectured that a panel data framework 'provides a better and more natural setting to control for this technology shift term'. He reformulated the model and estimated rates of convergence using a dynamic panel data model.

The resulting estimated rates of conditional β-convergence were significantly faster (approximately 4.5 to 9 per cent per annum, depending on the sample, giving half-lives of 15 to 8 years respectively) and the implied share of capital fell towards more realistic values without the inclusion of human capital. Islam suggested that his correction for the omitted variable bias, present in the cross-section regressions, highlighted the differences in the aggregate production functions across a sample of nations.

This superior panel approach to growth empirics adopted by Islam has been criticised by Lee, Pesaran and Smith (1998). They conceded that accommodating level effects for individual countries through heterogeneous intercepts (that is, 'fixed effects') would correct for the bias in conventional cross-section estimates. Nevertheless, they had earlier (Lee, Pesaran and Smith, 1997) proposed an alternative approach 'to allow for heterogeneity in level effects, growth effects, and speeds of convergence' (Lee, Pesaran and Smith, 1998, p. 319). Heterogeneity in growth rates makes the issue of β-convergence meaningless in an economic context: the speed of a country's convergence to its own steady state does not offer an explanation of cross-country variances in output per capita over time. Long-run cross-country productivity distributions will depend on the cross-country distribution of exogenous growth rates in technology. They claimed that the convergence literature has 'misdirected attention from the more fundamental issue of the determination and diffusion of technological growth' (1998, p. 323).

In reply, Islam (1998) argued that his earlier work addressed the untenable and unappealing assumption 'that all countries have identical production functions and differ only in the value of the variables of this function and not in its parameters' (p. 325). The introduction of heterogeneity in the intercepts was supported by this interpretation – it allowed the steady-state level of productivity to vary across countries, but assumed countries had the same steady-state growth rate. He concluded that attempts to allow further parametric heterogeneity in the growth model are the result of the tensions between the *within* and *across* country dimensions of convergence. The assumption of homogeneity in growth paths in the steady state was essential for *across* country convergence to have any meaning. These developments are discussed in greater detail below and are taken into account in the empirical work.

THE TRANSITION ECONOMIES AND BETA-CONVERGENCE

The unexpected depth and length of the depression in the eight Transition Economies immediately after their independence in 1989 had a critical impact on their growth performance relative to the OECD. Furthermore, the speed and method of imposing public sector reforms (privatisation, essentially) has shown considerable cross-country variation with unclear implications for economic performance (Svejnar, 2002). The fact that many of these economies now wish to join the euro as soon as possible has heightened the importance of the convergence process. However, there are difficulties involved in modeling using accession country data; specifically, there are relatively few time-series observations and significant structural changes have taken place.

Kočenda (2001), for example, analysed the convergence of eleven Transition Economies on the assumption that the institutional shift for accession to the EU has conditioned necessary adjustments in macroeconomic fundamentals. The basis for this proposition comes from studies of convergence among participants in ERMII. Sarno (1997), using cointegration econometric techniques, found robust evidence for long-run convergence in both nominal and real exchange rates among ERMII members compared with non-ERMII countries. He concluded that the ERMII had been effective in reducing the tendency towards exchange rate misalignment among its members. Kočenda and Papell (1997) found convergent inflation rates, using monthly data from 1959 to 1994 for the entire EU, with the remaining OECD as a control group. They found that while all EU economies experienced faster converging inflation rates than the non-EU benchmark group, participation in the narrow ERMII bands substantially accelerated this process for Germany and the Netherlands.

Kočenda (2001) used monthly data from January 1991 to December 1998 for real industrial output, monetary aggregate (M1), producer and consumer prices, and nominal and real interest rate spreads. His panel unit-root methodology found the greatest degree of convergence in growth rates of real output for the five original members of the Central European Free Trade Agreement (CEFTA)[4] and five economies[5] identified as the most suitable candidates to join the EU in 1998. This is unsurprising given that four of the economies (Czech Republic, Hungary, Poland and Slovenia) are included in both groups.

Kutan and Yigit (2004) extended the work of Kočenda (2001) by 'considering a more stable, post-1993 period and by adopting a more recent panel estimation technique . . . allowing heterogeneity in convergence rates' (p. 23). Using the same macroeconomic variables, but for the period January 1993 to December 2000, they confirmed convergence, albeit weaker than Kočenda estimated. In Kutan and Yigit (2005), Germany became the benchmark against which convergence in industrial production between the original EU members and the Transition Economies TE10[6] could be assessed. Analysing monthly data from January 1993 to December 2004, the TE10 had made significant progress in real convergence with Germany.

Brüggemann and Trenkler (2005) tested for convergence using the Bernard and Durlauf (1995 and 1996) method of cointegration and recursive cointegration (to allow for a catching-up process). Monthly data for industrial production for the Czech Republic, Hungary, and Poland were used to make comparisons with Germany. They found industrial production had not yet converged to German levels and the recursive cointegration technique revealed no evidence for improved convergence over the period 1990–2002.

Sarajevs (2001) employed both classical cross-section and dynamic panel data methods to study real income convergence between the EU15 and TE11[7] (including the TE8 that are now members of the EU). His study used annual data for the period 1991 to 1999. He found some evidence for σ-convergence (distribution of income) and conditional β-convergence in his sample that was robust for the majority of estimation methods and this offered positive support for the enlargement process. Estimated half-lives of β-convergence, however, were very long depending on the estimation method used, and partly explained by the inadequacies inherent in the data and the short data set. When he considered groups of countries, the Baltic group showed σ-convergence whereas, rather ominously, the EU15 exhibited little or no β-convergence, most noticeably after 1995. Sarajevs concluded that overall lack of firm evidence for real income per capita convergence between the Transition Economies and the rest of the EU made it advisable for the former to pursue 'growth-enhancing policies rather than concentrate their efforts on nominal convergence with Maastricht criteria' (p. 23).

More recently, growth theory has begun to take into account the interaction between economies through trade or technology transfer. The impact this has on distribution dynamics in per capita income across a broad spectrum of countries has led to a search for clusters of economies or convergence clubs. Quah (1999), for example, develops a model where 'idea-sharing coalitions' are an important aspect in the evolution of income distributions. Su (2003) utilised a convergence-clustering algorithm on the time series of real income per capita (spanning 1900–1987 and 1885–1994) concluding that four to five convergence clubs, each consisting of two to four members could be detected in a sample of 15 OECD countries. More relevant to this study, Boreiko (2003) assessed the readiness of the accession countries for Economic and Monetary Union (EMU) using a 'fuzzy cluster' analysis on variables determined by the (nominal) Maastricht criteria and Optimum Currency Area theory. Estonia and Slovenia emerged as the leaders in terms of both nominal and real convergence whereas the Czech Republic, Hungary and Poland achieved substantial convergence only in arguably the important real variables.

THE CLASSICAL CONVERGENCE MODEL

The conventional concept of β-convergence can be derived from the neoclassical model of economic growth, utilizing a Cobb–Douglas production function (see, for example, Sala-i-Martin, 1996):[8]

$$Y_{i,t} = A_{i,t} K_{i,t}^{\alpha} L_{i,t}^{(1-\alpha)} \tag{1}$$

where $Y_{i.t}$ is country *i*'s output at time *t*, $K_{i,t}$ and $L_{i,t}$ are the stock of capital and labour in that country respectively, and $A_{i,t}$ is the level of technology. Diminishing marginal product of the factor inputs, constant returns to scale, and perfectly competitive markets (implying that factors are paid their marginal products) are assumed. Technical progress occurs at an exogenously determined rate that is the same for all countries. In a cross-sectional approach, technology is assumed to be constant across countries. However as this is unrealistic in the context of the Transition Economies, as we noted above, a panel data approach with 'fixed-effects' can correct for some of the bias arising from this assumption. Consequently, some estimates of the rate of convergence using panel estimation techniques are presented, as well as those from estimating the original, or classical, convergence model.

In the conventional model, absolute β-convergence occurs if poor economies tend to grow faster than rich ones, or, rather, countries with a lower capital–labour ratio grow faster than those with a higher ratio. It is assumed that all economies converge towards the same steady-state level of GDP per capita. To test empirically absolute β-convergence, either of the following two equations is estimated:

$$\Delta \ln y_{i,t+T} = c + b \ln y_{i,t} + \varepsilon_{i,t+T} \tag{2}$$

or

$$\ln y_{i,t+T} = c + (1+b) \ln y_{i,t} + \varepsilon_{i,t+T} \tag{3}$$

where y is the level of productivity, $\Delta \ln y$ is the exponential growth rate over the period *T,* which is the length of period under consideration, for example, so many years, quarters or months, and ε is the error term. If $b <$ 0 and $(1 + b) < 1$, there is absolute β-convergence.

The neoclassical growth model does not necessarily imply that different economies converge to the same steady-state level of income per capita. Therefore, to investigate the concept of conditional β-convergence, a vector of conditioning variables, as noted above, must be included in equations (2) and (3) to control for differences in the steady-state level of productivity of country *i*.

Equations (2) and (3) may be estimated using cross-sectional data with the growth rate of productivity measured over a period of several years – 25 in the case of Mankiw *et al.* (1992) and pooled 5-year periods in the case of Islam (1995). Alternatively, it may be estimated using time-series data (annual, quarterly or monthly) by, for example, panel estimation techniques.

The Speed of Convergence and the Half-Life

In order to give a more intuitive interpretation of the regression coefficients, it has become common practice to calculate the speed of convergence and the half-life. The latter is the time taken to reduce the productivity gap by one half. Clearly, if there is no convergence, then the half-life is undefined. If we consider equation (3), it may be straightforwardly shown that $(1+b) = e^{-\lambda' T}$, where λ' is the speed of convergence, measured in units of the frequency of the data. It follows that λ' equals $-\ln(1+b)/T$. Consequently, when our cross-sectional data are used, the period T is 7.75 years (the start of the data set is 1994Q1 and the terminal date is 2002Q4). When panel data are used with quarterly data, T is 36 quarters and the speed of convergence is measured in per cent per quarter. For convenience, we always report the speed of convergence in terms of years and this is denoted by λ. Hence, $\lambda = \lambda' f$, where f is the frequency of the data: $f = 1$ for yearly data; 4 for quarterly data; and 12 for monthly data.

To calculate the half-life, we start with the equation for the rate of increase of productivity over a given period T, which is given by:

$$\ln y_{i,t+T} - \ln y_{i,t} = \lambda' T \tag{4}$$

The level of initial productivity is assumed to be half the terminal level, i.e., $y_{i,t} = 0.5 y_{i,\, t+T}$. Consequently, substituting this equation into equation (4) we obtain

$$\ln y_{i,t+T} - (\ln 0.5 + \ln y_{i,t+T}) = \lambda' H \tag{5}$$

where H equals the half-life measured in terms of the frequency of the data. It follows that $H = (-\ln 0.5)/\lambda' = (\ln 2)/\lambda'$. Consequently, the half-life in years is given by $H^* = (\ln 2)/\lambda' f = \ln(2)/\lambda$. Again, we report the annualized half-lives in all the results.

Data

The data cover fourteen of the EU15 (Luxembourg is excluded for reasons of data availability) and the eight Transition Economies (TE8) that joined the EU in May 2004. Country abbreviations and the definitions of various groups of economies are given in Table 13.1.

Quarterly seasonally-adjusted data for real per capita GDP covering the period 1994Q1 to 2002Q4 was obtained from Datastream and the central bank of the relevant country. Quarterly data were used to resolve econometric problems found in the literature from insufficient observations, as this gives 36 time-series observations for each of the 22 countries. These problems arise because the reliability of economic data between independence in 1989 and 1994 is uncertain and annual data (which many studies use) may lead to short sample period problems. The downside of this strategy is that quarterly data may be dominated by noise, and so the results must be treated with caution. However, some studies use even higher frequency data, for example, Kočenda (2001) and Brüggemann and Trenkler (2005) use monthly data.

Table 13.1 Country abbreviations and country groups

Code	Country	Code	Country
AUS	Austria	**ITA**	Italy
BEL	Belgium	**LAT**	Latvia
CZR	Czech Republic	**LIT**	Lithuania
DEN	Denmark	**NED**	Netherlands
EST	Estonia	**POL**	Poland
FIN	Finland	**PGL**	Portugal
FRA	France	**SPA**	Spain
GER	Germany	**SVK**	Slovak Republic
GRE	Greece	**SVN**	Slovenia
HUN	Hungary	**SWE**	Sweden
IRE	Ireland	**UK**	United Kingdom
EU14	AUS, BEL, DEN, FIN, FRA, GER, GRE, IRE, ITA, NED, PGL, SPA, SWE, UK	**EURO11**	AUS, BEL, FIN, FRA, GER, GRE, IRE, ITA, NED, PGL, SPA
TE8	CZR, EST, HUN, LAT, LIT, POL, SVK, SVN	**EURO5**	AUS, BEL, FRA, GER, NED
TE5	CZR, EST, HUN, POL, SVN	**TE4**	CZR, EST, LAT, SVN

For the short sample period of 7.75 years, many traditional variables shown to be important in long-term analysis of economic growth exhibited insufficient variation across countries and over time to be used here. For example, in the Transition Economies, rates of secondary school enrollment, traditionally used to measure human capital investment, are high and do not show significant variation. This is also true for the West European economies. The problem for the Transition Economies is not a deficiency of human capital as measured by enrollment rates, but rather a failure to develop the skills demanded by a market economy. Unfortunately, the rate of investment in physical capital is not available for the Transition Economies.

Following Sarajevs, therefore, consumer price inflation and the broad-money-to-GDP ratio are used to control for domestic economic developments that impact on the steady state of the economy, reflecting the 'dynamics of macroeconomic stabilisation and development of the financial sector' (Sarajevs, 2001, p. 12). Change in the real effective exchange rate is included as a proxy for external factors, particularly changes in relative international productivity and competitiveness. The data were obtained from Deutsche Bank and the central bank of the relevant country.

We start by estimating the classical convergence model. Table 13.2 presents the results for the classical cross-section approach to absolute and conditional β-convergence respectively. Absolute β-convergence is the strong prediction of the Solow–Swan model implying that economies will converge to the same steady state in terms of income per capita. This hypothesis can be detected in a number of statements referring to the long-term prospects for EU enlargement.

Using the full sample of 22 countries there appears to be no evidence of absolute β-convergence. The estimate of *b* is 0.073 which implies divergence, although this value is not significantly different from zero (the probability level is 0.387). One of the problems for the Solow–Swan model is that four of the Transitional Economies (Hungary, Lithuania, Poland, and the Slovak Republic) have negative productivity growth rates over the period. This would imply that they were above their steady-state capital–labour ratios, which is implausible to say the least. Removing these countries, the reduced sample of countries (the EU14 & TE4) gives the expected negative sign and an implied half-life of 239 years, but the coefficient is again not statistically significant. The sample of just the EU14 does not have a statistically significant rate of convergence, either. Including the average, but not the standard deviation, of the conditioning variables produced a small improvement in the results (Table 13.2 (II) and (III)).

Table 13.2 Cross section: absolute and conditional β*-convergence, 1994 to 2002*

(I) Absolute convergence

	b	*t-value*	λ	*t-value*	*Probability*	*Half-life*	$\bar{R}^2$
EU14 & TE8	0.073	0.855	n.a.	n.a.	n.a.	n.a	−0.004
EU14 & TE4	−0.022	−0.331	0.003	0.371	0.387	239	−0.053
EU14	−0.140	−1.423	0.019	1.319	0.212	36	−1.423

(II) Conditional convergence: Means of the conditioning variables

	b	*t-value*	λ	*t-value*	*Probability*	*Half-life*	$\bar{R}^2$
EU14 & TE8	−0.301	−2.086	0.046	1.735	0.101	15	0.292
EU14 & TE4	−0.347	−2.245	−0.055	−1.801	0.095	13	0.249
EU14	−0.314	−1.486	−0.049	−1.224	0.252	35	−0.089

(III) Conditional convergence: Standard deviation of the conditioning variables

	b	*t-value*	λ	*t-value*	*Probability*	*Half-life*	$\bar{R}^2$
EU14 & TE8	−0.059	−0.520	0.008	0.504	0.621	88	0.002
EU14 & TE4	−0.145	−1.382	0.020	0.879	0.396	34	−0.203
EU14	−0.141	−1.465	0.020	1.357	0.208	35	−0.066

Notes: The TE4 countries are CZR, EST, LAT and SVN. The t-values are derived from the White heteroscedasticity – consistent standard errors.

DYNAMIC PANEL DATA ESTIMATES

Bernard and Durlauf (1996) have pointed out that there is a difference in the assumptions underlying the classical cross-section convergence model and the model using time-series data. 'Cross-section tests turn out to place much weaker restrictions on the behavior of growth than the associated time-series tests. As a result, the cross-section tests can reject a no convergence null hypothesis for data generated by economies with different long-run steady states. . . . Time series tests do not spuriously reject the no convergence null for data generated by multiple long-run equilibria . . . However, the time series approach requires that the economies under analysis are near their long-run equilibria. . . . The tests may therefore be invalid if the data are largely driven by transitional dynamics' (p. 171). They argue that an important advance over both would be to integrate the two approaches. This is an advantage of the use of panel data.

In estimating the panel data model, a number of important issues must be addressed. The choice between a fixed effects and random effects specification is determined by Islam's (1995) observation that country-specific effects are correlated with the explanatory variable, making a random effect specification unsuitable. In choosing the appropriate dynamic panel data estimators the choices include: Least Squares with Dummy Variables (LSDV); Instrumental Variable Methods by Anderson–Hsaio using lagged levels or differences of the regressand (AHL, AHD); Two and Three Stage Least Squares (2SLS, 3SLS); Exact Maximum Likelihood Estimator (MLE); One- and Two-step Generalized Method of Moments (GMM1 and GMM2); and Minimum Distance Estimator (MD).

As Sarajevs (2001, p. 17) has pointed out, advanced estimators 'that make use of all the available information (MD, exact MLE, GMM2, 3SLS) are asymptotically the same and dominate simpler and less efficient estimators (such as LSDV, AHL, or AHD, one-step GMM1, 2SLS)'. However, he points out that Monte Carlo studies on the performance of various estimators with small samples and noisy data suggest that simpler estimators such as LSDV or 2SLS are to be preferred. Simple estimators such as LSDV and AHD also perform best in terms of bias and root mean square error. The optimal weighting matrices used in more advanced estimators must be estimated from the data, but this simply adds more noise to the estimator (Islam, 1995). Although the presence of a lagged dependent variable makes LSDV an inconsistent estimator with regard to N (the number of cross-sections), it is consistent in the direction of T (the number of time periods) (Islam, 1995, p. 1138). Islam found that LSDV performed very well in his estimations. Consequently, LSDV is chosen as the preferred estimation technique and these are the results we report in the chapter.

Switching to a dynamic panel data model yields a significant increase in the estimated rates of absolute and conditional β-convergence as can be seen from Table 13.3.

We estimated both absolute and conditional convergence for 11 different subgroups of the EU and Transitional Economies. In all cases there is a statistically significant rate of convergence, but what is surprising is the rapid speed, with half-lives of between just under one to three years, with the majority taking a value of around 2 years. This is perhaps unexpectedly fast, given the degree to which the Transition Economies lag behind the European Union countries in terms of their initial productivity levels. However, our results are not out of line with those of Sarajevs (2001, Table 11). Using yearly data and a shorter period, he found half-lives of 2 to 3 years for absolute convergence for various combinations of EU and Transition Economy countries. For some smaller groupings he found that there was no convergence at all. On the other hand, he found a much larger half-life of 78 years when he tested for no conditional convergence for his full sample, but there was no convergence for his remaining sub-samples.

Table 13.3 Panel data: absolute and conditional β*-convergence rates and half-lives (years), 1994 to 2002*

(I) Absolute convergence

	b	*t-value*	λ	*Half-life*
EU14	−0.067	−3.851	0.278	2.49
EURO11	−0.055	−3.281	0.228	3.04
EURO5	−0.221	−5.106	0.997	0.70
TE8	−0.111	−2.485	0.473	1.47
TE5	−0.097	−2.787	0.407	1.70
EU14 & TE8	−0.094	−3.271	0.397	1.75
EU14 & TE5	−0.087	−5.066	0.365	1.90
EU14 & TE4	−0.063	−2.561	0.260	2.67
EURO11 & TE8	−0.094	−2.943	0.395	1.75
EURO11 & TE5	−0.085	−4.549	0.356	1.95
EURO5 & TE8	−0.116	−2.719	0.494	1.40
EURO5 & TE5	−0.124	−4.237	0.528	1.31

(II) Conditional convergence

	b	*t-value*	λ	*Half-life*
EU14	−0.071	−3.825	0.296	2.34
EURO11	−0.057	−3.261	0.234	2.96
EURO5	−0.218	−5.080	0.984	0.70
TE8	−0.115	−2.840	0.488	1.42
TE5	−0.075	−2.586	0.312	2.22
EU14 & TE8	−0.097	−3.621	0.409	1.67
EU14 & TE5	−0.089	−5.081	0.375	1.85
EU14 & TE4	−0.062	−2.149	0.257	2.70
EURO11 & TE8	−0.096	−3.303	0.404	1.71
EURO11 & TE5	−0.085	−4.555	0.355	1.95
EURO5 & TE8	−0.120	−3.130	0.509	1.36
EURO5 & TE5	−0.116	−4.107	0.492	1.41

Notes: All values of λ are significant at the 1 per cent confidence level. The t-values are derived from the White cross-section standard errors.

In order to check for the stability of the results the conditional equations are re-estimated progressively moving the initial period forward by a year from 1994 to 1997. The last period of 5 years, 1997 to 2005 was considered to be the minimum length of time to give reliable results. Five samples were selected, namely EU14 & TE8; EU14 & TE5; EU14 & TE4; EU5 & TE8; and EU5 & TE5. As may be seen from Table 13.4, the estimates proved to be very stable both over time and across different groups. The speed of

Table 13.4 *Conditional* β-*convergence: various time periods*

	Period	*b*	*t-value*	λ	*Half-life*
EU14 & TE8	1994–2002	−0.097	−3.469	0.409	1.70
	1995–2002	−0.071	−3.227	0.294	2.36
	1996–2002	−0.080	−2.644	0.335	2.07
	1997–2002	−0.112	−3.046	0.476	1.46
EU14 & TE5	1994–2002	−0.089	−3.998	0.375	1.85
	1995–2002	−0.099	−3.599	0.418	1.66
	1996–2002	−0.109	−2.792	0.461	1.50
	1997–2002	−0.149	−3.048	0.644	1.08
EU14 & TE4	1994–2002	−0.062	−2.149	0.257	2.70
	1995–2002	−0.077	−2.561	0.320	2.17
	1996–2002	−0.095	−2.523	0.399	1.74
	1997–2002	−0.137	−3.085	0.589	1.18
EU5 & TE8	1994–2002	−0.120	−3.281	0.509	1.36
	1995–2002	−0.077	−3.319	0.322	2.15
	1996–2002	−0.077	−2.427	0.322	2.16
	1997–2002	−0.105	−2.724	0.446	1.56
EU5 & TE5	1994–2002	−0.116	−4.321	0.492	1.41
	1995–2002	−0.143	−3.704	0.616	1.12
	1996–2002	−0.139	−2.875	0.597	1.16
	1997–2002	−0.168	−2.769	0.597	1.16

Notes: All values of λ are significant at the 1 per cent confidence level. The t-values are derived from the White cross-section standard errors.

convergence was always statistically significant and the half-lives were again small, ranging from just over one to just over two years.

Next, attention is given to the idea of convergence clubs (groups of two to five countries) within the full sample of twenty-two countries. Table 13.5 shows a range of conditional β-convergence test results for clubs within the EU14 and TE8 separately and then across the two groups. The country clubs are arranged so that estimates of half-lives for the clubs are in ascending order.

For the intra-EU14 club the results support the emergence of 'core' clubs. Belgium, France, and Germany form one club and Belgium, Germany and the Netherlands another, with both exhibiting high rates of conditional β-convergence. The Scandinavian economies return the lowest half-life of 0.6 years. At the opposite extreme, the EU 'periphery' nations exhibit relatively poor convergence both amongst themselves (see Greece,

Table 13.5 Panel data: conditional β*-convergence, estimates for clubs, 1994 to 2002*

Country clubs	*b*	*t-value*	λ	*Half-life*
(I) Intra-EU14				
DEN, FIN, SWE	− 0.243	− 4.112	1.112	0.62
BEL, GER, NED	− 0.224	− 3.401	1.016	0.68
BEL, FRA, GER	− 0.215	− 3.416	0.969	0.71
DEN, SWE, UK	− 0.176	− 2.227	0.773	0.90
GRE, ITA	− 0.139	− 1.746	0.599	1.16
BEL, FRA, NED	− 0.134	− 3.225	0.577	1.20
BEL, FRA, UK	− 0.124	− 2.585	0.531	1.31
FRA, PGL, SPA, UK	− 0.088	− 3.034	0.370	1.87
FRA, GER, ITA	− 0.073	− 1.497	0.303	2.29
BEL, GER, NED, UK	− 0.065	− 2.604	0.269	2.58
BEL, FRA, NED, UK	− 0.064	− 2.170	0.264	2.63
FRA, GER, UK	− 0.052	− 1.916	0.215	3.22
IRE, UK	− 0.049	− 1.092	0.200	3.46
PGL, SPA	− 0.040	− 0.909	0.165	4.21
GRE, IRE, PGL, SPA	− 0.026	− 1.359	0.106	6.55
(II) Intra-TE8				
EST, LAT, LIT	− 0.183	− 2.311	0.810	0.86
EST, LAT, LIT, POL	− 0.165	− 2.311	0.723	0.96
CZR, POL, SVK	− 0.150	− 1.572	0.652	1.06
SVK, SVN	− 0.104	− 1.767	0.439	1.58
HUN, SVN	− 0.098	− 2.813	0.414	1.68
EST, SVK, SVN	− 0.083	− 1.910	0.345	2.01
CZR, HUN, POL	− 0.079	− 3.476	0.330	2.10
HUN, SVK, SVN	− 0.065	− 2.323	0.271	2.56
(III) Intra-EU14 & TE8				
AUS, GER, POL	− 0.272	− 5.684	1.270	0.55
DEN, EST, FIN, LAT, LIT, SWE	− 0.185	− 2.976	0.816	0.85
FRA, GER, POL	− 0.158	− 4.147	0.686	1.01
AUS, ITA, SVN	− 0.158	− 2.827	0.686	1.01
EST, GER, HUN, LAT, LIT, POL	− 0.136	− 2.473	0.584	1.19
CZR, FRA, GER, POL	− 0.127	− 4.401	0.545	1.27
GRE, ITA, SVN	− 0.119	− 1.970	0.509	1.36
FRA, GER, ITA, POL, SPA	− 0.108	− 3.864	0.457	1.52
AUS, CZR, GER, HUN, POL	− 0.098	− 4.189	0.415	1.67
AUS, CZR, HUN, POL	− 0.096	− 3.979	0.405	1.71
AUS, HUN, SVN	− 0.095	− 2.818	0.398	1.74
AUS, GER, HUN	− 0.091	− 2.746	0.382	1.82
FRA, GER, ITA, POL, SPA, UK	− 0.083	− 3.800	0.348	1.99
CZR, GER, HUN, POL	− 0.080	− 3.479	0.331	2.09
GER, HUN, SVK, SVN	− 0.067	− 2.334	0.278	2.49

Note: The t-values are derived from the White cross-section standard errors.

Ireland, Portugal, and Spain; or Portugal and Spain) and in clubs with 'core' EU countries (see Germany and Ireland or France, Ireland and Spain as examples). Nevertheless, even the longest half-life is only 6 years.

Within the TE8, the Baltic States have the fastest rate of β-convergence. Hungary, the Slovak Republic and Slovenia have the slowest rate of conditional β-convergence. However, the range of half-lives of between 1 and 2.5 years is extremely rapid and lies well within some of the clubs in the EU14. The most interesting set of results for clubs consists of both Transition Economies and the original members of the EU. The club containing Austria, Germany and Poland shows the fastest rate of convergence for any reported (a half-life of 0.5 years). The Scandinavian/Baltic club also performs well (a half-life of 0.8 years). The club with the five largest members of an enlarged Eurozone in population terms (France, Germany, Italy, Poland and Spain) has a convergence half-life of 1.5 years. The range of half-lives obtained in the intra-EU14 and TE clubs (0.5 years to 2.5 years) falls well within the range of those obtained for intra-EU14 clubs. This suggests that the entry of the Transition Economies will be no more destabilizing to the EU than the presence of the existing 'peripheral' economies, on the basis of existing economic policies. However, any rapid move towards fulfilling the Maastricht criteria on the part of the Transition Economies could change this situation. In particular, there may be the need to curb rapidly rising budget deficits and higher inflationary expectations that follow from the Balassa–Samuelson condition.

A Caveat

The results show perhaps implausibly high rates of convergence and remarkably short half-lives. However, the exact results are extremely sensitive to small changes in the estimated regression coefficients and hence may be greatly influenced by a relatively small bias on these estimates. Table 13.6 reports the associated speed of convergence and half-life for given estimates of $(1+b)$ when the frequency of the data is quarterly and yearly, respectively. These hypothetical results, of course, hold for any data set. For quarterly data, it can be seen that if the estimate of $(1+b)$ falls from 0.999 to 0.900, then the half-life falls dramatically from 173 to just over 1.5 years. The use of yearly data is nearly as sensitive; the half-life falls from 692 to just over 6 years. Thus, the value of the half-life is extremely sensitive to even small biases on the coefficients, and given the relatively large standard errors associated with λ, it is difficult to have much confidence in the precise estimates of the half-life. Moreover, in the case of absolute convergence, if $(1+b)$ is unity, there is a unit root and by definition no convergence. However, it is well established that when testing for this hypothesis then the

Table 13.6 Estimates of $(1+b)$, the speed of convergence (λ) and the half-life

	Quarterly data		Yearly data	
$(1+b)$	λ*	*Half-life**	λ	*Half-life*
0.999	0.004	173.20	0.001	692.80
0.990	0.040	17.24	0.010	68.97
0.950	0.205	3.38	0.051	13.51
0.900	0.421	1.64	0.105	6.58
0.850	0.650	1.07	0.162	4.27
0.800	0.893	0.78	0.223	3.11
0.750	1.151	0.60	0.288	2.41

Note: * λ and the half-life figures using quarterly data are annualized.

t-statistic does not have the approximate normal standard distribution even for large samples and so it is necessary to use the Dickey–Fuller test for a unit root. This is discussed in the next section.

UNIT ROOTS AND CONVERGENCE

The use of panel data for the estimation of whether or not there is convergence, and, if there is, for the determination of the speed of convergence, has obvious similarities to the testing for unit roots in panel data. As we noted above, the absolute convergence approach, given by equation (3) but using time-series data, assumes that all countries experience the same rate of convergence (or divergence) and have the same steady-state level of productivity. The coefficient on the initial level of productivity is common to all the countries. This is analogous to testing whether or not there is a common unit-root process.

The power of a unit-root test for a relatively small panel such as the one that we used here is considerably higher than for the individual country time-series. A failure to reject the null hypothesis of a unit root is equivalent to failing to reject the hypothesis that there is no absolute convergence. The null hypothesis is tested using the procedures of Levin, Lin and Chu (2002) and Breitung (2000), where full descriptions of the methods adopted can be found. Both are based on the augmented Dickey-Fuller equation, which in this context is given by:

$$\Delta \ln y_{i,t} = \phi \ln y_{i,t-1} + \sum_{j=1}^{k} \rho_{i,j} \Delta \ln y_{i,t-j} + \mu_{i,t} \qquad (6)$$

where the variables are as before and μ is the error term. The null hypothesis to be tested is that $\phi = 0$, or that there is a unit root. (ϕ is analogous to the coefficient *b* in the panel data estimates.) If this is not rejected, then, as we have noted, there is no statistically significant convergence. If, on the other hand, the null hypothesis is rejected, so that there is no unit root, then, if $\phi < 1$, there is convergence and we can compute the rate of convergence and the half-life as before. It should be noted that as with the earlier panel data estimations only the convergence rate of the whole group could be inferred. The panel data specifications that were estimated included fixed effects, so that the procedure corresponds closely to the testing for absolute convergence carried out above, but with the addition of the lagged values of $\Delta \ln y_i$ to remove any serial correlation.

The results are reported in Table 13.7, equations (i) and (ii). The optimal number of lags, *k*, was determined by the Schwarz Information Criterion. It can immediately be seen that the results are contradictory, with the Levin, Lin and Chu (LLC) unit root statistic always rejecting the null hypothesis of a unit root, whereas the Breitung statistic does not reject it for any grouping of countries. The average half-life given by the LLC estimation is around 3 years. This means that it would take around a decade for convergence to around 90 per cent of the EU average to occur, which is not implausible. However, the Breitung procedure not only fails to reject the null hypothesis, but it gives an estimated coefficient of ϕ that is actually positive.

It is interesting to consider the case of the EU14 & TE5. This, it will be recalled, was the full sample of countries less the Transition Economies with negative growth rates. The classical convergence estimates gave statistically significant estimates of convergence but with a long half-life, over two centuries. On the other hand, the LLC estimates suggest a statistically significant rate but with a half-life of 4 years.

Hadri's (2000) test statistic was estimated and the results are reported in Table 13.7, equation (iii). This test is based on a common unit-root process, but unlike the other tests, takes the null hypothesis of no unit root. In all cases, the heteroscedastic-consistent Z-statistic rejected this hypothesis, which again suggests that there is no convergence.

Im, Pesaran, and Shin (IPS) (2003) provide a test for unit roots in heterogeneous panels (Table 13.7, equation (iv)). In other words, in equation (6) the test allows the coefficient ϕ to vary according to the country concerned. The procedure essentially tests for a unit root on each cross-section *separately* and then calculates what IPS term the *t-bar* test, which is based on the average of the individual, augmented Dickey–Fuller tests. Although this test has become popular, IPS issue a word of warning. Rejection of the null hypothesis does not necessarily imply a rejection of a unit root for all

Table 13.7 Panel data unit root tests, 1994 to 2002

Group		*Unit root statistic*	*Probability*	*φ (or b)*	*λ*	*Half-life*	*Convergence implied?**
EU14 & TE8	(i)	−2.473	0.007	−0.050	0.205	3.38	Yes
	(ii)	1.291	0.902	0.006	−0.025	..	No
	(iii)	13.802	0.000	..	..	..	No
	(iv)	0.258	0.602	..	..	..	No
EU14 & TE5	(i)	−2.469	0.007	−0.051	0.208	3.33	Yes
	(ii)	0.927	0.823	0.005	−0.019	..	No
	(iii)	13.211	0.000	..	..	..	No
	(iv)	0.362	0.641	..	..	..	No
EU14 & TE4	(i)	−1.762	0.039	−0.040	0.165	4.19	Yes
	(ii)	0.846	0.801	0.004	0.018	..	No
	(iii)	13.852	0.000	..	..	..	No
	(iv)	1.374	0.915	..	..	..	No
EURO11 & TE8	(i)	−2.040	0.021	−0.047	0.195	3.56	Yes
	(ii)	1.038	0.850	0.005	−0.021	..	No
	(iii)	12.245	0.000	..	..	..	No
	(iv)	0.640	0.739	..	..	..	No
EURO11 & TE5	(i)	−2.014	0.022	−0.048	0.196	3.54	Yes
	(ii)	0.639	0.739	0.004	−0.014	..	No
	(iii)	11.554	0.000	..	..	..	No
	(iv)	0.786	0.784	..	..	..	No
EU5 & TE8	(i)	−2.402	0.008	−0.069	0.286	2.42	Yes
	(ii)	0.706	0.760	0.005	−0.020	..	No
	(iii)	9.944	0.000	..	..	..	No
	(iv)	−0.647	0.259	..	..	..	No
EU5 & TE5	(i)	−1.382	0.084	−0.066	0.272	2.54	Yes
	(ii)	0.857	0.804	0.007	−0.026	..	No
	(iii)	7.971	0.000	..	..	..	No
	(iv)	−0.121	0.452	..	..	..	No
EU14	(i)	−1.954	0.025	−0.045	0.184	3.76	Yes
	(ii)	1.133	0.871	0.007	−0.027	..	No
	(iii)	13.098	0.000	..	..	..	No
	(iv)	0.897	0.815	..	..	..	No
EURO11	(i)	−1.315	0.094	−0.036	0.146	4.76	Yes
	(ii)	0.269	0.606	0.002	−0.001	..	No
	(iii)	11.364	0.000	..	..	..	No
	(iv)	1.670	0.953	..	..	..	No
EU5	(i)	−0.131	0.448	−0.042	0.170	4.07	Yes
	(ii)	0.381	0.649	0.004	−0.015	..	No
	(iii)	7.459	0.000	..	..	..	No
	(iv)	0.921	0.822	..	..	..	No

Table 13.7 (continued)

Group		*Unit root statistic*	*Probability*	*ϕ (or b)*	*λ*	*Half-life*	*Convergence implied?**
TE8	(i)	− 1.729	0.042	− 0.074	0.309	2.24	Yes
	(ii)	1.486	0.931	0.014	− 0.055	..	No
	(iii)	5.545	0.000	..	..	..	No
	(iv)	− 1.052	0.146	..	..	..	No
TE5	(i)	− 1.997	0.023	− 0.100	0.420	1.65	Yes
	(ii)	0.884	0.812	0.011	− 0.042	..	No
	(iii)	3.814	0.000	..	..	..	No
	(iv)	− 1.153	0.125	..	..	..	No

Notes:
Equations:
(i) Levin, Lin and Chu estimating procedure.
(ii) Breitung estimating procedure.
(iii) Hadri estimating procedure.
(iv) Im, Pesaran and Shin estimating procedure.
* Convergence is implied if the regression coefficient takes a negative sign and is statistically significant at the 10 per cent confidence level or better.

the cross sections, but only for a certain proportion of them. The test does not provide any evidence as to the size of this proportion as the number of cross-sections tends to infinity, nor does it provide any information about the individual countries for which the null hypothesis is rejected. Hence, we cannot calculate the implied convergence speeds and half-lives for the individual countries. It can be seen that the IPS test does not reject the null hypothesis for any of the various groupings.

However, there are theoretical considerations as to whether or not allowing ϕ to vary is an appropriate test of the convergence model. As we have seen, the Solow model assumes that ϕ and the implied speed of convergence, λ, are constant across countries in the sample. In fact, λ is not constant to the extent that it is a function of the growth of employment (often proxied by population) that will differ between countries. But the convergence literature generally ignores this point.[9] Of perhaps greater importance is that allowing ϕ to vary implicitly allows the exogenous rate of technical change – or the steady-state rate of productivity growth – to vary between countries. (The fixed effects effectively allow the level of total factor productivity – or approximately per capita income – to differ between countries.) However, Islam (1998, p. 326) recalls that in Islam (1995) he noted that ' "By being more successful (through the panel framework) in controlling for further sources of difference in the steady state levels of income, we have at the same time, made the obtained convergence *hollower*" [p. 1162]. When in addition, heterogeneity in growth rates is allowed, convergence becomes, in essence, an empty construct. The higher rates of convergence obtained from such an exercise have little potency, if any.'

There are two points to be made concerning this argument. The first is that the hypothesis that each country has the same ϕ and the same intercept should be explicitly tested; a rejection of this hypothesis does not make convergence 'hollower'; it merely rejects the central hypothesis of the Solow model that all countries share the same production function and, as a consequence, the same technology and rate of technical progress.

Secondly, Felipe and McCombie (2005) have shown that once an allowance is made for variations in these two factors, the statistical fit is bound to improve. Essentially all that is being estimated is the national income identity, given the stylized facts that the growth of capital is approximately equal to the growth of output (the capital–output ratio is constant) and factor shares are roughly constant.

CONCLUSIONS

In this chapter, we have estimated the rate of beta-convergence for fourteen European Union countries and eight Transition Economies using both the classical and panel data estimation techniques. The original rationales for the specified equations are the models of Solow (1956) and Swan (1956), which predict convergence. Failure to find convergence implies a rejection of the Solow–Swan models and is compatible with the *AK* endogenous growth model. However, the results were rather inconclusive. There was no evidence of absolute convergence when the classical convergence model was estimated, although when four of the Transition Economies with negative growth rates were dropped from the sample, there was significant convergence, but with a half-life of over 200 years. When the means of the conditioning variables were introduced, significant convergence for the full sample could be found, with an implied half-life of 15 years. Nevertheless, while there was convergence for the EU14 countries, the result was not statistically significant.

The panel data estimates found significant absolute and conditional convergence for the full sample and various clubs of countries. However, the half-lives were short, being generally between one and three years. However, if there was no convergence there would be a unit root, and in this case the conventional statistics are no longer valid. Consequently, we tested the panel data for a unit root and at the same time estimated the speed of convergence and the half-life (where appropriate). The results were ambiguous. The Levin, Lin and Chu estimating procedure suggested significant absolute β-convergence for nearly all the sub-samples with half-lives of between approximately two and four years. The Breitung procedure, however, suggested no convergence, and neither did the Hadri statistic, which has the null of no unit root. All these tests are based on the assumption of a common unit root. The Im, Pesaran, and Shin technique allows ϕ in equation (7) to vary separately with the country. The results invariably suggested that on average it was not possible to reject the null hypothesis of a unit root, but this does not mean that this is true for all countries.

However, we must question the plausibility of the Solow model that underlies the interpretation of these specifications. Are the assumptions that all countries are on their production frontier (technically and allocatively efficient), have access to the same level of technology and have a common rate of (exogenous) technical change credible? It is very debatable that this is true for the Transition Economies. The slower, indeed negative, growth of some of the Transitional Economies may well be due to the dislocation and structural change caused by the movement to a free market economy. This may well affect each Transitional Economy to a different extent. Part of the differences in technical change may be due to a technological catch-up phenomenon. The effect of this may differ between countries depending upon their social capability. Thus, it may well be that such cross-country convergence analysis can tell us relatively little about 'why growth rates differ' and what is needed is a greater understanding of the disparate factors affecting the growth rates of the Transitional Economies at the individual country level.

NOTES

1. The EU15 countries are Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain, Sweden, and the United Kingdom.
2. The EU15 less Denmark, Sweden, and the UK.
3. As we are dealing with a production function, the data should refer to productivity, or output (GDP) per worker. In practice, data limitations mean that per capita GDP is used as a proxy for productivity.
4. Czech Republic, Hungary, Poland, Slovakia, and Slovenia.
5. Czech Republic, Estonia, Hungary, Poland and Slovenia.
6. TE8 plus Bulgaria and Romania. The combinations of Transition Economies are listed in Table 13.1.
7. TE10 plus Albania.
8. The existence of the aggregate production function is thus taken as a maintained hypothesis and aggregation problems, which are severe, are ignored.
9. Under the assumptions of the Solow model, $\lambda = (1-\alpha)(n+g+\delta)$ where, it will be recalled, n is population (employment) growth, g is the exogenous rate of technical change and δ is the rate of depreciation. The convergence model can be easily reparameterized to take account of this problem.

REFERENCES

Barro, R.J. and X. Sala-i-Martin (1999), *Economic Growth*, Cambridge, MA: MIT Press.

Baumol, W.J. (1986), 'Productivity Growth, Convergence and Welfare: What the Long-Run Data Show', *American Economic Review*, **76**, pp. 1072–85.

Bernard, A.B. and S.N. Durlauf (1995), 'Convergence in International Output', *Journal of Applied Econometrics*, **10**, pp. 97–108.

Bernard, A.B. and S.N. Durlauf (1996), 'Interpreting Tests of the Convergence Hypothesis', *Journal of Econometrics*, **71**, pp. 161–73.

Breitung, J. (2000), 'The Local Power of Some Unit Root Tests for Panel Data', in B.H. Baltagi (ed.), *Nonstationary Panels, Panel Cointegration, and Dynamic Panels, Advances in Econometrics*, **15**, Amsterdam: Elsevier Science, pp. 161–77.

Boreiko, D. (2003), 'EMU and Accession Countries: Fuzzy Cluster Analysis of Membership', *International Journal of Finance and Economics*, **8**, pp. 309–25.

Brüggemann, R. and C. Trenkler (2005), 'Are Eastern European Countries Catching Up? Time Series Evidence for Czech Republic, Hungary, and Poland', SFB 649 Discussion Paper **2005-014**, Humbolt University, Berlin, Germany.

DeLong, J.B. (1988), 'Productivity Growth, Convergence, and Welfare: Comment', *American Economic Review*, **78**, pp. 1138–54.

Dickey, D. and W.A. Fuller (1979), 'Distribution of the Estimators for Time Series Regressions with a Unit Root', *Journal of the American Statistical Association*, **74**, pp. 427–31.

Felipe, J. and J.S.L. McCombie (2005), 'Why are Some Countries Richer than Others? A Skeptical View of Mankiw-Romer-Weil's Test of the Solow Growth Model', *Metroeconomica*, **56**, pp. 360–92.

Hadri, K. (2000), 'Testing for Stationarity in Panel Data', *Econometrics Journal*, **3**, 148–61.

Im, K.S., M.H. Pesaran and Y. Shin (2003), 'Testing for Unit Roots in Heterogeneous Panels', *Journal of Econometrics*, **115**, pp. 53–74.

Islam, N. (1995), 'Growth Empirics: A Panel Data Approach', *Quarterly Journal of Economics*, **110**, pp. 1127–70.

Islam, N. (1998), 'Growth Empirics: A Panel Data Approach – A Reply', *Quarterly Journal of Economics*, **113**, pp. 324–9.

Kočenda, E. (2001), 'Macroeconomic Convergence in Transition Countries', *Journal of Comparative Economics*, **29**, pp. 1–23.

Kočenda, E. and D. Papell (1997), 'Inflation Convergence within the European Union: A Panel Data Analysis', *International Journal of Finance and Economics*, **3**, pp. 189–98.

Kutan, A.M. and T.M. Yigit (2004), 'Nominal and Real Stochastic Convergence of Transition Economies', *Journal of Comparative Economics*, **32**, pp. 23–36.

Kutan, A.M. and T.M. Yigit (2005), 'Real and Nominal Stochastic Convergence: Are the New EU Members Ready to Join the Euro Zone?', *Journal of Comparative Economics*, **33**, pp. 387–400.

Lee, K., M.H. Pesaran and R. Smith (1997), 'Growth and Convergence in a Multi-Country Empirical Stochastic Solow Model', *Journal of Applied Econometrics*, **12**, pp. 357–92.

Lee, K., M.H. Pesaran and R. Smith (1998), 'Growth Empirics: A Panel Data Approach – A Comment', *Quarterly Journal of Economics*, **113**, pp. 319–23.

Levin, A., C-F. Lin and C-S.J. Chu (2002), 'Unit Root Tests in Panel Data: Asymptotic and Finite Sample Properties', *Journal of Econometrics*, **108**, pp. 1–24.

Mankiw, G.N., D. Romer and D. Weil (1992), 'A Contribution to the Empirics of Economic Growth', *Quarterly Journal of Economics*, **107**, pp. 407–38.

Quah, D. (1999), 'Ideas Determining Convergence Clubs', available from http://econ.lse.ac.uk/staff/dquah/currmnul.html, accessed June 2006.

Sala-i-Martin, X. (1996), 'The Classical Approach to Convergence Analysis', *Economic Journal*, **106**, pp. 1019–36.

Sarajevs, V. (2001), 'Convergence of European Transition Economies and the EU: What do the Data Show', Bank of Finland Institute for Economies in Transition (BOFIT) Discussion Paper No. 13.

Sarno, L. (1997), 'Policy Convergence, the Exchange Rate Mechanism and the Misalignment of the Exchange Rates', *Applied Economics*, **29**, pp. 591–605.

Solow, R.M. (1956), 'A Contribution to the Theory of Economic Growth', *Quarterly Journal of Economics*, **70**, pp. 65–94.

Su, J-J. (2003), 'Convergence Clubs Among 15 OECD Countries', *Applied Economics Letters*, **10**, pp. 113–18.

Summers, R. and A. Heston (1988), 'A New Set of International Comparisons of Real Products and Price Levels Estimates for 130 Countries, 1950–1985', *Review of Income and Wealth*, **34**, pp. 1–26.

Svejnar, J. (2002), 'Transition Economies: Performance and Challenges', *Journal of Economic Perspectives*, **16**, pp. 3–28.

Swan, T.W. (1956), 'Economic Growth and Capital Accumulation', *The Economic Record*, **32**, pp. 334–61.

14. Knowledge externalities and growth in peripheral regions

Fabiana Santos, Marco Crocco and Frederico Jayme Jr

INTRODUCTION

The discussion of externality has assumed a central position since the emergence of the so-called New Endogenous Growth Theory (NEG) theory. In these models, which follow the neoclassical approach to economic growth, the widespread existence of externalities is essential for the appearance of increasing returns at the aggregate level, which offset decreasing returns at the firm level. Among different kinds of externalities, the literature has dedicated a significant amount of effort to discussing the knowledge ones (Romer, 1990; Grossman and Helpman, 1991; Aghion and Howitt, 1992). In these models it is assumed that knowledge externalities can be spread over any kind of space.

The aim of this paper is to discuss and deny the validity of this assumption. Although this point has already been discussed by some scholars in the heterodox tradition (Nelson, 1998; Martin and Sunley, 1998, among others), we would like to bring into discussion a new perspective that analyzes the validity of this assumption in peripheral regions/countries. It will be argued that there are some peripheral structural conditions that constrain the generation, transfer and absorption of knowledge externalities. Above all, it will be argued that the construction of 'space' in the periphery is determinant for the absence of widespread diffusion of this kind of externality. This conclusion implies that the generality of the NEG theory is very difficult to assume.

I. ENDOGENOUS GROWTH THEORY: GENERALITIES

Endogenous growth theory or New Growth Theory (NGT) argues that increased returns to scale are the key element to explain its theory. It is the outcome of externalities that can arise from specific types of investment: R&D, investment in capital goods and human capital (Romer, 1986, 1990; Jones, 1995). Since technology is the engine of economic growth, in the Romer (1986) version of NGT, ideas play a central role in his model by means of their externalities. It assumes that ideas are public goods, as long as they are nonrivalrous. Therefore, as Jones (1998) highlights, this nonrivalry generates increasing returns to scale and imperfect competition, which is the key to understanding the spillovers of ideas in NGT. Some models argue that spending on R&D can generate a sustainable growth rate of output as this type of spending generates more and better, either final and intermediate, goods. The increase in better intermediate goods works to increase the overall productivity of the productive sector (Romer, 1990; Grossman and Helpman, 1991; Aghion and Howitt, 1992).

Other models, like Romer (1986), assume that the degree of technological development of an economy is directly related with the amount of capital goods of this economy. So, there is a direct relationship between the amount of capital goods and technological development, as the process of learning by doing operates to increase technological knowledge. When an isolated firm increases its own stock of capital, it is at the same time increasing the stock of capital of the whole economy and the knowledge that has been produced by the use of this new capital good spills over to the rest of the economy.

Usually, the endogenous growth models assume that there is 'constant or decreasing returns at the level of the individual firm but with positive spinoffs between them'. New growth models have their results depending on the existence of increasing returns to scale, which are derived from the existence of externalities. The latter comes in the form of spillover effects for the economy as a whole: education, invention and learning networks.

As the famous textbook emphasizes:

> One of the main contributions of new growth theory has been to emphasize that ideas are very different from other economic goods. Ideas are nonrivalrous: once an idea is invented, it can be used by one person or by one thousand people, at no additional cost. . . . In particular, the nonrivalry of ideas implies that production will be characterized by increasing returns to scale. (Jones, 1998, p. 86)

In sum: knowledge externalities are the key for understanding increasing returns to scale in NGT. One important aspect to stress is the fact that NGT does not split with the neoclassical theory of growth. In fact, NGT intends to contribute to a better modelling of growth using the same instruments of the neoclassical framework, such as externalities, as well as increasing returns in a typical production function with perfect substitutability between labour and capital. While in the Solow–Swan model technology is totally exogenous and available, in the NGT growth model technological progress is driven by research and development.

II. KNOWLEDGE EXTERNALITIES AND THE TRANSFER PROCESS – THE IMPORTANCE OF GEOGRAPHY

The central point of the argument is the understanding that externalities do not flow over space. For the majority of the types of externalities, especially those related to knowledge, the features of the space surrounding the places of its generation and of its absorption are fundamental, or essential, for their 'operationality'. By 'operationality' we mean the unpaid side-effects of one producer's output or input on other producers. An externality does not exist until the moment that one producer takes the advantage (or disadvantage in the case of a negative externality) of the action of another producer. Until this moment, the mere fact that, for example, one producer spends money in R&D to generate new knowledge does not mean that an externality is created. When another producer uses this new knowledge, without paying for it, in her/his productive process, only then occurs the transformation of this new knowledge into externality. This means that the uncontrolled outcome of a productive action of one producer only becomes an externality when it has an economic value for other producers.

The point that we would like to stress here is that both the space surrounding the producer that has generated this 'uncontrolled outcome' and the space surrounding the producer that uses this outcome are equally fundamental to define the economic value of this 'uncontrolled outcome'.

Having this comment in mind, we believe that the concept of 'centrality' can help to improve our understanding of the 'externality phenomenon' in contrast to more simplistic views that completely disregard the underlying conditioning factors for the 'externality phenomenon' to take place.

II.1 Centrality and the Geographical Dimension of the Transfer Process

The regional economics literature has highlighted that the development of a series of activities – particularly services – is essential for the generation, transfer and absorption of knowledge. These activities are directly related to the emergence of urban densities that represent minimum scales for the emergence of external economies stemming from urban agglomerations.[1] This process allows diversification and accessibility of several kinds of services and goods, since they make up the confluence and overlapping of market areas.[2] The analysis of such a possibility requires the understanding that urbanization may be characterized by two movements: *concentration* and *centralization*. *Concentration* is related to urbanization in cities. *Centralization*, in turn, as Christaller (1966) has shown, consists of the unequal development of urban centres, implying relative concentration of economic activities in large urban centres. Christaller (1966) argues that a large urban centre relies on high-quality, complex, specialized – central – services that provide it with economic higher efficiency than that found in smaller centres. The author's major concern refers to the formation of urban-centre networks as well as the reasons for the existence of different city sizes and its irregular distribution over space. Therefore, the author develops the notion of *central goods* and *service*s and *central place networks*.

The 'centrality' characteristic of a *central place* stems from a region's quality of supply of service, which may have a relationship with high population density and economic activities[3] so as to allow this region to supply central goods and services, such as knowledge intermediaries,[4] wholesale and retail trade, banking, business organizations, administrative services, education, entertainment facilities, and so forth. That is to say, a *central place* would play the role of a *locus* of central services for itself and for the immediately neighbouring areas (supplementary region). From this definition of central place, Christaller admits the existence of a *hierarchy of central places*, in accordance with smaller or greater availability of goods and services that need to be centrally localized (central goods and functions). The position of an urban region in the hierarchy of central places is defined by the size of its market area and degree of complexity and essentiality of goods and/or services it provides to its polarised area.

It is widely recognised in the literature on regional economics that 'centrality' is essential to the appearance of externalities that are derived from the diversification of the industrial structure. This point is particularly emphasized by Jacobs (1966) through the concept of *economic reciprocating system*. It is defined as the process of diversification of the productive system associated with the introduction of new kinds of products in different kinds of sectors, made possible by the development of the exportation sector. This process allows the urban region to increase its economic performance as it increases its exports of goods and service. This will attract diversified firms to the region, thereby working to increase the agglomeration externalities of the local and, hence, making the region even more attractive to other business activities and people. Moreover, as an urban region moves (upwards) in the hierarchy (and thus becomes of higher centrality order) it displaces other region(s). This is a process that, left to its own course, will increase regional disparities and turn the space more fragmented or fractured.

What has been argued here is that the concept of centrality is fundamental for the occurrence of some forms of externalities, especially knowledge spillovers and the transfer process. These kinds of externalities are present in many neoclassical models of endogenous growth as those analyzed above. The importance of centrality to the occurence of these kinds of externalities can be visualized from the discussion of two special features: a knowledge demand and the existence of knowledge-intermediaries.

First of all, it is worthwhile noticing that in the majority of the studies on knowledge spillovers the role of knowledge demand is usually overlooked. As pointed out by Howells (2002, p. 879), this is a result of the fact that these studies assume a 'traditional "public good" notion of knowledge and its costless characteristics. . . . On this basis, demand in a sense need not be considered, since knowledge would somehow permeate to those who needed it'. Taking this fact into account, one can argue that for a knowledge spillover to become an externality it is necessary to have the existence of some economic activities that use it in the productive processes. In other words, it is necessary to have the existence of an economic opportunity for the application in the productive process of this new knowledge.[5] This can equally happen in areas that show a significant degree of specialization – as conceptualised by the Marshallian industrial districts, *clusters*, or Italian industrial districts – or in areas that encompass several clusters (that is, have a highly diversified productive structure). However, it is possible to assume that in areas with diversified productive structures, due to a high degree of centrality, the opportunities for the use of that new knowledge are greater than in places of lower ranking and thus, knowledge spillovers will in fact occur and become an important externality of the place.

Knowledge–intermediaries, in turn, can be defined as *conduits of knowledge-transfer* and, hence, significantly contribute to the emergence and diffusion of knowledge externalities. It can take both a formal shape – like specialist service design, research, engineering and consultancy firms – and an informal one – like membership in a learning society or industry association, or attendance at conferences and workshops. The existence of these *conduits* is directly associated with the degree of centrality of a specific region, as the latter is determined by the supply of more sophisticated, complex and central goods and services. That is to say, a place that can supply those types of services is a central place of a higher rank. In this sense, it is possible to argue that the higher the centrality, the easier is the emergence of knowledge externalities.

From what has been said so far, one can assert that the more central places exist within an economic territory, a country for example, and the higher their rank, the easier it becomes for the knowledge externality to spread around this economy and to impact positively on its performance. In other words, the spatial dimension of knowledge demand has impacts on the scale of knowledge spillovers. The size of knowledge demand in the locale of its generation and the existence of knowledge-intermediaries can determine whether a knowledge spillover will be lost (undiscovered) or ignored, and to what extent. The neoclassical new endogenous growth models seem to assume that every economic space has a sufficient number of central places for the effect of knowledge externalities to generate increasing returns in aggregate to compensate the diminishing returns on the firm level. In other words, there is an assumption of non-segmented space in new endogenous growth theories.

II.2 Knowledge Spillovers and 'Absorptive Capabilities'

Another feature that is essential for the outcome of a spending in R&D (or from a learning-by-doing process) to be transformed into externality is the capacity of this outcome to be incorporated by other producers. This capacity, in its own turn, depends on two aspects: the way this outcome is divulged and the capability of other producers to understand and absorb it.

The first aspect is related to the channels of transmission of knowledge, especially the technological ones, inside a society. To be spread to the whole economy, knowledge diffusion requires that channels of communication among agents be perfect, in the sense that once a specific knowledge is generated, it can be passed easily and quickly. However, this assumption does not take into account the concept of *knowledge base*. This concept is related to the characteristics of the knowledge used in an innovation. According to Dosi (1988, p. 224), various sorts of pieces of non-excludable knowledge are used in the solution of most technological problems: universal versus specific; public versus private; and articulated versus tacit.

Universal knowledge refers to knowledge that has a large applicable understanding, based on principles that are well known and pervasive, whilst specific knowledge denotes that particular to some activities. Moreover, there is that knowledge that is public in the sense that it is available in scientific and technical publications, as opposed to knowledge that is private because it is protected by laws (patents). Moreover, in the case of public or codified knowledge, it is necessary that the access to this new knowledge be equally distributed over sectors and regions, which implies the existence of homogeneous access to this knowledge. This implies the existence of a uniform distribution of universities, colleges and research centres, which can educate people to deal with new technologies. It also requires the widespread existence of libraries, bookshops, and technological assistance. Finally, some sorts of knowledge are well articulated, and for the most part are written down in manuals, books and so on. In contrast, there is also that kind of knowledge that is tacit, meaning that it comes from an unarticulated experience and practice. Given the relevance of tacit knowledge to our discussion, we think that a further analysis of this concept is worthwhile.

The concept of tacit knowledge has been synthesized by Polanyi (1958; 1967) in the following statement: '*We can know more than we can tell*' (1967, p. 4; italics in original). Basically, the meaning of tacit knowledge can be understood when we realize that we can recognize the face of our neighbours without being able to explain how we recognise the face. In other words, 'perception is determined in terms of the way it is integrated into the overall pattern' (Nonaka and Takeuchi, 1999, p. 216). Polanyi argues that knowledge acquisition is 'the outcome of an active shaping of experience performed in the pursuit of knowledge' (Polanyi, 1967, p. 6).

Polanyi stresses the importance of experience, self-involvement and commitment to the understanding of tacit knowledge when he identifies tacit knowing as indwelling. As pointed out by Nonaka and Takeuchi:

> To know something is to create its image or pattern by tacitly integrating particulars. In order to understand the pattern as a meaningful whole, it is necessary to integrate one's body with the particulars. Thus, indwelling breaks the traditional dichotomies between mind and body, reason and emotion, subject and object, and knower and known. Therefore, scientific objectivity is not a sole source of knowledge. Much of our knowledge is the fruit of our own purposeful endeavours in dealing with the world. (Nonaka and Takeuchi, 1995, p. 60)

While explicit knowledge can be expressed in a systematic and formal way in the form of hard data, scientific formulae, codified procedures or universal principles, tacit knowledge, as Polanyi has pointed out, is highly personal and hard to formalise. In this case, proximity or contact face-to-face is a necessary condition for its diffusion.

From the previous discussion, it is possible to argue that the diffusion of tacit knowledge requires proximity (geographical and cognitive), in a way that allows a network to be constructed, like in the Marshallian industrial districts. As tacit knowledge is not expressed in a formal code, its transmission is based on the share of cultural values, informal codes, routines, or in other words the share of institutions in a broad sense[6] (formal and informal). These institutions are geographically localized, giving the transfer process of tacit knowledge a strong local dimension. Moreover, even codified knowledge requires tacit knowledge to be learned. In the words of Howells (2002, p. 876):

> . . . tacit knowledge, situation and locational context do play a significant role in the use and spread of *codified* knowledge. Thus, although codified knowledge may be more ubiquitous and accessible, its interpretation and assimilation are still influenced by geography.

This contrast to the conventional view of neoclassical endogenous growth models, which assume that knowledge is a 'public good' (and, hence, non-excludable and nonrivalrous) that can *flow* freely, without any costs and frictions, between individuals (or firms) (Howells, 2002). Another form of knowledge prized by these models is that knowledge embodied in goods (notably capital goods). In this case, *flow* of knowledge (spillover effects) rather than *sharing* of knowledge is made possible by (free) trade relations[7] (Park, 1995; Coe and Helpman, 1995).

The discussion above indicates that the diffusion of knowledge externality is strongly influenced by the quantity and the quality of channels of communication of scientific knowledge and by the degree of proximity between the 'producers' and the 'users' of this knowledge.

Another important basic feature for knowledge to be transformed into externality is the capacity of potential users to understand and incorporate this knowledge into their productive processes, which depends on their absorptive capability. According to this approach, knowledge is not a good that anyone can pick up from the shelf. The introduction of a new piece of knowledge is surrounded by what has been labelled 'dynamic uncertainties' (Camagni, 1991). As put forward by Lawson (1999), these uncertainties would be related to: (1) information complexity and difficulty in identifying useful information, which requires a '*searching function*'; (2) the problem of *ex ante* inspection of the qualitative characteristics of inputs, equipment, and so forth, which requires a '*screening function*'; (3) the difficulty in processing available information, which requires a '*transcoding function*'; and (4) the difficulty in assessing the results of actions taken both by the firms' and other agents in their relationship (competitors, suppliers, and so on), which requires a '*coordination mechanism*'. The firm capabilities to deal with these uncertainties will vary among sectors, size and location, meaning that the absorption potential of an externality will vary over space.

One can summarize the discussion above saying that whether or not an 'uncontrolled outcome' resulting from a learning or R&D process will be transformed into a widespread externality (that is, a knowledge spillover), will depend upon the existence of a system of innovation. In Lundvall (1992, p. 2) words:

> . . . a system of innovation is constituted by elements and relationships which interact in the production, diffusion and use of new, and economically useful, knowledge and that a national [regional, local] system encompasses elements and relationships, either located within or rooted inside borders of a nation [regional, local] state.
>
> This definition makes it clear that a system of innovation is both a social system and is spatially defined, 'including all parts and aspects of the economic structure and the institutional set-up affecting learning as well as searching and

> exploring – the productive system, the marketing system and the system of finance present themselves as sub-systems in which learning takes place'. (Lundvall, 1992, p. 12)

II.3 Bringing Back the 'Space' – Knowledge and Geography

Taking this theoretical discussion into account, it is possible to argue that some neoclassical new endogenous growth models assume the existence of both well distributed central places and well organized national systems of innovation as a natural feature of all economies. This is an assumption necessary to make their theory a general one, capable of being applied in any space. Our argument is that this is a highly unsatisfactory approach, to the extent that, in our view, *territory* is a social space that goes beyond its physical geographical endowments. This means that it is impossible to analyze a specific space without understanding the conventions, values, rules and institutional arrangements that define its social forms of production. This means that history is an essential feature of every space and defines its social forms of production. In this sense, it is impossible to assume that central places and national systems of innovation are ubiquitous.

One can assume for theoretical purposes that, in general, space in developed countries is more homogenous (that is, the urban hierarchy is less fragmented or more horizontal than vertical), due to some similar features of their development. Developed countries show a degree of urbanization, income distribution and a system of innovation that, although not identical, can be assumed to be very similar. Moreover, it can be assumed that the spaces in developed countries are endowed with those conditions necessary for the occurrence of externalities derived from knowledge. That is, in developed countries, there exists a balanced distribution of central places and a system of innovation that works to facilitate the generation, diffusion and absorption of knowledge externalities.

However, this homogeneity of space is not found in peripheral countries in comparison with the developed ones. Moreover, this lack of homogeneity is likely to happen within peripheral countries, as will be discussed next.

III. EXTERNALITIES AND PERIPHERAL SPACES

The major question to be answered is: What would be the conditions in force in peripheral 'spaces' that impair knowledge externalities to be generated, diffused and absorbed by economic agents located in this space?

The answer is necessarily related to the need for amplifying the analytical range of studies of knowledge externalities so as to embed peculiarities associated with the peripheral condition of the country and that of the location itself (internally related to the country). We believe that peripheral development constraints may provide elements for the understanding of the potentialities and limits to the spillover of knowledge externalities. In what follows, two aspects related to these constraints are discussed: the construction of capabilities and urban spaces in peripheral countries.

III.1 Capabilities and the Transferability of Knowledge

First of all, it is important to make it clear that in our view peripheral countries do not innovate in the sense of being capable of shifting the frontiers of knowledge – an attribute of the core. Rather, peripheral countries do invest in knowledge acquisition effort, that is, to acquire, master and, sometimes, improve upon existing knowledge, borrowed from the core.

The fact that knowledge is not a 'free, public good' and, relatedly, that the appropriation and transfer processes are not automatic, passive and costless – but rather require minimum social capabilities and active actions to absorb and process it – imply that the 'non-excludability' and 'high mobility' of knowledge assumptions that sustain endogenous growth models' knowledge spillover effects are hard to accept for peripheral countries. This is because a place's (country or region of a country) social and technological capabilities together with the whole set of institutions (summarised by the concept of systems of innovation) that support the building and development of its capabilities are fundamental to determining whether or not it will be capable of benefiting from the externalities of existing (borrowed) knowledge.

Evolutionary *catching-up* models (including the concept of systems of innovation) in conjunction with Cepal's contributions on the problems of generation of technical progress in the context of core–periphery relations are helpful to the understanding of the reasons why widespread knowledge externalities, which are necessary to generate endogenous growth, are not always possible and, accordingly, why it is so difficult for a peripheral country to become an innovation-generating space.[8]

Evolutionary *catching-up* models based on technological diffusion have already shown that latecomer countries benefit from positive externalities of access to technologies coming from leader countries at the technological frontier, provided that they meet the threshold precondition of the so-called 'minimal social absorption capacity' (Abramovitz, 1986).[9] Countries below a threshold level would be excluded from the benefits of knowledge spillovers and, hence, from the opportunities brought about by technological *catching-up*. As technologies are becoming increasingly more demanding in terms of the capabilities they require to be adopted, the periphery will always be in a disadvantaged position to the extent that structural factors are difficult to change and the process of knowledge transfer does not have any in-built forces to reverse cumulative causation.

In fact, the incomplete character of peripheral systems of innovation (which captures the institutional dimension of peripheral development) helps to explain why capabilities are underdeveloped (Albuquerque, 2000). As Albuquerque (2000) maintains, 'incomplete national innovation systems' are characterised by: (a) scientific and technological infrastructure of a relatively small scale; (b) atrophy of the 'T' in the binomial S&T; (c) distribution of R&D spending skewed towards the public sector, which leads to the atrophy of the 'D' in the binomial R&D owing to the small presence of the private sector; and (d) significant inter-sectoral heterogeneity of technological development favouring sectors based on natural resources, where former state companies are concentrated.

One may ponder that in view of the growth of integrated production systems, with facilities at different levels of technological complexity, the need for building local capabilities in peripheral countries is reduced. However, one must consider the importance of 'technological isolation effect', associated with limited spillover effects to the peripheral host country. As it is well known, technological effort in R&D – which by its own nature, demands a significant locational indivisibility – is ultimately concentrated in TNCs' parent countries (the core). Conversely, TNCs transfer those simpler technologies which only require the efficient use of the capabilities existing in these countries. In fact, they have no interest in investing to create more advanced capabilities in peripheral countries. Thus, subsidiary firms located in peripheral countries would perform simpler strategic functions (basically manufacturing), fundamentally requiring operational capabilities. In this regard, the conjunction of information- and knowledge-poor environments of the subsidiaries' sites in a peripheral country with subsidiaries' dependence on knowledge transfers from the parent may create a 'technological isolation effect', to borrow Howell's (2002) words, characterised by 'little information and knowledge interaction with its local environment'. This means that the potential of learning, the scope for technological upgrading, and knowledge spillovers are considerably limited. In other words, the construction of capabilities of peripheral countries based on the transfer of knowledge produces an environment that does not facilitate the widespread generation of the knowledge externalities as assumed by the neoclassical endogenous growth. There are, therefore, important institutional constraints at work that check endogenous growth models' pretence of offering a general theory equally applicable to peripheral and core countries.

III. 2 The Construction of Centralities in the Periphery

Another element to be considered in a broader analysis would be related to the constitution of a peripheral urban space endowed with a complex service network, necessary for the generation, transfer and absorption of knowledge. As shown above, this is a process that is directly related to the construction of central places inside a region. This construction, in its turn, implies the centralization and concentration of services over the space. In the words of Lemos, 'urban concentration and centralization are nothing but the major way through which capitalism accelerates the market area growth, in order to guarantee the productivity development of the tertiary sector' (Lemos, 1989, pp. 293–4).

Such processes give rise to unequal development not only among countries but also among regions of a country, determining the emergence of polarizing regions and polarized regions. In order to understand this process of unequal regional development, it is necessary to understand that this process is essentially constrained by a country's income dimension and the inequality of its distribution in space: the greater the income spatial distribution, the greater the possibility of the emergence of several central places. In this way, compared with core countries and given the dimension and inequality of income distribution in peripheral countries, one may expect that the possibility of emergence of central places is naturally smaller in the latter. This would be the factor explaining the existence of a number of *incomplete* urban nuclei, in the sense that they are not able to embed a complex service sector as well as few *complete* urban agglomerations in the periphery. Moreover, the gap – in terms of the quality of the services supplied – among these few *complete* urban agglomerations and those many *incomplete* urban nuclei is very large.

Associated with the previous aspect of small urban density, it is also relevant to take into account that, in peripheral conditions, the urban nucleus' surroundings are usually that of subsistence (meaning that the diversification and quality of services and goods is low, as well as the level of income) when the region is lagged in the national context. In this case, the tertiary concentration and centralization does not follow a territorially contiguous urban hierarchy, and a strong segmentation of such a hierarchy in the regional surroundings occurs, mainly through the absence of medium-sized urban centres which would be able to absorb complementary industrial activities sustained by the supply of services in the urban centre pole. This means that there is a low productive complementariness between the pole and its vicinity and that social immersion (backward and forward linkages) is very weak.

Thus, the small service diversification and quality – especially in the case of those modern ones, which function as knowledge-intermediaries and inputs – and the strong segmentation of urban space are unable to feed and sustain knowledge externalities. This feature of peripheral countries/regions is a constraint to the widespread occurrence of increasing returns as theorised by the proponents of endogenous growth theory (Romer, 1990; Grossman and Helpman, 1991; Aghion and Howitt, 1992, among others).

IV. FINAL REMARKS

In this paper we have discussed one hypothesis that is fundamental in the neoclassicals'endogenous growth theory. In these models, it has been assumed that knowledge externalities can be spread over any kind of space. The heterodox literature has already challenged this assumption (Nelson, 1998; Martin and Sunley, 1998, among others). Even among geographers this assumption has been denied (Feldman, 1994). Most of this literature has argued that knowledge spillovers are geographically confined and are related to the amount of knowledge–generating inputs. Our argument goes further in two dimensions. First, we have argued that one important element of this geographical constraint to the generation and diffusion of these knowledge externalities is the degree of *centrality* shown by a region. It is this centrality characteristic that facilitates both the knowledge–generating inputs and knowledge–intermediaries (the *conduits of knowledge-transfer*). Moreover, local capabilities are also fundamental for the absorption of knowledge externalities. In the same way, these local capabilities are geographically constrained and their building is influenced by the degree of centrality.

Second, we have argued that both the construction of local capabilities and *centrality* have structural constraints in peripheral countries. These structural constraints impose serious problems to the neoclassical endogenous growth theory assumption that knowledge externalities are easily widespread diffused over the space. Taken together, these arguments make it very difficult to accept the generality of the NEG theory.

NOTES

1. As Lemos says, it is important here to distinguish a city from an urban center. 'The concept of city involves geographic-populational idea, while by "urban" or "urbanization", we understand the formation process – capitalist – of a "complex of services"' (Lemos, 1989, p. 216).
2. Market areas is defined here both in the Weberian sense (Weber, 1929), that is the locus where several economic transactions occur, and the Löschian sense (Lösch, 1954), a localized space whose property is the accessibility to a given service.

3. It is important to note that the concept of centrality should not be confused with the concept of urbanization. Although it is possible to argue that there is some relationship between both concepts, the idea of centrality implies the supply of special kinds of services, usually more sophisticated. A region with a large population without this kind of service will have a lower degree of centrality than another region with less population but with a supply of more sophisticated services, especially the productive ones.
4. Knowledge intermediaries will be discussed in more detail later.
5. It is important to note this knowledge demand can assume two forms. First it can appear in the form of a market for knowledge, implying the existence of some transaction process, like the purchase of a catalogue or a scientific book, the hiring of qualified personnel, and research agreements or contract R&D. Secondly, it can take place in non-market terms, like informal trading and reciprocal knowledge sharing via joint-venture operations (Howells, 2002).
6. See in this regard the concept of 'relational proximity' as found in Amin and Cohendat (1999, 2000).
7. Although it should be recognised that international trade plays a central role in spillover effects between countries, it is debatable, as the discussion above has shown, whether all features of knowledge can be embodied in goods. In this regard it is important to consider that for some types of knowledge the transfer process is via non-market mechanism. Moreover, it must be considered that underlying structures for the successful absorption of these spillovers are required (as captured by the concepts of national system of innovations, social and absorptive capabilities).
8. We are not saying that it is impossible for a latecomer country to climb the technological ladder, as the cases of Japan, Germany and more recently East Asian countries such as Korea, Taiwan and Singapore seem to prove. These cases indicate that purposeful State action is *sine qua non* to determine the outcomes of the catching up. This partly explains the difference between Latin American and East Asian achievements in the catching-up process.
9. In other words, a minimum level of basic social capital, such as the physical infrastructure (telecommunications, transport and electricity networks) and an organized education and health system.

REFERENCES

Abramovitz, M. (1986), 'Catching up, forging ahead, and falling behind', *Journal of Economic History*, **66** (2), 385–406.

Albuquerque, E.M. (2000), 'Domestic patents and developing countries: arguments for their study and data from Brazil (1980–1995)', *Research Policy*, **19** (9), 35–52.

Aghion, P. and P. Howitt (1992), 'A model of growth through creative destruction', *Econometrica*, **60** (2), 323–52.

Amin, A. and P. Cohendat (1999), 'Learning and adaptation in decentralised business networks', *Environment and Planning D*, **17**, 87–104.

Amin, A. and P. Cohendat (2000), 'Organisational learning and governance through embedded practices', *Journal of Management and Governance*, **4**, 93–116.

Camagni, R. (1991), 'Local milieu, uncertainty and innovation networks: towards a new dynamic theory of economics space', in R. Camagni (ed.), *Innovation Networks: Spatial Perspectives*, London: Belhaven-Pinter.

Christaller, W. (1966), *Central Places in Southern Germany*, Englewood Cliffs, NJ: Prentice Hall.

Coe, D.T. and E. Helpman (1995), 'International R&D spillovers', *European Economic Review*, **39**, 859–87.

Dosi, G. (1988), 'The nature of the innovative process', in G. Dosi et al. (eds), *Technical Change and Economic Theory*, London: Pinter Publishers.

Feldman, M.P. (1994), *The Geography of Innovation*, Dordrecht: Kluwer.

Grossman, G.M. and E. Helpman (1991), *Innovation and Growth in the Global Economy*, Cambridge, MA: MIT Press.

Howells, J. (2002), 'Tacit knowledge, innovation and economic geography', *Urban Studies*, **39** (5–6), 871–84.

Imai, K. and H. Itami (1984), 'Interprenetration of organization and market', *International Journal of Industrial Organization*, **2**, 285–310.

Jacobs, J. (1966), *The Economy of Cities*, New York: Vintage Books.

Jones, C. (1998), *Introduction to Economic Growth*, New York: W.W. Norton and Company.

Jones, C. (1995), 'R&D based models of economic growth', *Journal of Political Economy*, **103**, 739–84.

Lawson, C. (1999), 'Towards a competence theory of the region', *Cambridge Journal of Economics*, **23**, 151–66.

Lemos, M.B. (1989), 'Espaço e Capital: um estudo sobre a dinâmica centro x periferia' (Space and Capital: a study about the dymanics of center x periphery), PhD Thesis, State University of Campinas, Campinas.

Lösch, A. (1954), *The Economics of Location*, New Haven, CT: Yale University Press.

Lundvall, B. (1992), *National Systems of Innovation*, London: Pinter.

Martin, R. and P. Sunley (1998), 'Slow convergence? The new endogenous growth theory and regional development', *Economic Geography*, **74** (3), 201–27.

Nelson, R.R. (1998), 'The agenda for growth theory: a different point of view', *Cambridge Journal of Economics*, **22** (4), 498–512.

Nonaka, I. and H. Takeuchi (1995), *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*, Oxford and New York: Oxford University Press.

Nonaka, I. and H. Takeuchi (1999), 'A theory of the firm's knowledge-creation dynamics', in A. Chandler Jr., P. Hagström and O. Sölvell (eds), *The Dynamic Firm*, Oxford: Oxford University Press.

Park, W.G. (1995), 'International R&D spillovers and OECD economic growth', *Economic Inquiry*, **33**, 571–91.

Polanyi, M. (1958), *Personal Knowledge: Towards a Post-Critical Philosophy*, London: Routledge and Kegan Paul.

Polanyi, M. (1967), *The Tacit Dimension*, London: Routledge and Kegan Paul.

Romer, P. (1986), 'Increasing returns and long-run growth', *Journal of Political Economy*, **94** (5), 1002–37.

Romer, P. (1990), 'Endogenous technological change', *Journal of Political Economy*, **98**, S71–S102.

Weber, A. (1929), *Theory of Location of Industries*, Chicago, IL: University of Chicago Press.

15. Knowledge, human capital and foreign direct investment in developing countries: recent trends from an endogenous growth theory perspective

Diana V. Barrowclough*

I. INTRODUCTION – THE LENS OF ENDOGENOUS GROWTH THEORY

This paper describes recent new trends in foreign direct investment (FDI) in developing countries from an endogenous growth theory perspective, with the aim of highlighting the importance of an enabling policy framework to help make the most of the opportunities that the new trends contain. The focus is on FDI in research and development (R&D) and in tourism: two new and dynamic economic activities that rely on high value-added forms of human capital, characterised by knowledge, experience and technical expertise. These forms of human capital are important for their contribution to the process of innovation and technological change, which in turn feeds into productivity, competitiveness and, ultimately, human and social development. The experiences of the countries described in each activity reflect their differing sources of comparative advantage, stemming for the most part from relative endowments of expertise, knowledge and experience. In the R&D sector, developing countries act as sellers of their endowments of human capital and expertise. In tourism, the same countries are buyers.

The insights of endogenous growth theory (EGT) can offer a useful conceptual framework with which to analyse the trend and its determinants, and to make some observations about the role of policy to help create a

* Dr Barrowclough is an economist at the United Nations Conference for Trade and Development in Geneva; and a member of the Cambridge Centre for Economic and Public Policy. The comments expressed here are not to be taken as the official views of UNCTAD. She can be reached at Diana.Barrowclough@unctad.org.

'virtuous' path. The EGT approach applies in five main ways. Firstly, endogenous assets of human capital help drive the new trend, and are likely to determine the growth path of the countries involved. Secondly, from this perspective the path of technical progress is explicit, in the sense that endowments of human capital contribute to technological change that is endogenously determined. This acts so as to attract foreign investors, who will add to the process of technical progress and growth, as well as being a source of technical progress and growth in its own right. Thirdly, there appears to be the cumulative path of causation and reinforcement that is characteristic of the EGT approach. Fourthly, the critical sources of human capital endowments that are driving this trend are, to a certain extent, able to be influenced by policy – an important feature of the EGT approach. Knowledge, expertise and technical skills are created by appropriate policies in education, training and R&D, in addition to the natural endowment effects that are created by demographics.

Finally, the examples lend support to arguments that *supply*-led approaches to endogenous growth explanations of technical change and economic growth and *demand*-led approaches need to be considered together. In the case of human capital, experience and skills, the characteristics of non-rivalness and increasing returns mean that the interrelation between supply and demand on the part of countries that are well endowed with these resources vis-à-vis countries that are not, may be more than usually self-reinforcing. It seems likely to contribute to a cumulative causal effect of the kind envisaged by Kaldor and others, where, to coin the phrase, 'history matters'. This means abandoning the standard neoclassical assumption that returns to capital are diminishing, as is characteristic in the EGT approach. Rather, intellectual resources are renewed and regenerated the more they are used: the process of doing R&D does not empty a country's reservoirs of R&D capacity, but rather increases them. The trend of FDI into human capital resources reinforces these effects, as it contributes directly and indirectly through spillover effects, in a dynamic manner. These mean that the EGT approach, coupled with H–O–S type insights relating to comparative advantage and trade, offers an appropriate framework of analysis.

On the other hand, our ability to evaluate the likely impact of the trends described below is still limited, not least because the trends are new and still highly concentrated. Also because we are looking at human capital endowments that are intermediate services, or inputs into the production of goods and services, rather than their outputs, it is difficult to measure their quantitative effect. The appropriate framework with which to examine impact would be through the national innovation system (NIS)[1], and it is expected that FDI will affect its levels of innovation, national competitiveness and export revenues and eventually growth, but it is highly unlikely that we will ever be able to make a clear causal link between FDI inputs into human capital, and these final outputs. The introduction of foreign investment adds a new layer of firms, processes and behaviours to the NIS, and this can bring with it potential costs as well as benefits. In the R&D sector, for example, benefits may include adding a more business-like orientation to public laboratories and researchers, including helping to turn inventions into innovations that can be produced and sold on global markets. In tourism, it can mean providing access to international training chains that will enhance the skillsets of local tourism workers. On the other hand, costs may include crowding out in local R&D labour markets; or the downsizing, either temporary or permanent, of local R&D enterprises. This appears to be less likely in tourism than in R&D but the point is that the implications of these effects may depend on whether one looks from the perspective of the employee, the foreign affiliate, competing or collaborating domestic firms, and the home or host nation. Once one goes beyond the simple observation of growth effects on national employment, productivity and competitiveness, which appear to be largely positive, more subtle effects may depend on factors relating to knowledge and intellectual ownership, including the contractual agreements between parties, and on the distribution of the products of human capital, such as intellectual property rights, or brand names and reputation.

The structure of the chapter is as follows. Section II outlines recent trends in FDI into human capital endowments of R&D in developing countries. Section III describes demand and supply determinants of the trend. Section IV highlights selected policy initiatives that have contributed to these determinants and Section V considers some of the future implications of the trend, in particular those that may contribute to its reinforcement. Section VI introduces the counter-point of recent trends in FDI in tourism in developing countries, where for the most part, the distribution of human capital assets is different, meaning that countries that are sellers of human capital and knowledge in the R&D sphere remain buyers of knowledge and expertise when it comes to tourism. Section VII concludes by summarising the main themes from an EGT perspective.

II. FDI AND R&D IN DEVELOPING COUNTRIES: RECENT TRENDS

When large international corporations locate their R&D activities in another country, in essence they are buying or investing in the process of learning and innovation. This differs from other transactions, where it is the output or the consequence of investments in learning and innovation that is being transferred. It also differs from historical forms of FDI, where firms invested in processes that were more related to goods, such as manufacturing, mining or agricultural processing. Those trends tended to be driven by host countries' comparative advantages in terms of relatively low-cost labour and raw materials, rather than in the high value-added skills of learning and knowledge. Secondly, transnational corporations (TNCs) are internalising these host country forms of human capital within their firm, through the creation of a foreign subsidiary, rather than simply outsourcing and sub-contracting to a firm or an expert who is external to the firm. Internalisation with a foreign subsidiary makes sense from an industrial organisation (IO) approach, given that R&D tends to be a very strategic activity, involving knowledge that firms prefer to hold close to the centre, and furthermore it usually embodies high transactions costs of transfer. The effect of this is to help link the domestic innovation system in both home and host countries into a globalised system that can offer potential benefits that will deepen and widen their human capital stocks. It may also affect human capital flows, for example, if the incentives to gain higher education and R&D skills change. What is unexpected is that this is occurring now in developing countries. The following pages outline the main trends as they are currently occurring.

a. Rapid Growth in FDI in R&D in Developing Countries

The most detailed data currently available come from the United States, and this stands as a benchmark that reinforces the picture gained more recently from surveys and case-studies. These figures are low compared to FDI in general, but important because R&D is an activity that is especially significant in terms of its potential for value-added and technical progress, compared say to manufacturing. US majority-owned foreign affiliates increased their investments in R&D activities in developing countries by around 200% since 1994 compared to increases of only 73% in developed countries. This has taken the share of their R&D activities in developing countries from 7.5% of total offshore R&D to 12% (see Table 15.1). This trend is all the more significant when considered in terms of the values spent: expenditure rose from $902 million to $2705 million, a 200% increase in dollar terms, and considerably more in terms of the purchasing power of the dollar, which fell significantly over the period. Most of this went into Developing Asia, which saw US foreign R&D expenditure rise five-fold from $408 million to $2113 million in 2002 before falling slightly in 2003. In terms of the share of the total, this was an increase from 3% to 10%. At the level of the individual country, some increases were particularly

Table 15.1 Growth in FDI in R&D in developing countries exceeds developed countries (United States MOFA, 1994–2003; millions of dollars)

	1994	1999	2003	Rate of change (%)
Developing Countries	902	2031	2705	199.9
Developing Asia	408	1400	1907	367.4
China	7	319	646	9128.6
Hong Kong	51	214	227	345.1
India	5	20	81	520.00
Singapore	167	426	516	209.0
Latin America	477	613	689	44.4
Brazil	238	288	326	37.0
Mexico	183	238	280*	53.0
Transition Economies	5	54	51	920.0
Developed Countries	10,975	16,113	18,935	72.5
Total (millions of $)	11,877	18,144	22,328	88.0

Note: * includes new EU members. Mexico data from 2002. MOFA = majority-owned foreign affiliates.

Source: United States Department of Commerce, Bureau of Economic Analysis (2006).

notable, for example China (where investment grew almost 100-fold) and India. The transition economies of central Europe, including new European Union members, also saw particular increases. Expenditure on R&D is of course an input, rather than an output that can be directly linked to measures of business success or national competitiveness, but nevertheless the intuition of EGT that these trends are likely to be significant is compelling. Africa, for example, hardly registers, with US foreign affiliates investing less than $30 million in R&D, the vast majority of which was in South Africa.

Other countries do not provide such detailed statistics, but the general picture from what is available corroborates the trend of internationalising R&D to developing countries. For example, over the period 1995 to 2003, the 20 largest TNCs in Sweden doubled their total expenditure in R&D abroad from $1.1 billion to $2.5 billion (taking the share of foreign-located R&D from 22% to 43% of the total). This was an increase of almost 100% in the eight years covered, however what is even more remarkable is the speed of the increase in investment in developing countries – which rose six-fold (more than 400%) over the same period. Even though the absolute amounts are still small, at only $0.18 billion, the relative increase has been marked, and the impact in the host countries is likely to have been considerable.

Finally, survey evidence suggests that these trends will increase in the future. For example, an UNCTAD survey (UNCTAD, 2005) of the world's largest business R&D investors found that more than two-thirds of the respondents expected to increase their R&D activities in foreign locations. The top three destinations were China, United States and India, followed by the United Kingdom and Germany in fourth and fifth place respectively. A survey by the Japan Bank for International Co-operation found that Japanese companies increased their number of R&D bases offshore to 310, an increase of more than 70% in the years from 2000 to 2004. Of this, bases established in developing countries tripled, to a total of 134. This was most marked in China, which now accounts for a quarter of all the R&D units. Finally, evidence from new greenfield FDI in R&D indicates that of the close to 2000 new R&D projects established through FDI in the two years from 2002 to 2004, the majority were in developing countries or emerging economies (LOCOMonitor database, 2005). Developing Asia and Oceania alone accounted for one half of the new projects (notably India and China).

b. R&D is 'Basic' as well as 'Adaptive' and Boosts Technical Progress

There is a wide variation in the nature of the R&D conducted in these various foreign affiliates and subsidiaries in developing countries, but some broad trends stand out, reflecting the nature of the comparative advantages of host countries (and in some cases, its absence). Developing Asia is mostly the site for R&D in computers and electronic products; India is software development; and Brazil and Mexico are in chemicals and transport. In contrast the Latin American experience with R&D through the form of FDI is mostly about what was been called 'tropicalisation': with the exception of Brazil and Mexico, R&D is mostly confined to the adaptation of technology or products for local markets, rather than creating new products for sale in global markets. In Africa, R&D through FDI is virtually non-existent, with the exceptions of Morocco and South Africa. In the new European Union economies of the Czech Republic, Hungary and Poland, the majority of the R&D conducted by FDI is associated with parallel manufacturing activities carried out in those countries by the foreign firms. The feature that really stands out in Developing Asia, therefore, is not only the scale of the foreign investment in R&D but also the fact that it is directed towards innovative R&D, where new products and processes are developed, aimed as much for global markets as for regional or local ones. (It is worth noting, however, that even adaptive R&D has development benefits, if it leads to finding more appropriate ways of using new technologies, or more efficient ways of using old ones.)

c. Developing Country Corporations are also Establishing R&D Affiliates Abroad

A final new trend that is emerging, although to a lesser degree, is that some of the leading developing countries are also shifting their R&D activities abroad. This is the most recent of all the trends to internationalisation of innovation systems, and as such is still occurring only on a very small scale, and by a handful of companies. In some cases the firms are investing in developed countries, mostly to gain access to the knowledge bases in the United States or Europe (for example, Chinese firms Huawei and Haier; as well as leading Indian software firms (Infosys and Wipro). They are also setting up R&D facilities in other developing countries, in an extension of the general South–South trend that is starting to mark world trade and investment patterns. In these cases, however, the firms tend more to be targeting their local markets, and so the nature of the work conducted is more about localisation and adaptation and not so much global R&D aimed for world markets.

III. THE TREND IS BOTH SUPPLY-LED AND DEMAND-LED

At one level, this trend is the logical extension of the fragmentation of manufacturing activities already seen in previous decades. Trans-national corporations are *de-verticalising*, as they break up the production process into constituent parts and carry each out in different locations depending on the advantages of each location. This often leads to the use of outsourcing or sub-contracting for the various stages of the process, in a Smithian vision of a chain of independent specialists conducting a stream of discrete activities. It can be seen as the final stage in the process of disintegration into individual parts that has marked the value and production chain of manufacturing in the last few decades.

Now it is the turn of services, and in particular the knowledge and human capital aspects of services production, to be broken down into separate parts and traded individually, rather than being conducted within a single firm. The fact that the R&D activities described in this paper are kept internal to the firm, rather than being outsourced, supports theories of asymmetric information and transaction costs, such as the stream of literature spawned by Coase, Teece and Williamson. Activities that are strategic, and that embody high transaction costs (including tacit knowledge as well as the uncertainties that are inherent in research) are those that are mostly likely to be retained within the firm. Of all the activities that are conducted at arms length from headquarters, R&D has been the last one to 'go' offshore, even with the improvements seen in information and communications technology (ICT).

Reflecting this, many people would have expected to see R&D activities remain in a cluster around a firm's headquarters. Now it seems that the 'cluster' around headquarters can be a virtual one, and that clusters of R&D-related activities, even if carried out by a number of different firms in another country, may be more important than clusters relating to the mixed production activities of a single TNC. What is more surprising than this, though, is that the clusters are emerging in developing countries that still lack basic infrastructure in other parts of their economy. This is unexpected, because R&D is an activity that needs the support of a strong innovation system, providing a network of enterprises and skills of a depth traditionally only found in developed countries.

Explanations for this outcome can be found by applying the perspective of an endogenous growth theory approach; in particular, by bringing together both the supply-led and demand-led influences that have fuelled this trend. Given that these trends are primarily about human capital, expertise and experience, the Kaldorian view that history matters (see Kaldor, 1985) seems to be particularly appropriate, not least in the light of the survey evidence described above, about likely trends in the future.

Supply factors driving the trend include the immense reservoirs of highly skilled R&D staff available in a selection of developing countries. Countries such as China, India and Republic of Korea have a powerful comparative advantage in R&D activities, and also a competitive advantage. They are able to supply skilled personnel in the large numbers that are required by corporations of the developed world, and at a fraction of the cost. In 2000–2001, China, India and the Russian Federation accounted for almost one third of all tertiary technical students in the world; and developing countries accounted for seven of the top ten countries in terms of numbers of tertiary technical students (see Table 15.2).

A firm wishing to set up a R&D lab in software development can cut its costs by up to two-thirds by locating in one of these countries. Of course, one could say that this is no different from the traditional labour–cost arguments for FDI in labour-intensive activities in manufacturing, but it is more than just a question of cost. The skillsets and educational backgrounds provided are at a very high level. For firms that need to employ several hundred highly-skilled researchers in a single lab, these countries offer their only possibilities to find the skillsets in the numbers required – the skills are simply not available in sufficient scale in developed home countries. The

Table 15.2 Top ranked countries in terms of total tertiary students, and technical students

Total tertiary students		Tertiary technical students	
Rank	Thousands	Rank	Thousands
1 United States	13,596	1 China	2,580
2 China	12,144	2 Russia	2,388
3 India	9,834	3 India	1,913
4 Russian Federation	7,224	4 United States	1,719
5 Japan	3,973	5 Korea	1,000
6 Indonesia	3,018	6 Japan	817
7 Korea	3,004	7 Ukraine	644
8 Brazil	2,781	8 Germany	637
9 Egypt	2,447	9 Indonesia	586
10 Philippines	2,432	10 Mexico	577
11 German	2,158	11 United Kingdom	496
12 Thailand	2,095	12 Brazil	468
13 United Kingdom	2,067	13 Spain	460
14 Mexico	2,048	14 Iran	456
15 France	2,032	15 Taiwan	369

Source: UNCTAD (2005) based on UNESCO statistics.

effect is further compounded with the return of the skilled 'diaspora', whereby highly skilled researchers are leaving their expatriate employment in the West to come home either to join the foreign subsidiaries, or to set up their own enterprises.

Demand factors include the ever-increasing need for innovative skills and capacities on the part of the corporations in the developed world. As tasks become increasingly complex and specialised, the need for highly trained or innovative staff intensifies. Moreover, competition means that innovation needs to be faster than before: the time-span of the product cycle is shortening. One of the few ways that developed world TNCs can keep up the pace and level of innovativeness required is to seek knowledge and R&D capacity offshore. It is not only about costs – it is also about the speed of innovation and product development required. This feature relates to global demand but there are also local demand conditions that contribute to the trend. TNCs that wish to seek markets for their products in the newly emerging economies (as opposed to productive assets or resources) may need to adapt their products to local needs and incomes. This has been called the 'tropicalisation' of innovation. While this kind of R&D offers different kinds of benefits in terms of skills and expertise development, it does nonetheless further reinforce the demand to tap into human capital assets in developing countries.

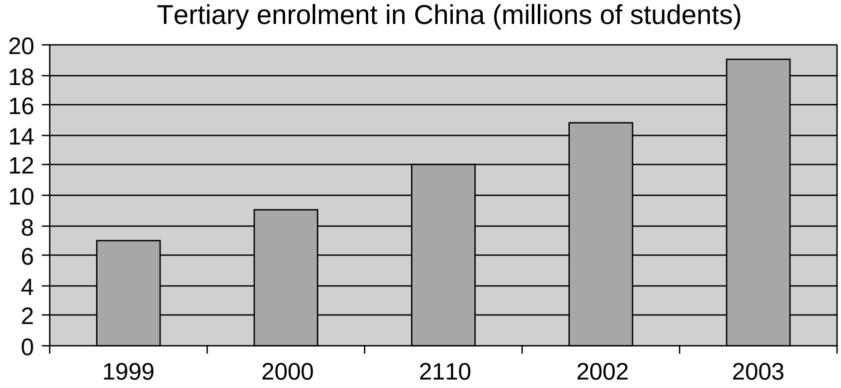
These demand and supply factors are also mutually reinforcing, because of the fact that human capital stocks of knowledge, experience, and innovative capacity are not depleted by consumption – rather they increase. The more that the endowments of human capital that reside in developing countries are used, the more they will be enhanced – especially if their use is in a challenging way. The more skills they will create, and the stronger will be their comparative advantage. This will be leavened, of course, to some extent if wages rise, and there is some evidence already of wage and price inflation in locations such as Bangalore, for example, but as described above cost is only one of the issues and the shortage of such skills in the developed countries is such that the demand appears set to continue for some time yet. It is notable that cost savings are often not the most important factor attracting foreign firms to invest in R&D skills offshore – often it is simply the case that the assets do not exist sufficiently at home. This appears to support the Kaldorian view that endogenous growth is a mutually reinforcing process. As described elsewhere in this volume, it may be a historically contingent process characterised by concepts such as cumulative causation and evolutionary change (Roberts and Setterfield, 2005).

These features are particularly appropriate given that the trend is about human capital, learning and experience. These elements of comparative advantage are characterised by non-rivalness, in the sense that one does not deplete human capital resources by using them – rather, we increase them through practice and learning. Such re-generative properties are seldom to be found in other resources or assets. They mean that returns are not only non-diminishing, they are actually increasing, helping to contribute to a kind of lock-in of causation. Skills and experience breed more skills and experience, in a mutually reinforcing cycle. This is in addition to the more conventional kind of advantage that stems from demographics. While many of the developing countries that have participated in this trend have the benefit of population size, some do not. These characteristics help contribute to a process of technical progress that becomes self-reinforcing, from both the demand and the supply sides.

IV. THE ROLE OF POLICY

Endogenous growth theory approaches tend to prioritise the role of policy, and this view is supported in the evidence that policy stances of developing countries, with regard to creating and reinforcing R&D capacity have been particularly important in this respect. Most countries offer a variety of incentives and promotional tools aimed to attract foreign investors, including low rates of corporate tax, tax holidays, or waiving duties on imported materials. However, in order to capture the virtuous benefits of FDI in R&D for which many countries hope, policies that are aimed at boosting domestic resources in terms of education, knowledge, research skills and entrepreneurship are particularly important. Having a large population is one obvious advantage, but this is not enough to create comparative advantages in the highly-skilled human capital sectors required to carry out R&D where the skills that are involved take years to nurture and to develop. The Asian success story in particular reflects a long-term commitment on the part of national policy makers towards boosting the knowledge, technical expertise and research capacity of their research workers. In China for example, the number of tertiary students has increased at an extremely high rate (Figure 15.1) in recognition of the economic value to be gained from investing in this form of productive capacity, in addition to the other social and cultural arguments in favour of higher education. Students in China and Singapore and other developing countries are also encouraged to gain additional experiences through post-graduate study abroad through a system of student grants and study programmes for working professionals.

Domestic policy tools that have been used by the developing countries featured in this paper include investing in public education, with a strong focus on technical skills such as engineering, mathematics and science; support for technological upgrading in the private sector, to enhance the scope of the work carried out by skilled employees and also to maintain



Source: UNCTAD (2005), derived from Ministry of Education, The People's Republic of China.

Figure 15.1 China has significantly increased its investment in human capital

competitiveness; and the promotion of science parks, R&D labs and incubators that will help promote innovation. Such incubators may be particularly important in the case of small and medium enterprises (SMEs). These can be especially innovative in many cases because their small size enables them to be fluid and responsive, but they are also hampered by finding it difficult to access capital, or need help with networking, marketing or business consulting.

Publicly financed venture capital funds have been used in many developing countries, given that SMEs are more than usually likely to find it difficult to get start-up capital, especially in the high risk areas of R&D. Also, most industrial policies that aim to promote the knowledge-based industries are now targeted at creating clusters, rather than supporting specific industries, so that the spillover effects can be dispersed more widely and effectively in terms of the knowledge assets that are created. The actual policies used vary according to the level of development of the country. Many of the countries described in this chapter specifically target R&D investors, using a combination of performance requirements (such as obligations to invest in training of local workers) as well as incentives. In some countries, there have also been specific policies to induce skilled diaspora to return home. The point is that human capital is as much a created asset as an inherited one, and the policy framework has important long-term implications. This is of course, a lesson that developed countries have also learned – in particular, for example, Ireland.

V. FURTHER IMPACT – DYNAMIC EFFECTS MAY BE SELF-REINFORCING

a. Human Resource and Employment

One of the most obvious impacts has been in terms of human resources. United States foreign affiliates (for which the best data exist, and which are used as a proxy for firms from other developed countries) increased their total R&D staff by more than 20% over the period 1994–1999, while domestic R&D staff for the same firms at home increased by only 3.5% (US National Science Foundation). Most of these foreign jobs were created in developed rather than developing countries, but anecdotal evidence suggests that the rate of growth is even larger in developing countries, especially in the years after 1999, for which data does not yet exist. In India, for example, other sources have estimated that some 40,000 R&D-related jobs were created through FDI-Greenfield investments in the years 2002–2004 (LOCOMonitor database). A typical example of this was General

Electric's new $80 million multi-disciplinary R&D centre in Bangalore, which employs 2,300 people. In China, the GE's new Global Research Centre opened in 2000 with 20 researchers, had 500 by 2003, and was expected to employ 1,200 by 2005 (UNCTAD, 2005). These are large numbers given the nature of the sector, and the other striking feature is the high level of skills and education required for these jobs, implying that the value-added content of the work is high. In the Chinese example above, more than 80% of the engineers have PhD degrees.

b. Reverse Brain Drain

Perhaps one of the most important effects may be that the development of new career opportunities in foreign affiliates and domestic firms appears to be contributing to a reverse brain drain effect that is seeing Indian and Chinese engineers, scientists and entrepreneurs return home in numbers that would have been unthinkable twenty years ago. Once attracted to work in universities, R&D institutions and TNC labs abroad, they are increasingly lured home by the entrepreneurial opportunities that are now apparent. In some cases, they are responding to direct incentives from government, but more generally it seems that they are given confidence by the fact that their countries are not only 'in the loop' of the global knowledge network, but are even starting to drive it. Some have retained their links with the TNC with whom they were originally associated at headquarters. For example, the Shanghai-based contract chemistry research and product development company, WuXi Pharma Tech, now employs more than 200 scientists and earns revenues of over $10 million. Founded by a Chinese national with experience in US TNCs, and with a senior scientific management team of Chinese nationals that were brought back from the USA, the firm still counts amongst its stable of customers the TNCs with whom many of its researchers were once employed. Again, this makes sense in the light of IO theory.

c. Moving Up the Value Chain

Another effect that is likely to contribute to the longer-term dynamic effects is that the enhanced investment in R&D capacity is helping countries to move up the value-chain in terms of domestic production processes. Adaptive and innovative R&D aimed at local markets can help to bring about process and product upgrading in domestic industries that will contribute to increased productivity and competitiveness. Innovative R&D for global markets may have more indirect effects through the spillovers of knowledge – not least the knowledge of how to manage large-scale R&D projects and how to professionalise the process of R&D. Developing countries can, of course, do this at their own pace, and many now have sufficient human capital resources to do so. The point is, however, that FDI offers a way to speed up the process, through leap-frogging bottlenecks such as lack of resources or low local demand.

d. Limitations

Of course, TNCs are in business, not philanthropy, and while an increasing number are following corporate social responsibility policies and the like, their strategic goals cannot be assumed to be the same as those of developing countries themselves. There will be limitations to the extent that the countries in which they invest can capture benefits from their presence. These do not necessarily affect the cumulative causation of the endogenous features described above, but they do affect the way that its benefits will be distributed between firms and between nations. In the first instance, TNCs may attempt to limit the potential for spillovers to the wider domestic economy, through restricting the work carried out in their foreign labs to non-core technology transfers. In addition, much of the knowledge and technology transferred will be tacit within the firm, and employees cannot readily pass it on to others outside the firm.

These points are, however, really instances where host countries fail to capture the full benefits possible. There may also be costs that are more direct. For example, an issue that is arising is the extent to which the host country can benefit from any intellectual property rights (IPRs) that are created through the R&D activities conducted in foreign-owned labs located in their countries. There is plenty of evidence that IPRs are boosted when foreign firms conduct research in host developing countries, as measured for example in the number of patents that have been created (see UNCTAD, 2005 for a fuller description of these effects). This is a sign that more innovative work is being carried out – or at least that it is leading to patentable outputs. It is a different question however, as to how the revenue streams and financial benefits from these innovations will be distributed. Aghion and Tirole (1994), for example, have provided a useful framework for interpreting the distribution of benefits accrued to the tasks of financing, creating, owning and using innovations, and these endogenous features will determine the extent to which developing countries are able to achieve the fullest benefits from their prior investments in national human capital endowments.

To a certain degree, this is a question of the nature of the contract between the various parties in the innovation: be they joint-ventures

Table 15.3 Potential human capital implications of FDI in R&D, for host developing countries.

Potential benefits	Potential costs
Increased investment in R&D.	Downsizing of existing local R&D, on either a temporary or permanent basis.
A more business-like approach.	Loss of control of R&D technology and processes.
Moving from 'invention' to 'innovation'.	Poor compensation for intellectual property rights.
Increased employment; increased complexity of employment tasks.	Crowding out in the R&D labour market.
Improved structure and performance of NIS.	Crowding out in R&D capital markets.
Knowledge spillovers.	Technology leakage.
Contribution to industrial upgrading.	Race to the bottom and unethical behaviour.

Source: UNCTAD (2005).

between universities in developing countries and TNCs located there; or between the national researchers actively engaged in discovering new patentable processes and products, and their foreign employers. When local partners do not have experience or equal bargaining power, they may not receive their fair allocation of the rights and responsibilities attributed to the innovations to which they have contributed. Lack of ownership of IPRs may mean that developing countries fail to receive the appropriate share of the revenue streams that can follow a successful innovation, and/or, it may make them dependent on foreign firms for their future technological progress. Researchers in developing countries often lack the experience and expertise to deal with these issues, and one priority for many is to strengthen their domestic institutions, to improve their ability to deal with IPRs effectively. Table 15.3 summarizes some of these issues.

VI. THE COUNTER EXAMPLE OF TOURISM

a. Economic Importance of Tourism Activities

Growing investment in tourism-related activities in developing countries is another new trend upon which many countries are pinning their hopes for advances in human and economic development. Tourism is seen by many as one of the most promising ways to boost employment, and to earn government revenues, and to earn foreign exchange. Developing countries are well endowed with some of the sources of comparative advantage in tourism, and are potentially able to exploit these advantages relatively quickly – much more quickly than in R&D activities, for example, where the necessary assets take decades to create.

The value of the sector is already evident. Tourism is the only services-related economic activity where the African countries as an aggregate group have a positive trade balance. For the majority of developing countries, tourism is one of their top five sources of foreign exchange, and for up to one third of developing countries, it is their main source of foreign exchange. In the Dominican Republic, for example, tourism exports rose from $500 m in 1985 to close to $8 billion by 2006 (WTTC, 2006); due in large part to FDI. Another relative newcomer, Tanzania, has seen tourism revenues more than double in the last five years, taking the sector's share of foreign exchange earnings from around 20% to close to 50%. Even in one of the last LDCs to come to global tourism, Bhutan, which has one airport and two aeroplanes serving a tourism industry that consisted in 2005 of 12,000 tourists in total, tourism is the single biggest earner of hard currency. Tourism is also a major employer in most developing countries, especially when supporting activities such as transport and agriculture are included. In some countries it can provide as much as one third of total employment. It is looking particularly attractive for natural resource-oriented developing countries that have experienced 'jobless growth' in recent years.

Foreign investment is typically seen as the primary engine of growth in this sector, and certainly FDI in tourism in developing countries and Least Developed Countries is gathering pace step-by-step with increasing tourist arrival figures, especially in the last five years. Hotels located in developing countries make up a small but rapidly growing proportion of the portfolios of hotels held by the world's largest hotel chains, and most hotel chains plan to increase rather than decrease their presence. This is also evident in the most recent trends of what is called South–South tourism and trade, where investment and consumption of tourism activities between developing countries rather than from north to south, is increasingly important. The World Tourism Organisation estimates that only one tenth of all tourism movements are truly 'international' – the majority of tourists travel within their own nation, or at least within their region. The rise of tourism consumption from the new middle classes emerging in some developing countries is attracting investment from southern hotel chains, as well as northern ones, for example.

b. Empirical Trends in Tourism

FDI is usually seen as a package of different elements, including equity but also knowledge, experience and access to markets. Developing countries usually want the whole package in order to gain, or augment, key resources. In the R&D example above the package almost invariably included equity capital, but in tourism it increasingly does not. TNCs participate in non-equity ways, meaning that they bring the other parts of the package but not finance. Because physical capital can be separated from human capital, including for example brand name and reputation, through the use of contracts, increasingly TNCs have a presence in a developing country through a management contract rather than equity. Local investors put up the capital, and purchase the hotel chain's assets of knowledge, management experience and access to markets. The TNC may open the door to global capital markets (for example, foreign banks are more willing to finance hotels in developing countries when a well-known foreign chain is involved) but essentially the TNC does not bring a significant share of the equity.

This puts the developing countries in the position of being buyers of TNC human capital endowments, such as management experience, and branding and marketing skills, where in the R&D sector the same countries were rather sellers. This reflects the relative scarcity of the endogenous endowments of the human capital required in order to make the most of their other natural endowments – such as beautiful natural scenery, climate, and cultural or religious attractions. Local capital is often available, but the expertise is lacking – even in countries such as China or India, where the highly skilled and highly trained R&D sector suggests that its lack is due more to lack of experience than lack of managerial talent. To put this in the EGT context described above, there are demand-led attractions bringing TNCs to developing countries (such as local investors seeking partners with management and expertise skills) as well as supply-led attractions such as, the opportunity to profit from growing local tourism markets, the destination's natural endowments, and so on.

Another supply-led feature that is prompting TNCs to seek non-equity participation in developing countries is the more general strategy whereby hotel chains are divesting themselves of equity in their hotels, and using the freed-up capital for other ventures. The strategy does not appear to be caused by perceptions of country or regional risk and is rather a reflection of the relative scarcity of expertise and knowledge and how this affects the opportunity cost of investing physical capital. This is also occurring within South–South investment trends: the new Kenyan hotels opening up in

Tanzania and Uganda, and the South African hotels opening in Rwanda, Uganda, Tanzania and India, are often bringing with them only their managerial expertise, because local investors are willing and able to provide the capital.

c. Implications

It is still too early to evaluate the implications of this trend, but UNCTAD surveys of hoteliers, investors and governments in developing countries suggest that the non-equity participation of foreign investors in hotels brings with it human capital opportunities that may be as valuable if not more so, than having provided equity finance. For example, the foreign affiliate may have international marketing and advertising reach that considerably raises the profile of the developing country. Bhutan is an example: the entry of two high-profile hotel groups dramatically increased international knowledge of the country and what it could offer, producing spillover benefits for all in the industry. Bhutan has been open, more or less, for at least a decade, but in the year since the establishment of the global chains, tourist numbers virtually doubled – most of whom stayed in local hotels and not the two that created the publicity.

On the other hand, a frequent criticism of foreign involvement in tourism is that local enterprises are excluded from the value-chain. It is argued that too high a proportion of the value created is lost to imports of raw materials, food and beverage purchases and repatriated profits; or does not arrive in the country at all as it is spent on air fares and tour operators' fees. This has been called leakage. However, the term needs to distinguish between what would have existed without the foreign investor. It should also be compared to the costs of tourism in general, as local hotel investors can also import inputs, pay off foreign debt, or move profits offshore. The more appropriate way to look at this problem is to see it as a failure to capture value – and hence the need to introduce policies that will enable the destination to capture more of what is created. Clearly one need is to invest in local human capital formation, so that local investors can also participate in the higher-value-added end of the tourism chain, being managers and suppliers of high value inputs, rather than simply financing the infrastructure. Developing countries will need to initiate policies that will promote local skill enhancement, through enterprise development and training schemes such as the formation of hotel management schools and so on; and they may also encourage training or procurement linkages through foreign and domestic hotels, following some of the examples learned from FDI in R&D activities.

VII. CONCLUSION

FDI is a source of investment finance and technology transfer that most governments, be they in developed or developing countries, actively seek. Of all the forms of FDI, R&D tends to be seen as the plum – the work is highly skilled, well paid, non polluting, and does not usually require large prior investment in public infrastructure such as airports or roads. It offers a way for host countries to access global research and innovation networks from which they might otherwise be divided; to retain or to attract back their most highly educated and trained citizens; and to move up the value chain. While being aware of the need for a cautious approach with respect to the intellectual property rights issues and other costs described above, it generally embodies all that most governments want, and unsurprisingly many countries place a high priority on its attraction. Already a few developing countries have a comparative advantage in this sector, thanks in part to policies in the past that have invested in education and other sources of the human capital that is needed.

In tourism, by comparison, countries have often been more cautious, seeming simultaneously to fear the effects of FDI as much as they court it. At base, this reflects concerns about the impact of tourism in general, rather than whether the investment is through foreign or domestic

Table 15.4 Investment and human capital in developing countries: R&D and tourism

FDI in human capital assets	R&D activities	Tourism activities (hotels)
FDI into developing countries	Developing countries sell skills in R&D, education and innovation, to foreign TNCs. TNCs bring capital, expertise and access to markets. *Scale: significant and growing*	Developing countries buy management, marketing, reputation and branding skills from TNCs. TNCs bring expertise and access to markets but not capital. *Scale: significant and growing*
FDI out of developing countries	Developing country TNCs buy R&D skills in developed countries abroad. *Scale: small but growing*	Developing country TNCs sell human capital skills to investors in other DCs and developed countries. *Scale: still very small*

enterprises; nowadays, generally tourism is seen as a priority sector much like R&D. Unlike R&D however, non-equity forms are typical and TNCs bring with them the managerial expertise and knowledge of international tourism markets that is needed rather than physical capital. Compared to the R&D story, however, developing countries have tended to concentrate their policy initiatives into attracting FDI and have not sufficiently invested in policies that will enable them to benefit from it. The EGT approach implies that more attention is needed to establishing a supportive policy environment that will help local tourism workers, managers and investors to gain the knowledge and human capital assets that are required, in order to help them participate more fully and equally in the tourism value chain.

NOTE

1. The NIS is the interrelation of firms, governments, and non-firm institutions such as universities and research centres that exist within a country, including tacit or intangible factors such as their incentives, their regulatory frameworks and their objectives. See Freeman 1995, Lundvall 1992 and Nelson 1993 for an overview of this literature.

REFERENCES

Aghion, P. and P. Howitt (1998), *Endogenous Growth Theory*, Cambridge, MA, MIT Press.

Aghion, P. and J. Tirole (1994), 'The management of innovation', *Quarterly Journal of Economics*', **109**(4), pp. 1185–209.

Balasubramanyam, V.N. et al. (1996), 'Foreign direct investment and growth in EP and IS countries', *The Economic Journal*, **106**, pp. 92–105.

Fines, B. (2000), 'Endogenous growth theory: a critical assessment', Working Paper number 80, School of Oriental and African Studies, Department of Economics, London.

Freeman, C. (1995), 'The national system of innovation in historical perspective', *Cambridge Journal of Economics*, **19**, pp. 5–24.

Kaldor, N. (1985), *Economics Without Equilibrium*, Cardiff, University College of Cardiff Press.

Lall, S. (1983), *The New Multinationals: The Spread of Third World Enterprises*, London, Wiley/IRM.

Lall, S. (1992), 'Technological capabilities and industrialization', *World Development*, **20**(2), pp. 165–86.

Lundvall, B.-Å. (1992), *National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning*, London, Pinter.

LOCOmonitor FDI database, OCO Consulting Ltd.

Nelson, R. (1993), *National Innovation Systems: A Comparative Analysis*, Oxford and New York, Oxford University Press.

Narula, R. and A. Zanefi (2004), 'Globalisation of innovation: the role of multinational enterprises', in J. Fagerberg, D. Mowery and R. Nelson (eds), *The Oxford Handbook of Innovation*, Oxford, Oxford University Press, pp. 318–45.

Roberts, M. and M. Setterfield (2005), 'What is endogenous growth theory?', Working Paper, CEPR conference, University of Cambridge, August.

Romer, P. (1994), 'The origins of endogenous growth', *Journal of Economic Perspectives*, **8**(1), pp. 3–22.

Romer, P. (1986), 'Increasing returns and long-run growth', *Journal of Political Economy*, **94**(5), pp. 1002–37.

UNCTAD (2005), *Transnational Corporations and the Internationalisation of R&D*, World Investment Report, New York and Geneva, United Nations.

UNCTAD (2005), 'FDI in tourism: the development dimension', Background note for Ad Hoc Expert Group meeting, Geneva, February.

United States Department of Commerce, Bureau of Economic Analysis (2006), 'Survey of US direct investment abroad', www.bea.gov/bea, accessed August.

Winter, S. (1987), 'Knowledge and competence as strategic assets', in D.J. Teece (ed.), *The Competitive Challenge: Strategies for Industrial Innovation and Renewal*, Cambridge, MA, Ballinger, pp. 159–84.

Zang, Y. (2005), 'Globalisation of R&D: new trends of economy globalisation', Paper presented at UNCTAD Expert Meeting on FDI and Development, Geneva, 24–26 January.

16. Is growth alone sufficient to reduce poverty? In search of the trickle down effect in rural India

Santonu Basu and Sushanta Mallick*

1. INTRODUCTION

In this paper we examine whether the trickle down effect has ever taken place in rural India.[1] One of the important sources of poverty is the existence of unemployment and seasonal unemployment in the rural areas of developing countries. The argument that growth alone will take care of poverty, referred to as the trickle down effect, appears to rest on the assumption that owing to the existence of a very large surplus labour supply, the initial rise in the growth of employment is unlikely to be accompanied by a rise in the wage rate. This assumption eliminates the possibility of the emergence of capital–labour substitution in the foreseeable future. Hence the argument can be made that growth will take care of poverty. In the case of India we know that in the past the government has changed its agricultural policy in a major way at least three times. This raises the question whether the trickle down effect ever took place in the rural areas. If not, the question is, why not? This is the subject matter of this paper.

In order to investigate this issue, the remainder of the paper has been divided into three sections. Section 2 examines why, if the trickle down effect was taking place in India, did the government intervene three times in the functioning of the agricultural sector? Section 3 examines whether there was a trickle down effect, by investigating whether there exists any empirical relationship between the changes in the poverty rate with either changes in output or with capital formation in Indian agriculture. Section 4 provides a theoretical explanation of why the benefits of growth did not trickle down to the poor in India. This is followed by the conclusion.

* We would like to thank I. Dutta, G.C. Harcourt, M. Hughes and T.K. Kumar for their helpful comments on the earlier version of this paper.

2. THE RELATIONSHIP BETWEEN CAPITAL FORMATION, GROWTH AND POVERTY

The argument of the trickle down effect appears to centre entirely on the growth of employment as a vehicle for reducing the incidence of poverty; therefore it is necessary not only to examine the impact of growth on poverty, but also to examine the relationship between the incidence of poverty and capital formation. It is the form of capital formation rather than its absolute value that will determine the growth of employment, and this will be reflected in the incidence of poverty. Therefore, if the trickle down effect did take place in India at any period we should expect to have not only a very high correlation between the capital formation and the GDP (*GDPAG*) but also a negative one with the incidence of poverty. Thus we need to examine the relationship between these three variables to see whether there exists any correlation between them. Our main aim is to examine whether poverty was consistently falling with the rise in the GDP and what its relationship was with capital formation.

Data on rural poverty (*POVRU*) is taken from the World Bank's web site (http://www.worldbank.org/poverty/data/indiadata.htm). Normally, there are two methods that are used for the measurement of poverty: head count ratio (*HCR*) and the poverty-gap index (*PGI*). The *HCR* indicates the proportion of the population living below the poverty line and the *PGI* mainly measures the depth of the poverty, that is the spread of the poor living below the poverty line. As our primary purpose is to examine whether the percentage of the rural population living below the poverty line over the years has consistently declined or not, rather than to measure the depth of the poverty, we decided to use the head-count ratio to examine whether any direct relationship can be established between the growth rate and poverty. The National Sample Survey (NSS), which provides poverty data, has not been conducted every year and consequently to calculate the trend in poverty levels one has to take these gaps into consideration. This means any calculation of the trend in poverty will always be accompanied by some degree of inaccuracy and therefore can lead to bias in interpretation; this is specially the case when the poverty figure fluctuates between every round.

For our chosen sample between 1951 and 1991, there are 12 missing values, and consequently we have filled the gap by interpolating from the observed values. A total of 41 observations covering the period 1951–1991 have been used for the calculation of the trend. We chose 1991 as the cut-off point for our sample mainly for three reasons. First, there was a change in agricultural policy in 1991. Secondly, although two more rounds of NSS data have been collected, the questionnaires have been changed, and as a result these data are not compatible with the previously collected data.[2]

Finally, it is known that in the early 1990s the number of people living below the poverty line increased by 13 million and subsequently the Indian government allocated Rupees (Rs) 350 billion to address this poverty (Mehta and Shah, 2003), and as the agricultural growth rate slowed down in the 1990s, the issue of the trickle down effect does not arise for this decade. Data on GDP in agriculture and capital formation in agriculture (CAPAG) are taken from several issues of National Accounts Statistics, published by the Central Statistical Organisation, India. GDP in agriculture has been used as a proxy for rural income and capital formation in agriculture has been used as an approximation (or indicator) of rural investment, with the aim being to measure the impact of the latter in the context of growth and poverty reduction.

Changes in the incidence of poverty are normally calculated with reference to changes in the per capita GDP. We chose not to use this method because in our opinion it portrays a very misleading picture. For example, when we consider the longer time period, it reveals that while per capita GDP was growing at 0.5% per annum, the incidence of poverty was falling by 0.8% per annum. This gives the impression that all you need to concentrate on is the growth rate, which in turn will take care of poverty; as will be shown later this is not necessarily the case. It is important to note that a rise or a decline in the growth of population may alter the value of per capita GDP, but a reduction in the incidence of poverty ultimately relies on an increase in the GDP growth rate (Dreze and Sen, 1995). Consequently, we decided to examine the impact of GDP on the level of poverty. There is a graph of the above three variables in Figure 16.1, and in Table 16.1, we present the trend rates of all three variables.

In relation to the longer time period, Table 16.1 suggests that poverty has fallen marginally with rises in both the growth rate and capital formation. Although this finding tends to support the view of those who claim that the trickle down effect worked in India, the result is not robust. Consequently, we decided to break down this long period arbitrarily into four sub-periods to examine whether a systematic trend can also be found between these

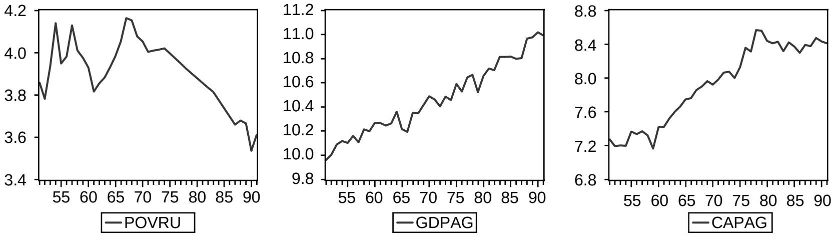


Figure 16.1 Rural poverty, GDP and capital formation

Table 16.1 Trend rates of rural poverty, GDP and capital formation in agriculture, 1951–1991 (%)

	Rural poverty	Agricultural GDP	Agricultural investment
1951–1991	**−0.8**	**2.4**	**3.7**
1951–1961	0.4	2.8	1.6
1961–1971	2.7	2.3	5.5
1971–1981	−1.7	2.6	5.9
1981–1991	−2.9	3.1	0.4

Note: Trend rates are calculated as the OLS regression coefficients on time.

three variables during these sub-periods. As all the policy changes overlap with each other between the 1960s and 1970s, any impact of policy changes upon these three variables to some extent will either be neutralised or minimised, thereby allowing us to investigate whether the trickle down effect did take place during these sub-periods. If it did take place we would expect to observe a systematic trend between these three variables: in the initial period, poverty may rise mildly with the rise in the GDP and capital formation, then it should start to fall.

Our result reveals that no systematic relationship can be found between the trend in poverty with either the growth rate or capital formation. It shows that while poverty was rising throughout the 1950s and 1960s with rises in both the growth rate and capital formation, poverty was falling with the rise in the growth rate subsequently. Interestingly, capital formation peaked in the 1970s and then declined in the 1980s, and poverty appeared to have declined at a faster rate during the last sub-period when compared to the previous periods.

This shows that there exists no systematic relationship between capital formation and the incidence of poverty. In fact, the puzzling feature of the 1950s and 1980s observations is that the growth rate was somewhat higher when the capital formation was low, raising a question about the nature of capital formation in the 1960s and 1970s. More importantly, the 1980s observations reveal that the poverty was falling at a faster rate compared to the previous sub-period. This raises an interesting question in relation to the 1980s, namely, what happened during this period to produce such a puzzling result?

In order to investigate this issue, we decided to divide the entire 40-year period into three sub-periods, now broadly following the policies that dominated each of these sub-periods, namely the pre-green revolution, green revolution and post-green revolution, to examine the impact of each policy upon the variables concerned and what caused the policy changes. The pre-green revolution period was dominated by the idea of implementing land reform policy and the promotion of cooperative based farming, while the green revolution was dominated by the adoption of High Yield Variety (HYV) seeds and an intensive application of inputs in the highest productive areas. The philosophy behind the green revolution was that technical progress sponsored by the state would eliminate institutional impediments. The post-green revolution period was dominated by the government's policy of increasingly divorcing itself from the strategy of the green revolution, and the promotion of the government's poverty alleviation programme. Although the poverty alleviation programme was officially launched in 1972, the effectiveness of this policy did not eventuate till the late 1970s and throughout the 1980s. From the mid-1970s the government increasingly alienated itself from the strategy of the green revolution, and instead of concentrating applications of inputs in the most productive areas, decided to spread the inputs more sparingly. The government, with the help of the nationalised banks, started to channel credit increasingly in favour of small and marginal farmers at a lower interest rate.[3] This, in turn, not only improved the small and marginal farmers' access to modern inputs such as fertilizer, but also made it possible to distribute fertilizer more sparingly. During the sixth and seventh five-year plan periods, that is from 1979 to 1983 and 1984 to 1989, the government undertook various programmes to address poverty. The government introduced the Integrated Rural Development Programme (IRDP), the National Rural Employment Programme and the Rural Landless Employment Guarantee Scheme, the latter two merging into Jawahar Rozgar Yojana.[4] Accordingly, we divided the sample. The first period is from 1951 to 1963. We end this period in 1963, as the green revolution marked a departure from the land reform policy and cooperative based farming. India adopted the HYV programme in 1964–65, so the second period is from 1964 to 1975. As the government changed its strategy from the mid-1970s onwards, we end the period of the green revolution in 1975. The third period begins in 1976 and ends in 1991, when the process of liberalisation started. By 'departure' we do not mean that the Indian government officially abandoned the policy of land reform and tenancy legislation, but that this policy, which dominated the minds of policy makers in the early years of independence, to some extent subsided. In fact, it was always the responsibility of each individual state to implement the policy. We present our results in Table 16.2 below.

Table 16.2 suggests that poverty was falling in all the three sub-periods, but the magnitude was very different for each period. For example, during the pre-green revolution period our results show that poverty declined at a rate less than the GDP growth rate, suggesting that the land reform, tenancy legislation and cooperative based farming had a negligible impact

Table 16.2 Trend rates of rural poverty, GDP and capital formation in agriculture during different policy regimes, 1951–1991 (%)

	Rural poverty	GDP in agriculture	Investment in Agriculture
1951–1991	**−0.8**	**2.4**	**3.7**
1951–1963	−0.3	2.4	2.6
1964–1975	−0.2	2.6	3.8
1976–1991	−2.7	3.1	−0.2

Note: Trend rates are calculated as the OLS regression coefficients on time.

on reducing poverty. This is largely because other than in a few states, the respective state governments were unable to implement the land reform and tenancy legislation policies effectively. The main aim during this era was to promote growth via redistribution, but in reality the scale of redistribution was insignificant and consequently we find that it was unable to make any appreciable impact upon poverty. As the 1950s policy was neither able to make any appreciable inroads in the reduction of poverty nor able to deliver high growth, this may suggest why the government decided to change its policy.

The next period is the green revolution, and the idea was that the states would sponsor technical progress, and that this progress would take care of the institutional barriers; therefore any barriers to the reduction of poverty would be automatically removed. Thus the issue of the trickle down effect does arise here.[5] But we find that the average poverty rate during this period declined by just 0.2%, which is even lower than the previous period, while the growth rate was marginally higher compared to the pre-green revolution period. This perhaps reconfirms all the early findings – poverty and inequality both rose at a faster rate in the green revolution belt, thereby slowing down the fall in the reduction of the incidence of poverty.[6] Interestingly, when we compare the relationship between capital formation and the rate of output growth, the relationship shows a very weak link. While capital formation grew by 1.2% per annum, more than that in its previous era, output grew merely by 0.2% more than in the previous period. This raises the question of what the nature of the capital formation was that had such a marginal impact upon the rate of growth of output? This, in turn led us to compare the relationship between capital formation and poverty. It shows that the number of people living below the poverty line declined at a slower rate per annum compared to the previous period, with a 1.2% rise in the rate of capital formation compared to the previous period. This was the period when large-scale mechanisation began, which was initially observed in Punjab and subsequently in many other parts of rural India (Rudra, 1992). The above findings therefore may explain why the government decided to change its policy yet again.

Let us consider our last period, where the government not only intervened with its employment-generating scheme, but also to allocate credit in favour of small and marginal farmers. Interestingly, we also observe that during this period the growth rate was much higher compared to any other previous periods, and capital formation was negative, implying that most of the government spending was in the form of the allocation of credit and inputs. In this period, the government concentrated more on the distribution aspect, and consequently we not only observe a higher growth rate, but also for the first time observe that poverty was declining in India at a much faster rate compared to the previous era.

An important point to note here is that when a longer time period is considered, it is not too difficult to find a declining trend in the incidence of poverty with the rise in the growth rate. But to conclude from this that the trickle down effect has taken place in India, is altogether a different matter. To prove that the trickle down effect took place, there is a need to show that at any point in time during this period either the government did not intervene regarding the distribution, or else poverty was falling consistently with the rise in the growth rate, irrespective of the government intervention. Our preliminary examination shows that it was the government redistribution policy in the late 1970s and 80s, which not only produced a higher growth rate but also reduced the incidence of poverty at a much faster rate. To reconfirm our preliminary findings, we need to undertake further rigorous empirical tests, to which we turn now.

3. EMPIRICAL TESTING OF THE POVERTY-GROWTH NEXUS

Re-confirmation of our preliminary findings depends upon whether in our empirical investigation we find the trickle down effect or not. With this in mind, we undertake further re-examination of the trickle down hypothesis, by investigating the impact of capital formation and GDP in agriculture on rural poverty using time series data from 1951 to 1991.

We can postulate from the analysis in the previous section that the rate of change in the incidence of poverty depends upon changes in capital formation and output in agriculture, or that these variables are dynamically related in a multivariate set up. In other words, we are investigating whether a change in the *CAPAG* causes a change in the *GDPAG*, which leads to a change in the incidence of poverty or vice-versa. While comparisons over a long time period between 1951 and 1991 suggest that the incidence of

Table 16.3 Correlation matrix

	POVRU	GDPAG	CAPAG
POVRU	1.00	− 0.65	− 0.45
GDPAG		1.00	0.90
CAPAG			1.00

Table 16.4 Pair-wise Granger Causality tests

Null Hypothesis:	Obs	F-Statistic	Probability
GDPAG does not Granger Cause POVRU	39	7.25644	0.00237
POVRU does not Granger Cause GDPAG			0.01987
CAPAG does not Granger Cause POVRU	39	2.47288	0.09938
POVRU does not Granger Cause CAPAG			0.65184
CAPAG does not Granger Cause GDPAG	39	1.63921	0.20911
GDPAG does not Granger Cause CAPAG			0.72801

poverty declined with economic growth, a sub-period analysis suggests that there is considerable variation in the rate of change in the incidence of poverty with a rise in the GDP, implying that the impact of the growth rate on the reduction in poverty may be negligible. An examination of the correlations with the stated period, presented in Table 16.3, reveals that the simple correlation of the level of poverty (or poverty index) with the *GDPAG* is –0.65, while the correlation of poverty with *CAPAG* is –0.45. Although the negative correlation seems to make sense from a theoretical standpoint, the results are not robust.

Consequently, we decided to undertake Granger Causality tests to assess the direction of the causality. Table 16.4 shows that there is a bi-directional causality between *GDPAG* and *POVRU*. This implies that income growth is a necessary pre-condition for the reduction in poverty, and also an increase in poverty may adversely affect economic growth.

The above bi-directional feedback result suggests that there is a need to undertake further tests in order to investigate whether there exists any long-term relationship between the variables, running from income growth to reduction in the incidence of poverty via higher capital formation. In this examination, first, we need to test the stationary properties of the time series involved. This is because many time series are non-stationary or integrated of order one, $I(1)$, implying the presence of a unit root. The presence of unit roots is tested for all three variables in logarithms using the augmented Dickey–Fuller (ADF) test. There is evidence

Table 16.5 *ADF unit root test results*

Variables	ADF in levels		ADF in 1st differences		
	Without trend	With trend	Without trend	With trend	~I( )
POVRU	−0.96	−2.62	−5.95**	−6.33**	I(1)
GDPAG	−0.26	−3.37	−7.01**	−6.99**	I(1)
CAPAG	−1.22	−1.15	−4.74**	−4.85**	I(1)

Notes: The ADF unit root test is given as follows. The aim is to test if H_0: $\alpha_2 = 0$ against H_0: $\alpha_2 \neq 0$ corresponding to y_t is integrated vs. y_t is not integrated, and the test is based on one lag.

$$\Delta y_t = \alpha_0 + \alpha_1 t + \alpha_2 y_{t-1} + \sum_{j=1}^{k} \beta_j \Delta y_{t-j} + u_t$$

Critical values are: 5%= −2.9399, 1%= −3.6117 (without trend)
5%= −3.5312, 1%= −4.2165 (with trend)
where y is the series under consideration, t is the time trend, α_0, α_1, α_2, and β_j are parameters, k is the number of lagged differences included to capture any autocorrelation, and u is the error term.

of a unit root in all the three series, which means they are all $I(1)$, or they are non-stationary. The results are presented in Table 16.5 below. When the variables contain such properties, running OLS regression will produce spurious results. However, Granger (1988) has argued, if the $I(1)$ series move together, they share a common stochastic trend and their linear combination is stationary, which indicates that they are cointegrated, implying the existence of a meaningful long-run equilibrium relationship.

As the Granger causality test is only valid in the context of any pair of variables, in order to undertake a multivariate cointegration test we need to use Johansen's cointegration technique (Johansen, 1988). As our variables are stationary in first differences (see Table 16.5) they support Johansen's VAR model in first differences. This technique tests the null hypothesis of no cointegration against the alternative of cointegration, and yields two likelihood ratio statistics for the number of cointegrating vectors, namely, the maximum eigen value and the trace statistics. The Johansen cointegration framework is used to examine the possible long-run relationships between agricultural output, capital formation in agriculture, and the rural poverty index. Table 16.6 shows the trace statistics (likelihood ratio) when testing for cointegration. The results show that there is a stable equilibrium relationship between poverty, *GDPAG* and investment in the long-run. The null hypothesis that $r=0$ is rejected means there exists a meaningful long-run relationship between these three variables.

Table 16.6 Testing for cointegration

No. of CEs	Eigenvalue	Likelihood ratio	95% Critical value
R=0	0.548960	34.33208*	29.68
R≤1	0.066964	3.280284	15.41
R≤2	0.014689	0.577129	3.76

Notes: R is the number of cointegrating equations; the test assumes linear deterministic trend in the data.
* L.R. test indicates 1 cointegrating equation at 5% significance level.

The estimated normalized cointegrating equation can be written as follows:

$$POVRU = 12.74 - 1.19 * GDPAG + 0.46 * CAPAG$$
$$(-7.888) \qquad (5.492)$$

The results suggest that rural poverty has been negatively affected by changes in the agricultural output, while capital formation has contributed to an increase in poverty. It is interesting to note that while 1% increase in agricultural output leads to 1.2% decline in rural poverty, capital formation increases poverty by 0.5% in the long run, *ceteris paribus*. Both these coefficients are highly significant.

As the results suggest the presence of a long-run relationship between these variables, the construction of an empirical model for examining the linkages among them will include an error-correction (*EC*) term. Therefore we will be deriving a short-run model from the above long-run relationship, which will allow us to assess the poverty response, following a shock to GDP or investment. The estimates of the poverty equation as part of the dynamic vector error-correction (VEC) model are as follows:

$$\Delta(POVRU)_t = -0.012 + 0.05\,\Delta(POVRU)_{t-1} + 0.41\,\Delta(GDPAG)_{t-1}$$
$$(-1.115) \; (0.384) \qquad (2.781)$$

$$-0.11\Delta(CAPAG)_{t-1} - 0.54EC_{t-1}$$
$$(-1.172) \qquad (-5.764)$$

where $R^2 = 0.502$; SE = 0.06; Sample (adjusted): 1953–1991; t-values in parentheses.

In the short run, an increase in agricultural output growth significantly increases poverty by 0.41%. But the coefficient of the error correction term is negative and significant, meaning that 54% of the deviations from the long-run equilibrium are being reversed in the following year. This indicates a very fast adjustment. It appears that such an adjustment can only take place provided some external body (government) intervened in order to reduce the poverty.

In order to reconfirm this result we undertook an impulse response analysis for poverty, along with the short-run equations for output growth and investment. This analysis allows us to further examine how an unanticipated shock to *GDPAG* would impact on the poverty rate. The anticipated shock analysis can be conducted using the estimated coefficients. In order to analyse the impact of unanticipated policy shocks (or innovations), Sims (1980) suggested employing the impulse response functions (*IRF*), which are obtained from the moving average representation of the VEC model. It has been argued that the distributed lag coefficients estimated using VEC do not provide a clear understanding of the implied dynamic behaviour of the model. The use of impulse response coefficients will enable us to analyse the dynamic behaviour of a variable when random shocks are given to other variables. Thus the IRF describes the effect of an innovation ('shock' of one standard error) in a given variable on the movement of itself or another variable in the system. The ordering of the variables in the VAR is important since the first variable is only affected contemporaneously by a shock to itself; the second is affected contemporaneously by shocks to the first and to itself, and so on.

Figure 16.2 reports the impulse responses to one-standard deviation shock in each of the variables in the system, which provides a better device to analyse the shocks. As expected, given the fact that we found the existence of cointegration, the impulse responses show that the short-run effects eventually decay back to a constant in the long run. The impulse responses exhibit the effects of shocks in terms of annual percentages. A 1% shock to *GDPAG* produces a negative impact on *POVRU* – declining

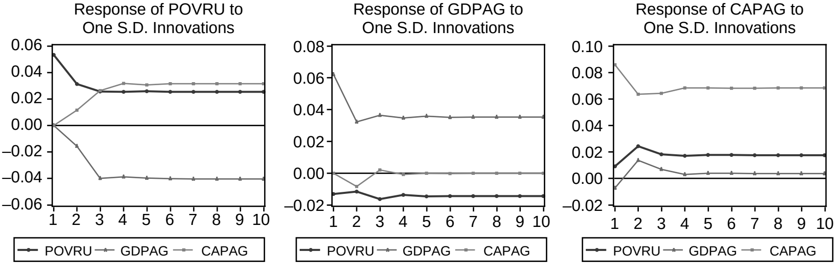


Figure 16.2 Short-run impulse responses

by 0.02% in the first year and 0.04% in the second year, after which the impact dies out eventually (first panel), and with a positive initial impact on *CAPAG* (third panel). The shock also affects the variable itself and the impact decays gradually, which is true for the other two variables as well. With regard to the effects of *CAPAG*, there is a positive response on *POVRU*, suggesting that it is the nature of capital formation that explains why poverty increases.

The above analysis also suggests that the reduction in poverty has a positive impact upon economic growth. The shocks to *POVRU* show that it reduces *GDPAG* (see second panel) and, as *CAPAG* rises following a shock to poverty (see third panel), it has a marginal positive impact on *GDPAG* subsequently.[7] All these impulse responses do reflect that the model is stable in the sense that they are smoothed out with a decaying response.

The above analysis suggests that when a longer period was considered, the argument that economic growth trickles down automatically to the poor has not been proven to be correct for India, as poverty appears to have increased following an increase in labour-saving capital formation, thereby preventing output growth from reducing poverty permanently. Therefore, the decline in the incidence of poverty from the late 1970s and the 1980s must have been the result of the government's anti-poverty measures. But the question that needs to be answered is why, in a labour-surplus economy, when the wage rate is low the growth rate is not accompanied by a growth in employment? It is to this issue we turn now.

4. AN EXPLANATION OF WHY THE TRICKLE DOWN EFFECT DID NOT TAKE PLACE IN INDIA

The issue of the trickle down effect principally emerged in the context of Indian agriculture during the period of the HYV programme (HYVP). The higher yield of the HYV and their shorter maturity opened up the possibility of multiple cropping, which offered to increase the rate of growth of output and also to open up the possibility for further employment growth. Therefore, it was reasonable to assume that the growth in agricultural output would reduce the price of food, and would increase the demand for labour, and these two effects in turn would directly work in favour of the poor. But this did not happen. The previous section reconfirms our point, made in the second section, that this was largely due to the increasing mechanisation during the green revolution period, which greatly concentrated not only on acquiring land-saving devices, but also on labour-saving devices, especially by large farmers. The puzzling pattern of this form of mechanisation does raise the question of why, in a labour-surplus economy, when it is unlikely that the growth of employment will be accompanied by a rise in the wage rate, would a landlord prefer to opt for labour-saving devices when the wage rate is so low? In order to understand this apparently peculiar behaviour by the landlords, we need to investigate the conditions under which the production system operates.

Prior to the introduction of the HYVP, Indian agriculture was predominantly cultivated by tenants and small farmers. Most of the literature on the rural economy concentrated on the distributional aspect of the economy, owing to the existence of high levels of poverty, and therefore the literature focused on the exploitative aspect of the relationship between landlords and tenants, whether the authors followed neoclassical (such as Mellor, 1968) or Marxian (such as Bhaduri, 1973) models. As a result, some of the subtle complexities under which the operational aspect of the production system was formed have been either neglected or not received much attention.

Two-thirds of the land area under cultivation used to depend directly upon rainfall, and as a result the return from agricultural investment always remained subject to the uncertain whims of nature. This uncertainty in relation to future income in the case of the rural sector primarily arises from four factors, two of which concern rainfall: the level of rainfall and the distribution of rainfall. Thus any cultivable area that is not under the control of an irrigation facility is more susceptible to crop failure if the rainfall is below normal. Even normal rainfall may have an adverse consequence upon the crop if the distribution of the rainfall affects the timing of the availability of water. These factors not only cause a variation in the yield per unit of land, but also cause a variation in labour input requirements. For example, if the crop production is below normal, especially due to the maldistribution of rain, less labour will be required during the harvesting period, but this cannot be predicted during the sowing period, implying that part of the labour force will remain idle. Similarly, if crop production is above normal, part of it may be lost in the absence of a reserve stock of labour to draw on at critical times. Also, heavy rainfall may cause an outbreak of fever during the sowing period. Thirdly, adverse financial circumstances may affect the health of the labour force during the harvesting period. Finally, the expectation of changes in the relative price ratios of competing crops introduces uncertainty in relation to the allocation of resources among different crops for individual farmers.

All of these factors amount to uncertainty in relation to future rates of return on investment in the agricultural sector. Landlords were aware that there is nothing they can do about the uncertainty that follows from the whims of nature. But the landlords recognised that they could reduce the uncertainty that otherwise would follow from the adverse financial circumstances of their labour force and from a possible lack of a steady supply of labour that might be required at critical times. Accordingly, they opened up credit facilities for their tenants and their labour force. It is this method of minimising uncertainty that resulted in the interlocking of the land, the labour and the credit market. Much of this credit market operates on the basis of a congruence of interest, where landlords know that in the absence of a credit facility, they would not be able to draw on labour at critical times.

Furthermore, if the health of the labour force becomes adversely affected due to malnutrition during the harvesting period, much of the harvest may be lost. As the labour force knows that it is the landlord who comes to rescue them on their rainy days, they feel obligated, and make themselves available to assist the landlords when the latter are in need, and this in turn also ensures the availability of these credit facilities. But this congruence of interest largely arose owing to the backward technology and poverty of the labour force.

Where there is backward technology, manpower is the most precious asset which can be bought, which means that in conjunction with wages, landlords had to offer other benefits in order to ensure a steady availability of labour. These benefits were not only non-accountable in wage bills, but some of them were not even accountable in monetary units, as it involved time and effort to develop and manage a relationship between landlords and tenants that ensured a steady flow of labour, which could be conceived of as quite costly (Basu, 1997). The uneasy equilibrium that had formed between the landlords and their tenants or labour force was not understood. Therefore, it was not recognised that the landlord would have the incentive to opt for technology which would not only reduce the uncertainty that follows from the whims of nature, but also the uncertainty that otherwise followed from the poorly nourished, and unsteadily available, labour force. This means landlords had an incentive to acquire both forms of devices: land-saving as well as labour-saving.

In addition to the above, the tenancy system in India principally emerged owing to the backward technology, which put a severe physical limitation on the number of acres of land that an individual farming family could cultivate. For example, in Punjab, a family of five labourers with a pair of bullocks can cultivate a maximum of 10 acres, beyond which it is physically not possible. As we move towards the eastern and southern belts of India, this physical capability reduces even further among farmers, and as a result a family of five would not be able to cultivate even 10 acres. For the sake of simplicity we assume that the average farming family at best can cultivate 10 acres of land. This means that any farmer whose landholding size is greater than 10 acres will be required to hire tenant farmers. Under the tenancy arrangements, normally the share of the harvest is divided equally between the landlords and their tenants. Now imagine a landlord whose holding size is 25 acres, which is not uncommon in the state of Punjab; in the absence of labour-saving devices, he/she had to lease out 15 acres of land. This means that, under the tenancy arrangement, landlords were losing harvest, equivalent to the output of 7.5 acres of land in every harvesting season. Given this scenario, with the availability of tractors, a landlord knows that his family labourers can now cultivate not only 25 acres, but even more, thereby preserving the entire harvest for their own family. This provides them with an additional incentive to opt for labour-saving devices; by opting for labour-saving devices landlords can change the tenancy arrangement and have much to gain from it.

Those who adopted the HYVP, who happened to be rich and middle class farmers, received most of the benefits from institutional credit agencies.[8] Initially, these farmers obtained loans at a cheaper rate, and with reduced collateral requirements for the installation of irrigational facilities and the purchase of inputs. The installation of irrigational facilities is a vital land-saving device, and it reduces the uncertainty that arises from the less than normal rainfall and the maldistribution of rain. This means landlords were not required to make provision for a steady flow of labour at critical times, and could plan their total requirements in advance. In addition to this, landlords recognised that if they extended HYV cultivation into that land which was cultivated by their tenants, they could benefit also from the increase in output per unit of land. However, under tenancy agreements, although the tenants' percentage share of the harvest might remain the same, in terms of the absolute amount there would be an increase, while landlords would bear the major input costs. Some of the landlords decided to offer these inputs as a loan, which in turn enabled them to take a greater share of the increased absolute amount of the harvest in the form of interest, which otherwise would have been available to the tenants. Other landlords, especially the large ones, recognised that, given the supply of labour, if they could change the tenancy arrangements they would not have to go through this elaborate system, and perhaps could take the entire gain that followed from this HYVP. This meant that now landlords would be cultivating with hired labour, and as a result, the landlords' operational holdings would increase, given the ownership size of the land. In the past, landlords' operational holdings used to be much smaller compared to their ownership holdings, while in the case of tenants it was the opposite.[9] But to change this arrangement, landlords required further mechanisation.

Although the initial allocation of loans was mostly invested in devices that could be categorised as land-saving, this caused the value of their land to increase, compared to that of those whose land size was not sufficient to make the installation of irrigation facilities economically viable.[10] In the process, this further increased their access to the loan market. The existence of cheaper rates reduced the repayment rate on loans, and this meant borrowers could borrow even larger amounts than was possible at a higher rate. Furthermore, as the government reduced the collateral requirements under the priority sector loan arrangements, they did not have to offer collateral of a higher value, as required under normal banking operations.

As a result, these borrowers received more loans of a considerably larger size than was possible in the absence of this policy, and subsequently they used these loans to purchase machinery categorised as labour-saving, such as tractors, harvesting and threshing machines. This opened up the opportunity for landlords to change the tenancy arrangements, but instead of increasing, reduced the scope of employment in the rural areas.[11] Consequently, the green revolution, instead of reducing the incidence of poverty, caused it to escalate in the rural areas, and the government had to intervene to directly attack poverty.

5. CONCLUSION

The analysis and econometric tests that have been undertaken in this paper suggest that there is little or no evidence to claim that the trickle down effect has occurred in India. Growth can only directly address the issue of poverty provided it is also accompanied by growth in employment. The green revolution technically offered a unique opportunity to address the issue of poverty owing to its multiple cropping possibilities. The higher yield of HYV and its shorter maturity period opened up scope for multiple cropping, which in turn not only offered a higher rate of growth in output but also a greater rate of employment. Thus it was anticipated that this in turn would reduce poverty, and consequently the claim was made that the technical progress sponsored by the state would automatically eliminate institutional impediments.

But this anticipation was based on the assumption that the existence of surplus labour is unlikely to put much pressure on raising wage rates, thereby ruling out the possibility that the growth in employment would bring capital–labour substitution. But some crucial factors were overlooked, and particularly that tenancy cultivation was a predominating feature of Indian agriculture while landless labourers constituted only 15 to 17% of the total labour force, and therefore wage rates were not a major part of the equation; rather it was the tenancy system itself. It was the backward technology and the uncertain conditions under which cultivation took place, arising from the whims of nature and the poverty of the rural workforce, which to some extent tied landlords to the rural workforce, that is the tenants. Landlords not only had a share in the good fortune of an abundant harvest but also often had to share disproportionately the fate of the poor harvest, as they had to feed their tenants. Therefore they always had the incentive to opt for technology that would free them from the fate of their tenants. This is irrespective of the adoption of the HYVP. The HYVP made it easier for landlords to adopt both technologies, that is, land-saving as well as labour-saving devices.

It is the investment in labour-saving devices that prevented the emergence of the trickle down effect. This suggests that, without investigating the system, perhaps it is not wise to make assumptions, even for a poorer country, based merely on observations of the existence of surplus labour that capital–labour substitution is unlikely to emerge.

Poverty that was declining with the higher growth rate during the late 1970s and throughout the 1980s was largely the result of a variety of government measures, including both direct anti-poverty measures and the adoption of a more egalitarian distribution of credit and inputs to smaller and marginal farmers. But the problem with these measures is that they do not form a permanent platform on the basis of which one can attempt to eradicate poverty. As we have observed, during 1991, following the financial reform, when the government attempted to address macroeconomic stability, this meant cutting expenditure on anti-poverty programmes and the removal of various subsidies that were given to agriculture, and poverty was on the rise again and the government had to intervene immediately. The sorry state of our knowledge is that we have yet to find a platform upon which to formulate a policy that will permanently improve the conditions of poor people. In an unequal society, growth alone does not provide such a platform.

NOTES

1. For earlier work on this issue see Ravallion and Datt (2002), Datt and Ravallion (1998), Mellor (1999), Fan *et al.* (2000) and World Bank (1995). These authors, including the World Bank (1995), in general claim that the trickle down effect does take place. Fishlow (1995) and Deaton (2001) have reservations about their claim, while Besley and Burgess (2000), Rao (1994) and Tendulkar (1998) argue that falling poverty in India in recent years has been largely due to government intervention.
2. See Deaton (2001) and Deaton and Dreze (2002) for more on this issue.
3. For further details on this subject see Desai (1988) and Haque and Verma (1988).
4. For further details on these issues see Rao (1994), Rath (1985), Dantawala (1985) and Hirway (1985).
5. It was argued that the technical progress stimulated by the state would itself eliminate the institutional impediments to progress, thus meaning the adoption of the HYV

programme would simultaneously solve the problems of food shortages and poverty. See Cummings and Ray (1969) for more on this issue.

6. See Bardhan (1974), Saini (1976), (Rudra, 1969), Ladejinsky (1969) and Basu (1982) for more on this issue.
7. The positive impact of capital formation on total output in agriculture can emerge only to the extent such capital formation takes place in areas, such as investment in irrigation, that is usually undertaken by the public sector; see Mallick (1993).
8. See Dasgupta (1976) and Basu (1982) for further on this issue.
9. See Rudra (1971), Dasgupta (1976) and Vyas (1976) for more on this issue.
10. For further details on this issue see Dasgupta (1976).
11. For further discussion on this issue see Bhalla (1987), Rudra (1992) and Rao (1994). For early studies on this issue see Rudra (1971), Binswanger (1978), Rao (1979) and Laxminarayan (1982).

REFERENCES

Bardhan, P.K. (1974), 'Inequality of Farm Income: A Study of Four Districts', *Economic and Political Weekly*, February.

Basu, S. (1982), *An Analysis of the High Yielding Variety Programme in India*, MEc dissertation, Sydney, Sydney University.

Basu, S. (1997), 'Why Institutional Credit Agencies are Reluctant to Lend to the Rural Poor: A Theoretical Analysis of the Indian Rural Credit Market', *World Development*, **25**, 267–80.

Besley, T. and R. Burgess (2000), 'Land Reform, Poverty Reduction, and Growth: Evidence from India', *Quarterly Journal of Economics*, **115** (2), 389–430.

Bhaduri, A. (1973), 'A Study in Agricultural Backwardness under Semi-Feudalism', *Economic Journal*, **83**, 120–37.

Bhalla, S. (1987), 'Trends in Employment in Indian Agriculture, Land and Asset Distribution', *Indian Journal of Agricultural Economics*, **42** (4), October–December.

Binswanger, H.P. (1978), *Economics of Tractors in South Asia: An Analytical Review*, Agricultural Development Council, New York, and International Crops Research Institute for the Semi-Arid Tropics, Hydrabad.

Cummings, R.W. and S.K. Ray (1969), 'The New Agricultural Strategy', *Economic and Political Weekly*, March, 29.

Dantawala, M.L. (1985), 'Garibi Hatao: Strategy Options', *Economic and Political Weekly*, March, 16.

Dasgupta, B. (1976), *Agrarian Change and the New Technology in India*, Geneva, UNRISD.

Datt, G. and M. Ravallion (1998), 'Why Have Some Indian States Done Better than Others at Reducing Rural Poverty?', *Economica*, **65**, 17–48.

Deaton, A. (2001), 'Counting the World's Poor: Problems and Possible Solutions', *World Bank Research Observer*, **16** (2), 125–47.

Deaton, A. and J. Dreze (2002), 'Poverty and Inequality in India: A Re-Examination', *Economic and Political Weekly*, September, 7.

Desai, D.K. (1988), 'Institutional Credit Requirements for Agricultural Production', *Indian Journal of Agricultural Economics*, **43** (3), July–September.

Dreze, J. and A. Sen (1995), *India: Economic Development and Social Opportunity*, Delhi, Oxford University Press.

Fan, S., P. Hazell and S. Thorat (2000), ‘Government Spending, Growth and Poverty in Rural India’, *American Journal of Agricultural Economics*, **82** (4), 1038–51.

Fishlow, A. (1995), ‘Inequality, Poverty and Growth: Where Do We Stand?’, in M. Bruno and B. Pleskovic (eds), *Annual World Bank Conference on Development Economics 1995*, Washington, DC, World Bank, pp. 25–37.

Granger, C.W.J. (1988), ‘Some Recent Developments in a Concept of Causality’, *Journal of Econometrics*, **39**, 199–211.

Haque, T. and S. Verma, (1988), ‘Regional and Class Disparities in the Flow of Agricultural Credit in India’, *Indian Journal of Agricultural Economics*, **43** (3), July–September.

Hirway, I. (1985), ‘Garibi Hatao: Can IRDP do it?’, *Economic and Political Weekly*, March, 30.

Johansen, S. (1988), ‘Statistical Analysis of Cointegrating Vectors’, *Journal of Economic Dynamics and Control*, **12**, 231–54.

Ladejinsky, W. (1969), ‘How Green is the Green Revolution?’, *Economic and Political Weekly*, **14** (39), September.

Laxminarayan, H. (1982), ‘The Impact of Agricultural Development on Employment: A Case Study of Punjab’, *The Developing Economies*, **20** (1).

Mallick, S. (1993), ‘Capital Formation in Indian Agriculture: Recent Trends’, *Indian Journal of Agricultural Economics*, **48** (4), October–December, 667–77.

Mehta, A.K. and A. Shah (2003), ‘Chronic Poverty in India: Incidence, Causes and Policies’, *World Development*, **31** (3), 491–511.

Mellor, J.W. (1968), ‘The Evolution of Rural Development Policy’, in J. Mellor, J. Weaver, U. Lele and M. Simon (eds), *Developing Rural India*, Bombay, Lalvani.

Mellor, J.W. (1999), *Pro-poor Growth: The Relation between Growth in Agriculture and Poverty Reduction*, Bethseda, Abt Associates Inc.

Rao, C.H.H. (1979), ‘Farm Mechanisation’, in C.H. Shah (ed.), *Agricultural Development of India: Policy and Problem*, Delhi, Orient Longman.

Rao, C.H.H. (1994), *Agricultural Growth, Rural Poverty and Environmental Degradation in India*, Delhi, Oxford University Press.

Rath, N. (1985), ‘Garibi Hatao: Can IRDP do it?’, *Economic and Political Weekly*, February, 9.

Ravallion, M. and G. Datt (2002), ‘Why has economic growth been more pro-poor in some states of India than others?’, *Journal of Development Economics*, **68** (2), 381–400.

Rudra, A. (1969), ‘Big Farmers of Punjab: Second Installment of Results’, *Economic and Political Weekly*, **4** (52).

Rudra, A. (1971), ‘Employment Patterns in Large Farms of Punjab’, *Economic and Political Weekly*, Review of Agriculture, June.

Rudra, A. (1992), *Political Economy of Indian Agriculture*, Calcutta, K.P. Bagchi & Company.

Sims, C.A. (1980), ‘Macroeconomics and Reality’, *Econometrica*, **48** (1), 1–49.

Saini, G.R. (1976), ‘Green Revolution and the Distribution of Farm Incomes’, *Economic and Political Weekly*, Review of Agriculture, March, 27.

Tendulkar, S. (1998), ‘Indian Economic Policy Reform and Poverty: An Assessment’, in I.J. Ahluwalia and I.M.D. Little (eds), *India’s Economic Reforms and Development: Essays for Manmohan Singh*, Delhi, Oxford University Press.

Vyas, V.S. (1976), ‘Structural Change in Agriculture and the Small Farm Sector’, *Economic and Political Weekly*, January, 10.

World Bank (1995), ‘The Social Impact of Adjustment Operations: An Overview’, Report No. 14/76, Operations Evaluation Department, Washington, DC, 30 June.

17. Strategy for economic growth in Brazil: a Post Keynesian approach

José L. Oreiro and Luiz Fernando de Paula*

1. INTRODUCTION

This chapter proposes a Keynesian strategy for economic policy that aims to achieve higher, stable and sustained economic growth in Brazil. Its main hypothesis is that the current poor growth performance of the brazilian economy is due to macroeconomic and structural constraints rather than the lack of microeconomic reforms (labour market, credit market, and so on), as liberal economists in Brazil have suggested.

The chapter is divided into four main sections, besides this introduction. The second section briefly discusses the main features of a new economic strategy (based on demand-side and supply-side policies) that aims to overcome the constraints on sustained economic growth. The third section discusses the current economic constraints on sustained economic growth in Brazil. In the fourth section, a simple version of the Harrod–Domar growth model is utilized in order to obtain the potential growth rate of the brazilian economy. Finally, the fifth section presents a new economic policy model for the brazilian economy, designed to achieve its potential growth rate. This policy should include both demand-side, and supply-side policies.

2. KEYNESIAN ECONOMIC POLICIES: A BRIEF VIEW

2.1. Definition of Keynesian Economic Policy

Contrary to orthodox economics, for which activist economic policies have no permanent effect on the real variables, such as employment and product,

* The authors would like to thank Phillip Arestis and G.C. Harcourt for their helpful comments. Financial support of CNPq, PRONEX (FAPERJ/CNPq) and Fundação Araucária is gratefully acknowledged.

Keynesian policies, in a broader sense, have as their main objective the achievement of full employment. In this connection, the meaning of Keynesian policy that we will adopt in this chapter is that in which 'policy implications arise from the perception of the role of aggregate demand in setting the level of economic activity and the lack of automatic forces in leading a market economy to full employment' (Arestis and Sawyer, 1998, p. 181). According to this view, a *laissez-faire* market economy normally exhibits elements of instability and, most of all, does not create a level of aggregate demand consistent with full employment. As a result, in monetary economies, full employment can only be achieved by accident or through state policies.

Based on the concept of non-neutrality of money and on the principle of effective demand, economic policy – according to the Post Keynesian approach – is able to affect the real variables of the economy both in the short- and long-run. Keynesian policy is related to the implementation of economic policies that intend to increase aggregate demand in order to create a stable environment that stimulates entrepreneurs to make *new* investments. Indeed, employment levels and the utilization of productive capacity depend crucially on the determinants of aggregate demand, particularly the entrepreneurs' investment decisions. In other words, economic policy should affect aggregate private investment, as it can create a safe environment that stimulates private agents to make more risky choices than just accumulating liquid assets. So, a 'good' policy happens when economic agents are stimulated to invest in capital assets. The sphere of government's action should not, however, overlap with the private sphere; indeed, it should help to create a stable and safe environment for private agents to act.

One should note that the objective of the economic policy in this approach is related to *macroeconomic stability*, a broader concept than just *price stabilization*, as it aims to reduce the uncertainties that are intrinsic to the business world. Government can reduce macroeconomic risks that affect the economy as a whole. Price stability and higher level of product and employment can be, under certain conditions, compatible; for this purpose, government should make use of broader tools of economic rather than just monetary policy. In order to reach multiple policy objectives – such as economic growth and price stabilization – it is necessary to have a greater co-ordination of macroeconomic policies (fiscal, monetary, exchange rate, and income policies). Government should evaluate the global impacts of the policies on their objectives as a whole; that is, Keynesian policies consist of concerted actions in a multiplicity of arenas. In this context, policy co-ordination is essential in order to achieve macroeconomic stability.

2.2. Constraints on Economic Growth[1]

There are many constraints, both from demand-side and from the supply-side, on the achievement of sustained economic growth. This objective on a long-term basis requires that those constraints are somehow sufficiently eased.

Aggregate demand constraint

As we have already stressed, a *laissez-faire* market economy does not create a level of aggregate demand consistent with full employment. According to the effective demand principle, the level of output and employment in an economy is determined primarily by the demand for goods. Low economic growth and high unemployment results from the lack of effective demand; such demand is determined by entrepreneurs' expectations of future demand, as they decide during each period of production what they are going to produce and how many people they are going to employ. In other words, the volume of expenditure determines the aggregate demand of an economy, while the level of employment depends on the agents' expected expenditure.[2] In sum, according to the Post Keynesian approach, there is a lack of automatic forces within a market economy working to ensure that the level of aggregate demand is compatible with the full employment of labour and the existing capital stock.

Inflation constraint

Inflationary pressures usually emanate from the real side of the economy. Indeed, the process of moving towards sustained economic growth always involves a fall in the unemployment rate, and most of the time rising productive capacity utilisation, which are likely to generate inflationary pressures and a climate of inflationary expectations. The spread of inflation pressures depends on the degree of monopoly of firms, which can allow them to increase the mark-up of prices relative to costs, and the degree of workers' organization, as every increase in money-wage rates not offset by productivity improvements raises production costs. Particularly, if unemployment rates shrink a great deal, it is easier for workers to obtain more liberal wage increases.

Post Keynesian economists agree that inflation is a symptom of a fight over the distribution of current income, as it is the result of attempts to alter the existing distribution of money income among economic agents of the same region, and/or interregionally, and/or internationally. In the Post Keynesian view, there are many and different causes for inflation, and consequently there are various types of inflation; for each type of inflation, a specific anti-inflationary tool should be used. For instance, *spot* or *commodity price inflation*, which occurs whenever there is a sudden and unforeseen change in demand or available supply for immediate delivery, can be avoided 'if there is some institution that is not motivated by self-interest but which will maintain a "buffer stock" to prevent unforeseen changes from inducing wild spot price movements. A buffer stock is nothing more than some commodity shelf inventory that can be moved into and out of the spot market to buffer the market from disruptions of offsetting the unforeseen changes in spot demand and supply' (Davidson, 1994, p. 158).

Balance of trade constraint

The balance of trade constraint arises when the level of economic activity is constrained to ensure that the level of imports is compatible with the level of exports, as any difference between imports and exports should be covered by borrowing from overseas, which in the long run can increase the external vulnerability of an economy (Paula and Alves, 2000, p. 597).

Developing countries particularly can face a structural problem in their balance of payments, due to the effect of what is known as Thirlwall's law.[3] This law states a link between rate of economic growth and the income-elasticity of imports and exports of an economy; it states that in the long run, demand-side variables play a key role in economic growth through the 'balance of trade constraint': a country cannot grow at a rate higher than what is consistent with its balance of trade equilibrium. The low income-elasticity of products of smaller aggregate value exported by developing countries vis-à-vis the greater income-elasticity of products imported from developed countries can generate structural deficits in the balance of payments of the former countries. These increasing deficits can result in a significant constraint for economic growth in developing countries, as the maintenance of a non-exploding deficit requires that the domestic growth rate is maintained below the world growth rate so that imports and exports grow in line with one another.

Capital account constraint

The capital account constraint arises when an economy is vulnerable to the changes in the liquidity conditions and/or changes in the mood of global players in the international financial market, no matter the reason. Indeed, as the experiences of the 1990s currency crises showed all around the world, under a context of high capital mobility, such crises can occur for reasons not related directly to deficits in the current account's balance of payments. In other words, economies with small (if any) current account deficit (over GDP) – a situation in which a country is seen as solvent from the balance of payments' point of view – can face a sudden stop in the capital inflows due to a shift in the international investors' expectations. Sunspots, herding behaviour or contagion effect can induce this shift.

Countries with (i) much larger and volatile capital flows in relation to the size of their domestic capital markets and economies; (ii) non-convertible currency; and (iii) low levels of international reserves, are generally more prone to face capital account constraints. In such countries, volatile capital flows can generate very high volatility on exchange rates. Indeed, there are many economic issues related to excessive volatility of exchange rates, particularly related to the management of exchange rate risk and macroeconomic policy (determination of interest rate and public debt).

Lack of capacity

Lack of capacity can constrain economic growth in the long run in two scenarios. During the upturn, high economic growth can fulfil the full productive capacity of an economy, a phenomenon that can result in inflation pressures, as we have already stressed. On the other hand, after a period of prolonged low growth, the volume of the capital stock may fall to less of what would be required to sustain economic growth, due to the uncertainty about the future, generating a low level of 'animal spirits' that affects entrepreneurs' investment decisions. Under these conditions, entrepreneurs' expectations should be stimulated (in their decisions related to fixed investments) by demand-side economic policies.

2.3. Keynesian Economic Policies

Post Keynesian policies, in order to overcome the constraints on full employment, put emphasis on the need of both demand-side and supply-side policies. However, aggregate demand and aggregate supply are not independent, as the current level of demand has direct effect on the future supply potential of the economy, that is in both investment and productive capacity.

Fiscal policy can have a strong impact on the level of economic activity, as it is a powerful tool to stimulate aggregate demand, triggering a multiplier effect on private income. Fiscal policy should be used to push the economy toward full employment, as it affects directly private income, and agents' expectations concerning the future, igniting their optimism. For this purpose, Keynes recommended public expenditure or investment rather than increasing consumption, because of its stronger multiplier effect. Public investment can create a safe environment that can stimulate investment on fixed capital.

Using as a starting point the distinction made by Keynes (1980) between *ordinary budget* (related to ordinary functions of public administration) and *capital budget*, the former should be balanced at all times or even in surplus (which would be transferred to capital budget), while the latter one could be transitorily unbalanced, although it should be balanced over the long run; that is, it should be adjusted according to the fluctuations of the level of aggregate demand. In other words, the capital budget must be operated in a contra-cyclical way, preventing high fluctuations in private investments through the implementation of a long-term fiscal stabilization programme. The pace of public investments should be set according to the need of sustaining aggregate demand, serving to offset exogenous cyclical changes in investment spending (Kregel, 1994–95, pp. 265–6).

Monetary policy operated by the management of the interest rates can also have a significant impact on the level of economic activity. The management of interest rates can be used in order to influence the private agents' portfolio in favour of both increases of production (using current productive capacity) and the acquisition of capital goods. The management of monetary policy can be used to provoke a shift in the relative prices of different assets, from the more liquid to the more illiquid assets, that is, leading changes in the portfolio decisions that can affect real variables of the economy (product and employment). Monetary policy acts through the anticipation of expected movements of the rates of interest (Carvalho, 1997, p. 45).

Monetary policy should give clear signals of central bank purposes for the private agents in order to incite them to act according to the objectives of the policy-makers. Clearer policy signals can leave private agents more confident and safe to act. Contrary to what became accepted by orthodox economists, Keynes and Post Keynesian economists defend openness, not secrecy, as a condition for monetary policy to be effective.

In global financial markets, financial market prices – including exchange rates – have been excessively volatile, as they fluctuate according to fads and fashions. Indeed, there is an extensive empirical literature which indicates that excessive volatility in exchange rates affect negatively some real variables, such as investment and output.[4] Aiming at achieving a stabilizing economic policy for sustained economic growth, national governments should adopt an exchange rate policy that aims to prevent excessive volatility in exchange rates. The greater degree of stability of exchange rates would encourage entrepreneurs to engage more freely in international production, investment and trading transactions. This suggests an adjustable peg system with arrangements to avoid high volatility in exchange rates, such as accumulation of foreign reserves so that central banks can make use of dirty floats, the use of capital controls by some developing countries, and so on. Furthermore, institutional and regulations – such as some sort of capital controls, financial supervision, etc. – can be required to ensure that the fragility of the financial system does not spill over into instability within the productive economy.

Supply-side policies have to deal with three sorts of issues: the problem of inflation, imbalance in the overseas current account, and the organisation of work.[5]

As we have already stressed, orthodox stabilization policy is only efficient in the maintenance of a sufficiently high unemployment rate; furthermore, in most cases, it attacks the symptom but not the cause of the inflation; that is, such a policy does not solve the problem related to the increase of production costs. Therefore, Post Keynesians suggest some kind of incomes policy as part of the required arsenal in a market economy. Incomes policy requires, however, the generation of some sort of consensus over the distribution of income among the economic agents (government, entrepreneurs and workers). If money-wage rates and gross margins could be somehow controlled, price levels would decline. For this purpose, some degree of centralisation and coordination of pay setting would be required. Furthermore, the success of an economic policy oriented towards the objectives of macroeconomic stabilisation, as we have defined above, can also contribute to price stabilisation. For instance, if economic policy succeeded in reducing the volatility of exchange rate and interest rate, the more stable macroeconomic environment would have positive effect on both economic growth (as investment decisions are stimulated by business environment and macroeconomic policy) and price stabilisation.

The requirement for a broad balance on the overseas current account at full employment implies the need to overcome somehow the structural problems of balance of payments that some countries face (mainly those that are producers of commodities), as increasing deficits can result in a significant constraint to economic growth, according to Thirlwall's law. To overcome the balance of trade constraint, public policies (mainly an industrial policy) should be adopted in order to create conditions for a country to decrease the income-elasticity of demand for imports and to increase income-elasticity of demand for exports. These efforts should involve the development of an ability to compete in a range of high technology sectors, and/or the technological improvement of some current industrial sectors, which in turn involves both investment in research and development and the formation of linkages between companies to develop the whole production system.

3. CONSTRAINTS FOR ECONOMIC GROWTH IN BRAZIL

The period following implementation of the stabilisation plan known as the 'Real Plan' – that is, from July 1994 onwards – was striking for a remarkable reduction in inflation, even after the major devaluation of January 1999. After two years of economic growth (1994–95) resulting from the initial effects of this stabilisation plan, based on an exchange rate anchor, GDP evolution disappointed previous expectations of sustainable economic growth after price stabilisation. Furthermore, the trend took a 'stop-go' pattern and, as a result, the formal unemployment rate has remained above 10% since 1997 (Table 17.1).

In fact, the brazilian economy has suffered the impact of a succession of crises: Mexico in 1995, Asian countries in 1997, Russia in 1998, its own crisis in late 1998 and early 1999[6] and, more recently, crises in Argentina since late 2001 and again external crises in Brazil in 2002–03. A wide range of factors have contributed to shaping a very unstable macroeconomic context: the perception of external vulnerability deriving from both the still-worrying levels of external indicators – although these indicators improved a great deal in 2004 due to the increase of exports and GDP – and the liberalisation of capital account;[7] semi-stagnation in the economy that has inhibited productive investments; and the central bank's adoption of very high short-term interest rates and the consequent growth in public debt (Table 17.1). Brazil's current macroeconomic constraints stem mainly from the period when an exchange rate anchor was adopted (1994–99) in a context of trade and capital account liberalisations that generated a notable degree of external fragility of the economy and consequently some serious macroeconomic imbalances (for instance, high foreign debt, rapidly growing internal public debt, and so on). Private sector expectations have dropped under the impacts of various external shocks, the weak performance of the brazilian economy, and the very high rates of interest. As a result, the rate of investment has reached levels (around 18–21% of GDP since the early 1990s) far below the 1970s ones when investment rate was around 21–23%.

The 1999 switch from an exchange anchor to a floating exchange rate regime plus an inflation target regime brought no significant improvement in the macroeconomic variables (see Table 17.1), although balance of payments have improved their accounts in 2003–04, due mainly to the increase in the trade balance surplus. One might have expected that adopting a floating exchange regime would ease down the interest rate more quickly in Brazil. Although the rate of interest did decline, it picked up again during 2001, in view of the turbulence on international markets (the Argentina crisis, the effects of 11 September 2001, and so on), and again in 2003 due to the market turbulence at the beginning of Lula da Silva's government.

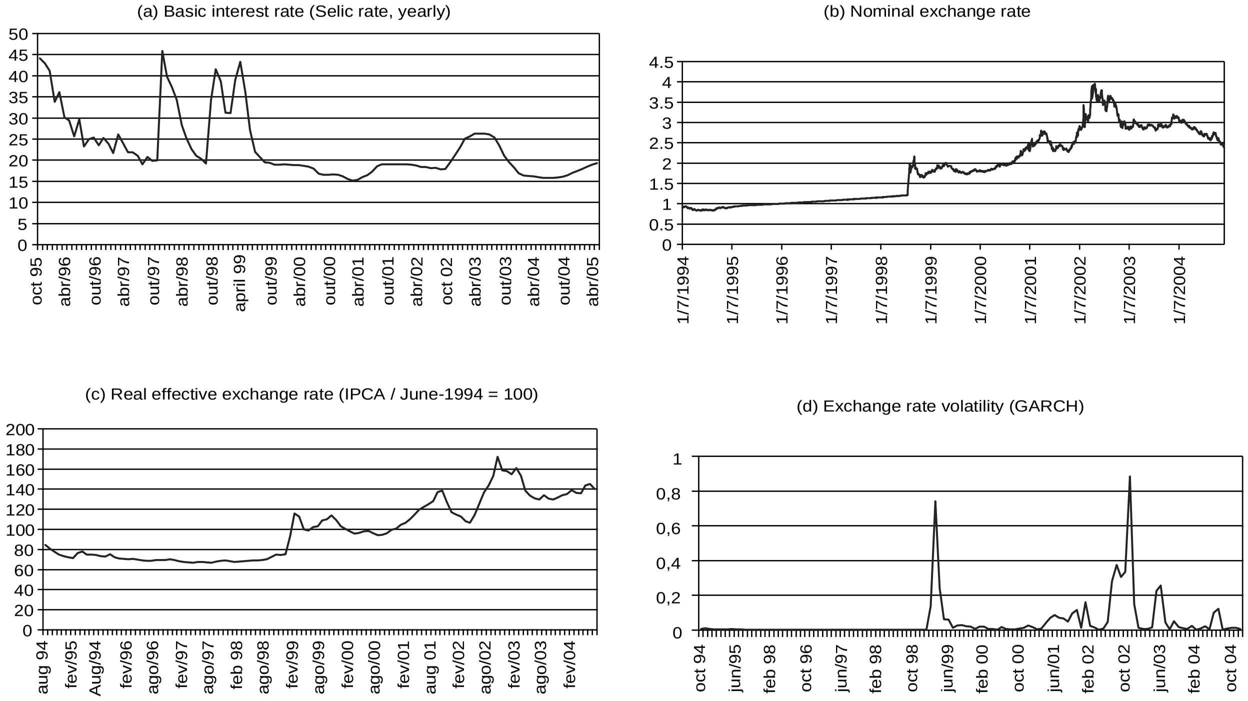
Indeed, the *modus operandi* of the inflation targeting regime, plus the adoption of a floating exchange rate regime under the conditions of high external debt and full opening of the capital account, has resulted in sharp

Table 17.1 Brazil – some macroeconomic data, 1991 to 2004

Year	Consumer price index (1PCA)	GDP growth – annual %	Investment rate (percentage of GDP)	Trade balance – US$ million	Current account – US$ million	Net Public debt-over-GDP	Real average income – Sao Paulo urban region (1985 = 100)	Formal unemployment rate* – Sao Paulo urban region (%)
1991	1,621.00	1.03	18.11	10,580	− 1,408	38.1	58.5	6.7
1992	472.7	− 0.54	18.42	15,239	6,109	37.1	61.3	8.0
1993	1,119.10	4.92	19.28	13,299	− 676	32.6	68.4	7.6
1994	2,477.10	5.85	20.75	10,467	− 1,811	30.0	65.9	7.8
1995	916.5	4.22	20.54	− 3,466	− 18,384	30.6	69.9	8.7
1996	22.4	2.66	19.26	− 5,599	− 23,502	33.3	71.5	9.2
1997	9.6	3.27	19.86	− 6,753	− 30,452	34.4	72.4	10.2
1998	5.2	0.13	19.69	− 6,575	− 33,416	41.7	71.5	10.8
1999	1.7	0.79	18.90	− 1,199	− 25,335	48.7	65.9	10.5
2000	8.9	4.36	19.29	− 698	− 24,225	48.8	62.3	10.0
2001	6.0	1.31	19.47	2,651	− 23,215	52.6	56.9	11.6
2002	7.7	1.93	18.32	13,121	− 7,637	55.5	51.6	11.4
2003	12.5	0.54	17.78	24,794	4,177	57.2	53.5	12.0
2004	9.3	5.18	19.58	33,693	11,669	51.8	52.3	10.0

Note: (*) Formal unemployment rate does not include informal unemployment.

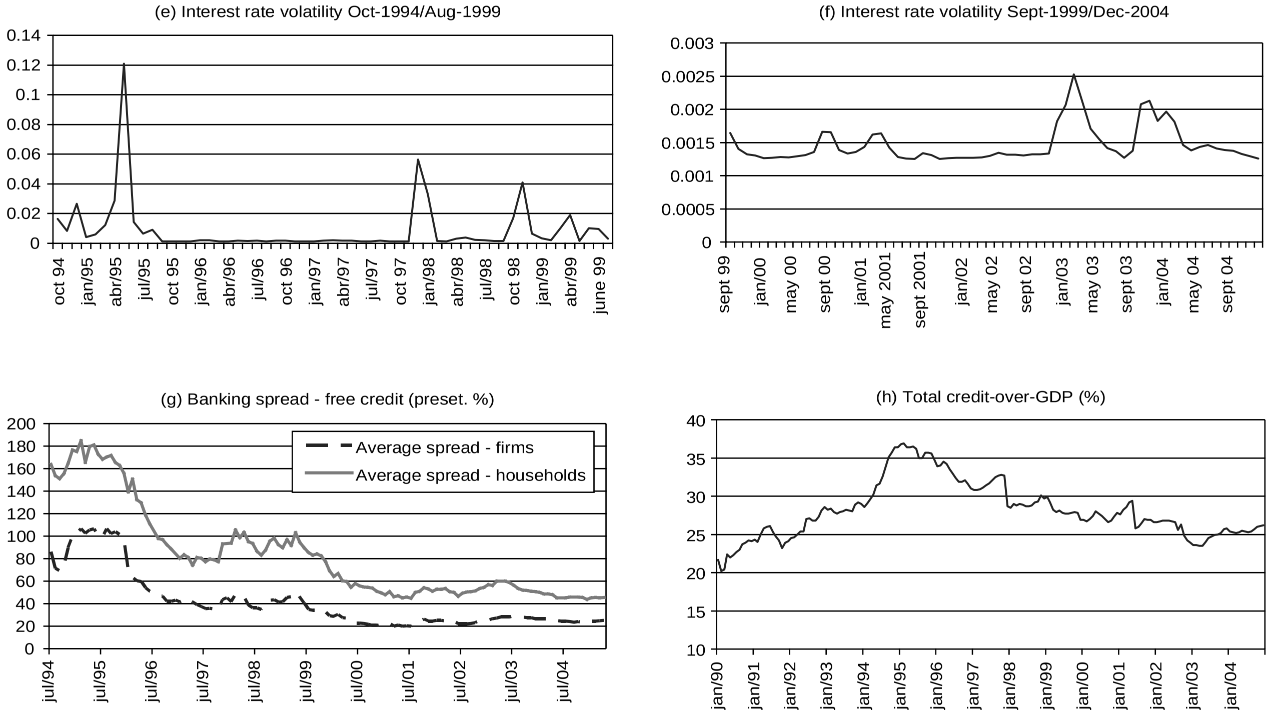
Source: IPEADATA.


(a) Basic interest rate (Selic rate, yearly)

(b) Nominal exchange rate

(c) Real effective exchange rate (IPCA / June-1994 = 100)

(d) Exchange rate volatility (GARCH)



(e) Interest rate volatility Oct-1994/Aug-1999

(f) Interest rate volatility Sept-1999/Dec-2004

(g) Banking spread - free credit (preset. %)

(h) Total credit-over-GDP (%)

Source: IPEADATA and Central Bank of Brazil. Figure 4 authors' calculations.

Figure 17.1 Some data for Brazil (1994–2004)

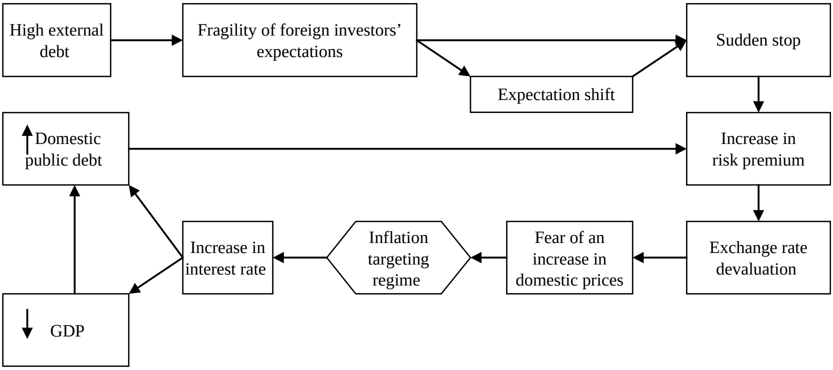

High external debt
Fragility of foreign investors' expectations
Sudden stop
Expectation shift
↑Domestic public debt
Increase in risk premium
Increase in interest rate
Inflation targeting regime
Fear of an increase in domestic prices
Exchange rate devaluation
↓ GDP

Figure 17.2 Macroeconomic framework of developing countries with high external indebtedness

instability of nominal exchange rate (Figures 17.1(a) and 17.1(d)). Capital outflows can induce a sharp exchange rate devaluation that can affect domestic prices ('pass through effect'), which in turn can jeopardize the Central Bank's inflation target. Under these conditions, the Central Bank is compelled to increase interest rates in order to seek to avoid both capital outflow and the pass through effect as it affects the aggregate demand. The Central Bank's reaction to exchange rate movements causes a decline in output and employment, increasing at the same time the volume of public debt (see Figure 17.2).

Therefore, Brazil's very high rates of interest are the result of high country risk[8] (due to marked external vulnerability and the risk of fiscal insolvency), and of adopting an inflation-targeted regime in a context of various macroeconomic constraints and a high level of internal debt. High interest rates have had two effects: (i) they have constrained economic growth, through the price of credit (loan rates) and entrepreneurs' negative expectations; and (ii) they have increased public debt, which is formed mainly by indexed bonds or short-term pre-fixed bonds. Indeed, the strong demand for hedges against exchange devaluation and interest rate changes in turbulent periods has influenced Brazil's internal public debt. The brazilian government has been obliged to offer exchange rate and interest rate hedges to buyers of securities, who charge high risk premiums to roll over the public debt. As a result, since the end of 1998, more than 50% of federal domestic securities have been indexed to the overnight rate, while more than 20% have been indexed to foreign exchange. In addition, the ratio of net public debt to GDP rose from 34.4% in December 1997 to 53.5% in December 2003; in 2004 this ratio declined due to both economic growth and exchange rate appreciation (Table 17.1).

The behaviour of the domestic public debt in Brazil has proved particularly vulnerable to changes in interest and exchange rates. Reducing the public debt depends on reducing the related financial burden by bringing down the interest rate or raising the exchange rate, and/or boosting the primary fiscal surplus. Thus, the Brazilian government has been forced to generate a high primary fiscal surplus (more than 3.5% of GDP), which stands in the way of any anti-cyclical fiscal policy, while the fiscal effort itself is partly neutralised by increases in the rates of interest or exchange.

Another reason why economic growth in Brazil has remained above its potential growth is that credit has declined since the beginning of 1995 (Figure 17.1(h)). One of the main factors preventing increased credit in Brazil lies in its very large banking spreads (Figure 17.1(g)), which explain, at least partly, the high profitability of the banking sector in Brazil (Paula and Alves, Jr, 2003). Although the banking spread has declined in recent years in Brazil, it is still very substantial by international standards: in 2000, the annual banking spread was 38.72% in Brazil, while it was 11.96% in Mexico, 2.75% in Argentina, 5.64% in Chile, 2.77% in the US, and 3.15% in the Euro area (Afanasieff *et al.*, 2001, p. 7).

4. REQUIREMENTS FOR THE SUSTAINED GROWTH OF THE BRAZILIAN ECONOMY

In this section we will use a simple version of the Harrod–Domar growth model in order to obtain an estimate of the *warranted growth rate* of the brazilian economy under the conditions imposed by the current economic policy. As we will see, the warranted growth rate under current economic conditions is no higher than 2.5% per year. This growth rate is clearly unsatisfactory for an economy in which population growth rate is around 1.8% per year and productivity growth is estimated at 2.6% per year. This means that the *warranted growth rate* of the brazilian economy is lower than the *natural* long-run growth rate. This 'disequilibrium' between warranted and natural growth rates of the brazilian economy is the main cause of the high unemployment rate and of the decreasing of real average income observed recently in Brazil (Table 17.1).

Let us start with an economy in which firms employ a Leontieff-type technology, the stock of capital being the limiting factor to firms' production level (see Marglin, 1984, Chapter 5). In this setting, the potential output of this economy is given by:

$$Y = \sigma K, \quad \text{and } \sigma \equiv \frac{1}{v} \tag{1}$$

Where v is the capital–output coefficient, that is the technical coefficient that shows the amount of 'capital' that is necessary for the production of one unit of final output.

Taking the time derivative of (1) and supposing a constant depreciation rate equal to c we arrive at the following expressions:

$$\dot{Y} = \sigma \dot{K} \tag{1a}$$

$$\dot{K} = I - cK \tag{2}$$

where I is the gross (planned) investment.

We will also suppose that households save a constant share s of their income. So, planned savings are given by:

$$S = sY \tag{3}$$

One requirement for a sustained growth of the economy in the long-run is the equality between aggregate output and effective demand. For this, it is necessary that planned investment be equal to planned saving. Taking for granted the occurrence of this equality, we can substitute (3) in (2) in order to get the following expression:

$$\dot{K} = sY - cK \tag{4}$$

After substituting (4) in (1a), we get:

$$\dot{Y} = \sigma(sY - cK) \tag{5}$$

Finally, after substituting (1) in (5) and dividing both sides of the resulting expression by Y, we arrive at the *fundamental growth equation* of the Harrod–Domar model given by:

$$g = \frac{\dot{Y}}{Y} = \frac{s}{v} - c \tag{6}$$

Equation (6) determines the *warranted growth rate*, that is, the growth rate of output that – if obtained – will assure the equality between effective demand and aggregate output over time.

In order to use equation (6) to estimate the *warranted growth rate* of the Brazilian economy, we must have realistic values for the following parameters: capital–output coefficient, investment and saving rates and depreciation rate of the capital stock.

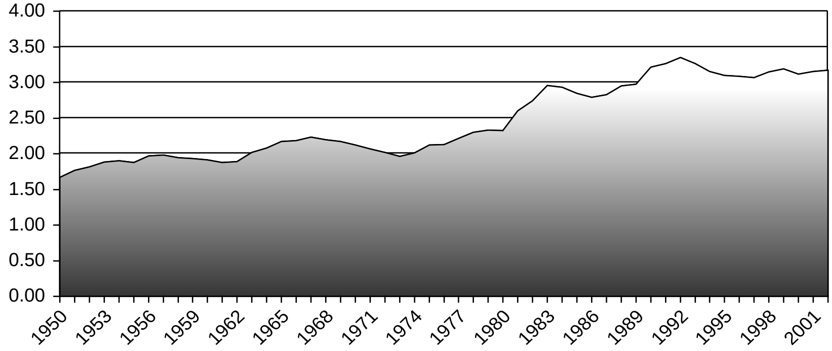


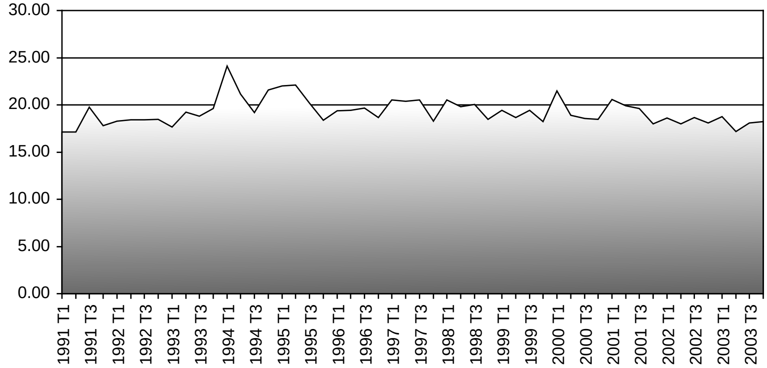
Figure 17.3 Capital–output coefficient in Brazil (1950–2002)

Estimates of the first two variables can be easily obtained at IPEADATA (www.ipeadata.gov.br). Capital–output coefficient shows a clear *upward* trend in the last fifty years, as we can see in Figure 17.3. Such an upward trend makes difficult, if not impossible, the occurrence of a reduction in the capital–output coefficient in the near future. However, taking a simple average of the capital–output coefficient in the period 1989–2002, we will arrive at a value equal to 3.16, which can be taken as the *minimum possible value* for this parameter in equation (6).

Investment rate, defined as gross capital formation divided by GDP, shows a remarkable stability in the period 1991–2003. This rate, according to IPEADATA, had fluctuated around 19.26% of GDP in this period, as we can see in Figure 17.4. So we will take 19.26% as a plausible estimate for the value of the parameter s in equation (6).

Unfortunately, we have found no estimates about depreciation rate of the capital stock for the brazilian economy, so we have no other option than to use the values of this parameter from other economies. Romer (2001, p. 25) estimates the depreciation rate of the capital stock for US economy as lying between 3 and 4% per year. So an average estimate for the depreciation rate of the capital stock for the US economy is around 3.5% per year. Based on some similarities of the industrial sectors of Brazil and the United States, we will use this value as an estimate for the depreciation rate of the capital stock of the Brazilian economy.

Taking $s = 0.1926$, $v = 3.16$ and $c = 0.035$ in equation (6), we get $g = 0.025$, that is an *equilibrium growth rate* equal to 2.5% per year. For several reasons, this growth rate is completely unsatisfactory for Brazil. First of all, Brazil grew at an average rate of 7.0% per year in the period between 1930 and 1980. Second, this rate is lower than the *natural growth rate*; that is, the



Source: IPEADATA.

Figure 17.4 Investment rate in Brazil (1991–2003)

Table 17.2 Productivity growth in Brazil, 1950–1997

Period	Average growth rate of labour productivity (%)
1950–1955	2.7
1955–1960	2.7
1960–1965	2.5
1965–1970	2.5
1971–1973	5.6
1974–1980	1.0
1981–1985	0.3
1986–1990	−0.8
1991–1997	7.1
1950–1997	2.62

Source: Franco (1999, p. 150).

required growth rate for full employment of the labour force. Estimating a population growth of 1.8% per year (see IPEADATA) and a productivity growth of 2.6% per year (Table 17.2), output must grow at a minimum rate of 4.4% per year in order to employ the new workers and those who have lost their jobs due to technological progress. An average growth rate lower than 4.4% per year implies that unemployment and/or 'underemployment' will increase over time. Last, but not least, this growth rate is clearly insufficient for brazilian economy to *catch-up* developed economies. The average growth rate of developed countries lies between 2.5 to 3% per year.

If the Brazilian economy grows at an average rate of 2.5% per year, then the income gap between Brazil and developed countries will be constant or will increase in the long-run.

In face of these arguments, we consider an average growth rate of 5% per year a desirable and realistic goal for the brazilian economic policy.[9] In order to achieve this goal, the investment rate – according to equation (6) – must increase to 27% of GDP.

5. HOW TO INCREASE THE INVESTMENT RATE? AN AGENDA OF REFORM FOR THE BRAZILIAN ECONOMY

As we have seen in the last section, the average investment rate in the last fifteen years was insufficient to generate robust growth for the Brazilian economy. This behaviour of the investment rate was mainly due to the 'economic policy model' adopted by brazilian policy makers since the beginning of the 1990s. This economic model was characterized by: (i) high nominal and real interest rates in order to achieve price stability; (ii) growing liberalization of the capital account in order to integrate Brazil with international capital markets; (iii) overvaluation of domestic currency;[10] and (iv) since 1999, an increasing primary fiscal surplus – generated mainly by the reduction of public investment – in order to stabilize the public debt–GDP ratio.

This *economic policy model* has only succeeded in achieving a low rate of inflation compared to the period of high inflation; that is before 1994. Indeed, since 1996, inflation rate in Brazil has been lower than 20% per year. However, public debt, as a ratio to GDP, increased from 30% in 1994 to almost 55% in 2004, with GDP growing at an average rate of 2.4% per year in the period 1995–2004. Price stability is, of course, an important goal of economic policy, but not the only one. A robust economic growth and stability of the public debt–GDP ratio are also very important.

In order to achieve a higher investment rate, the economic policy model must be changed. Nominal and real interest rates must be reduced for entrepreneurs to increase private investment. The primary fiscal surplus must also be reduced. Brazil needs to increase public investment in infrastructure to generate positive externalities for private investment. Nominal and real exchange rates must be kept at competitive levels in order to generate a sustained current account surplus, which is required to reduce the amount of external debt and the level of external fragility of the brazilian economy.

The challenge is to make these changes compatible with: (i) price stability; and (ii) stabilization or reduction in the level of public debt. Brazil spent almost 15 years fighting against very high inflation rates. The reduction in inflation rates obtained after the 'Real Plan' was a very important achievement, and must be maintained. Stabilization in the level of public debt is also important. Brazil simply cannot stand with public debt–GDP ratios higher than 50%. With public debt as high as the ratio to GDP, almost all the efforts of the financial sector were devoted to finance public debt, thereby causing a reduction in the level of banking credit to finance private expenditures. As we saw in Section 3, Brazil had a very low credit to GDP ratio. The main reason for this is that banks prefer to buy public bonds, which are very liquid and profitable, rather than to incur the risks of lending money to private enterprises (Paula and Alves Jr, 2003).

An alternative economic policy model for the brazilian economy[11] should be based on the following principles:

1. Adoption of a *crawling-peg exchange rate regime* in which the devaluation rate of domestic currency was set by the Central Bank at a rate equal to the difference between a *target inflation rate* (determined by National Monetary Council – CMN) and *average inflation rate* of Brazil's most important trade partners, namely United States, European Union, China, Japan and Argentina.
2. Adoption of *market-based capital controls* in order to increase the autonomy of the Central Bank to set nominal interest rates according to domestic objectives (mainly to promote a robust growth) and to avoid the likelihood of speculative attacks on the brazilian currency.
3. Reduction of the nominal interest rate to a level compatible with a real interest rate of 6.0% per year.
4. Reduction of primary fiscal surplus from the current 4.5% of GDP to 3.0% of GDP on average for a period of 10 years. This reduction must be used to increase public investment in the same amount.

The first principle of the 'alternative economic policy model' entails the abandonment of the current *Inflation Targeting Regime* (hereafter ITR). As we know, in the ITR, monetary policy is directed only to price stability. For the workings of this system, however, there must be a *floating exchange rate regime*. This exchange rate regime has not worked well in the brazilian case. First of all, since the adoption of such a regime, at the beginning of 1999, there was a huge volatility in the nominal exchange rate as we saw in Section 3. This volatility increases exchange rate risks and the *uncertainty* surrounding investment decisions. Secondly, this system was not capable of avoiding the problem of exchange-rate over-valuation. For instance, the nominal exchange rate between the US dollar and the brazilian currency ('real') fell from R$3.50 in June of 2003 to R$2.20 in December of 2005, an appreciation of almost 37% in 30 months. Such a huge appreciation in the nominal exchange rate can soon reduce sharply the current account surplus, thereby increasing the level of Brazilian external debt.

Adoption of a *crawling-peg exchange rate regime* will reduce the *exchange rate risk* – contributing to an increase in private investment – and to the maintenance of the nominal exchange rate at competitive levels, provided that the *initial level of the nominal exchange rate* – that is, the level set in the first day of the new regime – is not over-valued.

Another interesting feature of the *crawling-peg exchange rate regime* is that it will serve as the *nominal anchor* for the brazilian economy, substituting ITR as a device for inflation control. If *Purchasing Power Parity theorem* (thereafter PPP) holds true, then the (effective) domestic rate of inflation (π)[12] is equal to exchange rate depreciation (Δe) *plus* international inflation rate(π^*). In the *crawling-peg exchange rate regime*, the Central Bank sets the rate of depreciation of domestic currency, maintaining domestic rate inflation at a level near the one dictated by *PPP*.

Accumulated experience during ITR shows that an implicit *target inflation* of 8.0% per year is a realistic goal for economic policy in Brazil. Supposing that the international rate of inflation lies between 1.5% to 2.0% per year, the Central Bank will set the rate of domestic currency devaluation at 10% per year under the *crawling-peg exchange rate regime*. A competitive level for the initial value of nominal exchange rate under this new regime should be R$3.20.[13]

The second principle of the 'alternative economic policy model' is the adoption of capitals control. Such control is necessary for two basic reasons. First of all, to increase private investment, a substantial reduction in the level of domestic interest rates is necessary. In fact, in the last six years (1999–2004), real interest rates were up to 11% per year. Under the actual open capital account situation of the brazilian economy, a sharp reduction in interest rates may cause a huge capital outflow, making it impossible for the Central Bank to control nominal exchange rate devaluation. To avoid this result, the implementation of controls over *capital outflows* is necessary. The second reason is that control over the *nominal exchange rate* may not be sufficient to avoid a substantial appreciation of the *real exchange rate* in the presence of huge *capital inflows*. These flows will make the Central Bank to increase the stock of high powered money due to the buying of foreign reserves, which is necessary to sustain the nominal exchange rate at the level determined by the monetary authorities. In the absence of sterilization, this may produce an excessive increase in aggregate demand that can generate inflationary pressures in the economy and, given the rate of depreciation of the nominal exchange rate, real exchange rate appreciation.

We propose the adoption of *market-based capital controls*, that is, the introduction of income taxes over the yield of foreign investment in brazilian assets.[14] These taxes should be proportional to the length of investment in these assets. For example, a one-year investment in brazilian assets should be taxed at a rate of 35% over all yields generated by these assets during this period. A two-year investment should be taxed at a much lower rate, for example, 28%. A three-year investment must be taxed at an even lower rate of 19%. The idea is to give to foreign investors a clear and strong incentive to make their investment in brazilian assets as long-term as possible in order to create *market incentives for the reduction of capital outflows*.

To reduce capital inflows, the introduction of reserve requirements over all capital inflows is necessary, except foreign direct investment, as done by Chile in the beginning of the 1990s. The idea is to oblige foreign investors to make a deposit of a fixed percentage of the value of their investment in brazilian assets in the Central Bank. These deposits will receive a zero yield over the entire investment period. This will reduce the *ex ante* yield of these assets for foreign investors, *creating a market incentive for the reduction of capital inflows*.

After the implementation of the *crawling-peg exchange rate regime* and *market-based capital controls*, it will be possible to reduce the level of domestic interest rates without producing an increase in inflation rate and/or a huge capital outflow. The relevant question now is: how much reduction in the level of interest rates is possible in economic terms?

In a regime of fully open capital account, the answer would be very simple: interest rates can be reduced to a level equal to the one dictated by *uncovered interest rate parity* – in other words, international interest rates plus the risk premium required for foreign investors to buy domestic assets plus the expected rate of depreciation of domestic currency.

In the brazilian case, the relevant international interest rates were the interest rates over US government bonds with the same maturity as the brazilian government bonds.[15] This rate is near 4.0% per year. The risk premium over brazilian sovereign bonds was near 450 basis points at the beginning of 2005. Supposing the validity of *PPP* in the long run, the expected rate of domestic currency depreciation must be equal to the difference between domestic and international rate of inflation. For a domestic rate of inflation of 8% and an international rate of inflation of 2%, the expected rate of currency depreciation should be equal to 6%. So, the nominal interest rate in Brazil could be reduced from the current 18.25% per year to 14.5% per year without producing a huge capital outflow or an increase in inflation rate which would imply a real interest rate of 6.5% per year.

With capital controls, however, a much higher reduction in the level of interest rates will be possible. So, it would be possible to reduce nominal interest rates to 12% per year, generating a real interest rate of 4% per year. However, the high level of public debt as a ratio to GDP may set a *downward limit* to the reduction in the level of nominal and real interest rates. It is true that brazilian government bonds have a great degree of liquidity, since secondary markets where these assets are traded – in Brazil or abroad – are well organized. This means that investors (both domestic and foreign) have a low required rate of return for investment in these assets. But Brazil is not the United States or Germany. Investors still have doubts about the *inter-temporal solvency* of the brazilian government. In this case, a very low real interest rate may make impossible for the Treasury to roll over the existing debt. Prudence dictates a certain degree of conservatism in the setting of nominal and real interest rate levels by the Central Bank.

This reasoning shows to us that a real interest rate of 6% per year, although still a high level, is a perfectly realistic value for the brazilian economy and should be the target of the monetary policy.

Once real interest rates are reduced to 6% per year, it will be possible to reduce the level of primary fiscal surplus. The required level of primary surplus is determined by a *government inter-temporal solvency condition.* This condition determines the *minimum level of primary surplus that is compatible with a constant public debt to GDP ratio*. This condition is given by the following equation:[16]

$$s = \left[\frac{r-g}{1+g}\right]b \tag{7}$$

where s is the primary surplus as a ratio to GDP, r is the level of real interest rate, g is the growth rate of real GDP, and b is the ratio of public debt to GDP.

Under the conditions imposed by the current economic policy model, we have $r=0.11$; $g=0.025$; $b=0.53$. So the minimum level of primary fiscal surplus must be 4.4% of GDP. However, a successful implementation of the alternative economic policy model may change the values of these parameters to: $r=0.06$; $g=0.05$; $b=0.53$. In this case, the minimum level of primary surplus can be reduced to 0.5% of GDP.

So the reduction of primary surplus from actual 4.5% of GDP to 3.0% of GDP is not only compatible with the inter-temporal solvency condition, but also with a cumulative reduction of public debt as a ratio to GDP. Under the conditions supposed by the alternative economic policy model, the public debt as a ratio to GDP will be reduced to 32% of GDP in 2012, as we can see in Figure 17.5.[17]

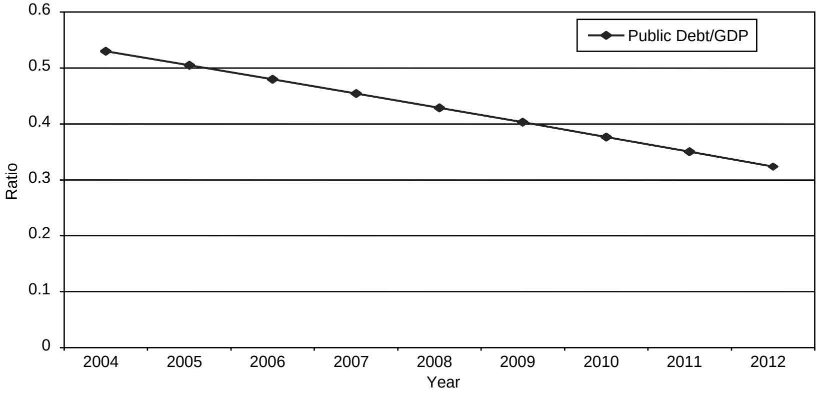


Figure 17.5 Expected dynamics of public debt as a ratio to GDP in Brazil under the alternative economic policy model

The reduction of primary surplus is essential for the increase in public investment. We propose that the *entire reduction in primary surplus be used to increase public investment*. In this case, public investment will be increased by 1.5% of GDP. Assuming a one-to-one relation between public and total investment, the average rate of investment will be increased to 20.67% of GDP and the potential growth rate will be increased to 3.04% per year.

However, there are good reasons to suppose that an increase in public investment will increase total investment (public *plus* private) at a rate greater than one to one. First of all, as recognized even by neoclassical growth theorists such as Barro (1990), public investment generates *positive externalities* for the private sector, so an increase in public investment will increase profits in the private sector, stimulating entrepreneurs to increase their investment spending. Secondly, a reduction in the primary fiscal surplus will certainly increase aggregate demand due to the well known *government spending multiplier*. In a *regime of excess capacity* such as the one that characterized the Brazilian economy since the beginning of the 1980s, firms will increase output in order to meet the additional demand for their products. The increase in the production level will generate also an increase in the degree of capacity utilization, stimulating firms to increase their investment spending in order to make the adjustment between effective and *desired* degree of capacity utilization (see Oreiro, 2004b). In other words, the increase in the level of capacity utilization will produce an increase in private investment due to the well known *accelerator effect*.

So we can assume that an increase in public investment will produce a higher than one-to-one increase in total investment. We do not have a precise estimate of this magnitude, but an 'educated guess' is that an increase in public investment will induce a 1.5 increase in total investment. Under these conditions, an increase in public investment by 1.5% of GDP will increase total investment by 2.25% of GDP. This means that potential growth rate of real GDP will be increased up to 3.3% per year. This effect, combined with the positive stimulus over private investment from the reduction in the level of real interest rates and from the elimination of uncertainty due to the exchange rate risk, will generate the required increase in the investment rate for the sustained growth of the Brazilian economy at a rate of 5.0% per year.

6. CONCLUSION

This chapter presented a Keynesian strategy of economic policy that aims to achieve higher, stable and sustained economic growth in Brazil. The basic features of this strategy are: (i) adoption of a *crawling-peg exchange rate regime* in which devaluation rate of domestic currency is set by the Central Bank at a rate equal to the difference between a *target inflation rate* and the *average inflation rate* of Brazil's most important trade partners; (ii) adoption of *market-based capital controls* in order to increase the autonomy of the Central Bank to set nominal interest rates according to domestic objectives (mainly to promote robust growth); (iii) reduction of nominal interest rates to a level compatible with a real interest rate of 6.0% per year; (iv) reduction of the primary surplus from the current 4.5% of GDP to 3.0% of GDP. These elements are fundamental for the required increase in the investment rate of the brazilian economy from the current 20% of GDP to the 27% of GDP needed for a sustained growth of 5% per year.

NOTES

1. We are following, in broader terms, the basic structure of the economic constraints for a full employment policy developed by Arestis and Sawyer (1998).
2. On Keynes's principle of effective demand, see among others Davidson (2002, Chapter 2).
3. See, among other references, Thirlwall (2002).
4. See, for instance, Guérin and Lachrèche-Révil (2003).
5. We refer again to Arestis and Sawyer (1998, pp. 190–1).
6. See Paula and Alves, Jr (2000) and Saad-Filho and Morais (2002) for an analysis of the 1998–1999 Brazilian currency crisis.
7. Although capital account has been gradually liberalized since the early 1990s, more recently it has been eased.
8. Bresser-Pereira and Nakano (2002) suggest that the causality between interest rate and country-risk may be inverse; since short-term interest rates have been very high, foreign creditors believe that country-risk is high.

9. A growth rate higher than the estimated natural growth rate of the brazilian economy for several years (one decade or so) is possible due to the existence of a high unemployment rate (more or less 10% of the labour force) and, more importantly, due to the existence of a very big informal (and low producitivity) sector in Brazil.
10. Except in the brief period between June of 2002 to June of 2003 due to the exchange rate crisis of the final period of the Cardoso administration (1994–2002).
11. The ideas shown here were originally proposed by Oreiro *et al.* (2003, ch 4).
12. Effective inflation rate may be different from *target inflation rate*, which is set by the National Monetary Council and is a reference for the nominal exchange rate devaluation, due to the occurrence of supply shocks.
13. This implies that during the transition from the actual *free floating exchange rate regime* to the *crawling-peg regime* there must be a *nominal exchange rate appreciation* of almost 19%. This will generate a *transitory increase* in the rate of inflation due to *pass-through effect* of exchange rate to prices. To avoid a *permanent increase* in the rate of inflation, it is necessary that *real wages* are reduced in order to make possible a *real exchange rate depreciation*. This means that during the transition from the old to the new exchange rate regime, nominal and real interest rates must be kept at high levels to force *unions* to accept a *reduction in real wage*. Once the new exchange rate regime is implemented and inflation has returned to its prior level, interest rates can be reduced.
14. This proposal was originally offered by Paula *et al.* (2003).
15. The average maturity of brazilian government bonds is around 30 months.
16. See Oreiro (2004a) for a detailed discussion of this condition. A similar – although not identical – condition can be found in Palley (2004).
17. This figure was obtained by the numerical simulation of the equation

$$b_t = \left[\frac{1+r}{1+g}\right] b_{t-1} - s_t \text{, taking } s = 0.03;\ r = 0.06;\ g = 0.05 \text{ and } b(0) = 0.53.$$

REFERENCES

Afanasieff, T.S., P.M. Lhacer and M.I. Nakane (2001), 'The determinants of bank interest spread in Brazil', in *Proceedings of XXIX Encontro Nacional de Economia*, ANPEC, Salvador.

Arestis, P. and M. Sawyer (1998), 'Keynesian economic policies for the new millennium', *The Economic Journal*, **108**, 181–95.

Barro, R. (1990), 'Government spending in a simple model of endogenous growth', *Journal of Political Economy*, **98** (5), 103–25.

Belaisch, A. (2003), 'Do Brazilian banks compete?', *IMF Working Paper* WP/03/113, Washington, DC.

Bresser-Pereira, L.C. and Y. Nakano (2002), 'Uma estratégia de desenvolvimento com estabilidade', *Brazilian Journal of Political Economy*, **3** (3), 146–77.

Carvalho, F. (1997), 'Economic policies for monetary economies: Keynes' economic policy proposals for an unemployment-free economy',. *Brazilian Journal of Political Economy*, **17,** 31–51.

Davidson, P. (1994), *Post Keynesian Macroeconomic Theory*, Aldershot, UK and Brookfield, USA Edward Elgar.

Davidson, P. (2002), *Financial Markets, Money and the Real World*, Cheltenham, UK and Northampton, MA, USA, Edward Elgar.

Franco, G.H.B. (1999), *O Desafio Brasileiro: ensaios sobre desenvolvimento, globalizaçãoe moeda*, Editora 34, São Paulo.

Guérin, J.-L. and A. Lahrèche-Révil (2003), *Exchange Rate Volatility and Investment*, Mimeo.

Harrod, R.F. (1939), 'An essay in dynamic theory', *Economic Journal*, **49**, 14–33.

IMF (2003), *Public debt in emerging markets: is it too high?*, Washington, DC, IMF.

IPEADATA, www.ipeadata.gov.br.

Keynes, J.M. (1973), *The General Theory and After, Part I: Preparation. Collected Writings of John Maynard Keynes*, vol. XIII, London, Macmillan.

Keynes, J.M. (1980), *Activities 1940–46 Shaping the Post World: Employment and Commodities. Collected Writings of John Maynard Keynes*, vol. XXVII, London, Macmillan.

Kregel, J. (1994–95), 'The viability of economic policy and the priorities of economic policy', *Journal of Post Keynesian Economics*, **17** (2), 261–77.

Marglin, S. (1984), *Growth, Distribution and Prices*, Cambridge, MA, Harvard University Press.

Oreiro, J.L. (2004a), 'Prêmio de risco endógeno, equilíbrios múltiplos e dinâmica da dívida pública: uma análise teórica do caso brasileiro', *Revista de Economia Contemporânea*, **8**.

Oreiro, J.L. (2004b), 'Accumulation regimes, endogenous desired rate of capacity utilization and income distribution', *Investigación Económica*, **63**.

Oreiro, J.L., J. Sicsú and L.F. Paula (2003), 'Controle da dívida pública e política fiscal', in Sicsú *et al*. (2003).

Palley, T. (2004), 'Escaping the debt constraint on growth: a suggested monetary policy for Brazil', *Brazilian Journal of Political Economy*, **24 (**1).

Paula, L.F. and A.J. Alves Jr (2000), 'External financial fragility and the 1998–1999 Brazilian currency crisis', *Journal of Post Keynesian Economics*, **22** (4), 589–617.

Paula, L.F. and A.J. Alves Jr (2003), 'Banking behaviour and the Brazilian economy after the real plan: a Post-Keynesian approach', *BNL Quarterly Review*, **227**, 337–65.

Paula, L.F., J.L. Oreiro and G. Silva (2003), 'Fluxo e controle de capitais no Brasil', in Sicsú *et al*. (2003).

Romer, D. (2001), *Advanced Macroeconomics*, New York, McGraw Hill, Second Edition.

Saad-Filho, A. and L. Morais (2002), 'Neomonetarist dreams and realities: a review of the Brazilian experience', in P. Davidson (ed.), *A Post Keynesian Perspective on 21st Century Economic Problems*, Cheltenham, UK and Northampton, MA, USA, Edward Elgar, pp. 29–55.

Sicsú, J, J.L. Oreiro and L.F. Paula (eds) (2003). *Agenda Brasil: Políticas Econômicas para o Crescimento com Estabilidade de Preços*, Manole, Barueri.

Thirlwall, A.P. (2002), *The Nature of Economic Growth*, Cheltenham, UK and Northampton, MA, USA, Edward Elgar.

Tobin, J. (2000), 'Financial globalisation', *World Development*, **28** (6) 1101–104.