4

THEORIES OF ECONOMIC GROWTH: OLD AND NEW

Introduction

Growth and development theory is as old as economics itself. The great classical economists of the eighteenth and nineteenth centuries were, in a sense, all development economists writing about forces determining the progress of nations as the countries of Europe embarked on the process of industrialization. The most famous of the early classical economists was Adam Smith, who published a path-breaking book in 1776 entitled *An Inquiry into the Nature and Causes of the Wealth of Nations*. This work is still widely cited today because Smith recognized how specialization in industrial activities could lead to increasing returns and big increases in labour productivity, compared with specialization in agriculture.

In this chapter we examine how theories of growth and development have evolved through time from the classical economists to modern-day thinking about factors that determine the pace of economic growth. Other great classical economists, apart from Smith, were Thomas Malthus, David Ricardo and John Stuart Mill, who were all pessimistic about the growth and development process because of the pressure of population on food supply and diminishing returns in agriculture, which, they argued, would reduce the rate of profit in industry. Eventually, a stationary state would be reached. Even more pessimistic was Karl Marx, who predicted the collapse of capitalism itself.

Modern growth theory started with (Sir) Roy Harrod's well-known 1939 paper, 'An Essay in Dynamic Theory' (Harrod, 1939). His main purpose was to make dynamic Keynes' static theory of income determination. In doing so, he showed how unstable economies can be in the short run, and how, in the long run, it is possible for countries to experience either prolonged periods of secular stagnation if the supply of saving exceeds the demand for it, or structural unemployment if the growth of the labour force exceeds the growth of capital (as it does in most developing countries).

The neoclassical response to the Harrod model was Robert Solow's well-known growth model (Solow, 1956), which attempts to show that if the prices of factors of production are flexible, and labour and capital are substitutable, it is possible for countries to achieve equilibrium growth at the so-called 'natural rate', determined by the growth of the labour force and labour-saving technical progress. What determines labour force growth and technical progress is left unexplained. Because of the assumption of diminishing returns to capital, investment itself does not matter for long-run growth. The model also predicts that capital-scarce countries should grow faster than capital-rich countries, leading to a convergence of per capita incomes across the world, because the marginal product of capital should be higher in poor countries than in rich countries.

As we saw in Chapter 2, however, convergence of living standards across countries is not apparent, and this led in the 1980s to the development of the so-called 'new' growth theory or endogenous growth theory, which relaxes the assumption of diminishing returns to capital by redefining capital to include improvements in human capital and new techniques of production through research and development (R&D) expenditure. In these 'new' models, convergence is only conditional, and investment matters for long-run growth because the marginal product of capital does not decline as more investment takes place.

We not only look at theory, but also empirics, and see how the neoclassical production function has been used to analyse the sources of growth in developing countries, and how 'new' growth theorists treat their models empirically.

An attempt is made to illustrate the contemporary relevance of the theories discussed. We shall, in fact, discover that the wheel has turned full circle, and that the most recent theories of endogenous growth rehabilitate many of the ideas of the old classical economists, particularly Adam Smith's emphasis on **increasing returns** associated with investment in manufacturing industry, and the general emphasis in both classical and Keynesian theory on the role of capital accumulation, and the embodiment of various forms of technical progress associated with it.

The chapter ends with a discussion of the topic of growth diagnostics and identifying binding constraints on growth.

Classical growth theory

The macroeconomic issues of the growth of output, and the distribution of income between wages and profits, were the major preoccupation of all the great classical economists, including Adam Smith, Thomas Malthus, David Ricardo and, last but not least, Karl Marx. This discussion starts with **Adam Smith**, because while Smith had a generally optimistic vision of the growth and development process, the later classical economists tended to have a more gloomy vision. This led historian Thomas Carlyle to describe economics as a 'dismal science' – not a sentiment, we hope, that will be shared by students reading this book.

Adam Smith

Born 1723, Kirkcaldy, Scotland. Died 1790. Professor of Moral Philosophy, University of Glasgow. Often described as the 'father' of modern economics. Famous for two main books: *The Theory of Moral Sentiments* (1759) and *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776). Strong advocate of free markets and free trade, but most important for recognizing the role that increasing returns play in the growth and development process, based on the principle of the division of labour or specialization – a characteristic of manufacturing industry in particular.

Adam Smith and increasing returns

One of Smith's most important contributions was to introduce into economics the notion of **increasing returns**, based on the **division of labour**. He saw the division of labour (or gains from specialization) as the very basis of a social economy, otherwise everyone might as well be their own Robinson Crusoe producing everything they want for themselves. This notion of increasing returns, based on the division of labour, lay at the heart of his optimistic vision of economic progress as a self-generating process, in contrast to later classical economists who believed that economies would end up in a stationary state owing to diminishing returns in agriculture and a falling rate of profit in industry. It was also in contrast to Marx, who believed that capitalism would collapse through its own 'inner contradictions', by which he meant competition between capitalists reducing the rate of profit and the alienation of workers.

Given the central importance of increasing returns, the essence of Smith's model is a simple one, and many of the features he emphasizes will be a recurring theme in this and other chapters. The growth of output and living standards depends first and foremost on investment and capital accumulation. Investment, in turn, depends on savings out of the profits generated by industry and agriculture and the degree of labour specialization (or division of labour). The division of labour determines the level of labour productivity, but **the division of labour is limited by the extent of the market**. The extent of the market, however, depends partly on the division of labour as the determinant of per capita income. We have here a circular cumulative interactive process, although not without constraints, as we shall see later.

The notion of increasing returns may, on the surface, appear to be relatively trivial, but it is of profound significance for the way economic processes are viewed. It is not possible to understand divisions in the world economy and so-called 'centre–periphery' models of growth and development (see Chapter 10) without distinguishing between activities that are subject to increasing returns on the one hand, and diminishing returns on the other. 'Increasing returns' means rising labour productivity and per capita income as output and employment expands, while 'diminishing returns' means falling labour productivity and per capita income and a limit to the employment of labour at the point where the marginal product of labour falls to the level of the subsistence wage. Beyond that point, there will be no more employment opportunities, and there will be disguised unemployment (see Chapters 3 and 5). Increasing returns are prevalent in most industrial activities, while diminishing returns characterize land-based activities such as agriculture and mining, because land is a fixed factor of production – and one of the few incontrovertible laws of economics is that if a variable factor is added to a fixed factor, its marginal product will eventually fall (the law of diminishing returns). Poor developing countries tend to specialize in diminishing returns activities, while rich developed countries tend to specialize in increasing returns activities, and this is one of the basic explanations of the rich country–poor country divide in the world economy. As we see later, it is the concept of increasing returns (or more precisely, non-diminishing returns to capital) that lies at the heart of the new endogenous growth theory.

Adam Smith (1776) gives three sources of the increasing returns to be derived from the division of labour:

> This great increase in the quantity of work, which, in consequence of the division of labour, the same number of people are capable of performing, is owing to three different circumstances; first to the increase of dexterity in every particular workman [what we now call learning by doing]; secondly, to the saving of time which is commonly lost in passing from one species of work to another; and lastly, to the invention of a great number of machines which facilitate and abridge labour, and enable one man to do the work of many.

That is, specialization provides greater scope for capital accumulation by enabling complex processes to be broken up into simpler processes permitting the use of machinery. But the ability to specialize, or the division of labour, depends on the extent of the market. Smith uses the example of the production of pins. There is no point in installing sophisticated machinery to deal with the different processes of pin production if the market for pins is very small. It is only economical to use cost-saving machinery if the market is large. If the market is small, there would be surplus production. To quote Smith (1776) again:

> when the market is very small, no person can have any encouragement to dedicate himself entirely to one employment, for want of power to exchange all that surplus part of the produce of his own labour, which is over and above his own consumption, for such parts of the produce of other men's labour as he has occasion for.

Smith (1776) recognized, however, that increasing returns based on the division of labour were much more a feature of industry than agriculture:

> the nature of agriculture, indeed, does not admit of so many subdivisions of labour, nor of so complete a separation of one business from another, as manufactures. It is impossible to separate so entirely the business of the grazier from that of the corn farmer, as the trade of the carpenter is commonly separated from that of the smith.

This does not mean, of course, that agriculture is unimportant in the development process. On the contrary. Even though industry offers more scope for the division of labour, it would be difficult for industry to develop at all without an agricultural surplus, at least in the absence of imports. Smith recognized that an agricultural surplus is necessary to support an industrial population, and labour released by improved productivity in agriculture can be used for the production of non-agricultural goods. So, agriculture is certainly important for industrialization from the supply side. On the demand side, it is the agricultural surplus that gives rise to the demand for other goods, which can be purchased with the excess supply of agricultural goods. As Smith (1776) put it: 'those, therefore, who have the command of more food than they themselves can consume, are always willing to exchange the surplus – for gratification of this other kind [manufactured goods]'. We have here a model of reciprocal demand between agriculture and industry, with industry demanding food from agriculture to feed workers, and agriculture exchanging its surplus for industrial goods. Balanced growth between agriculture and industry is essential for the growth and development process to proceed without impediment. Many later models of economic development reflect this insight (see Chapters 5 and 9). For a full exposition of Smith's vision of the development process, in which aggregate demand and aggregate supply interact in a cumulative expansionary process, together with structural change, see Kim (2015).

The division of labour is limited by the size of the market. This is a central axiom of Smith's model. The size of the market will be partly limited by restrictions on trade; hence Smith's advocacy of free trade and laissez-faire, internally and externally. Goods must be able to be exchanged freely between industry and agriculture. But demand for industrial goods can also come from abroad, and Smith (1776) recognized the role of exports in the development process:

> without an extensive foreign market, [manufactures] could not well flourish, either in countries so moderately extensive as to afford but a narrow home market; or in countries where the communication between one province and another was so difficult as to render it impossible for the goods of any particular place to enjoy the whole of that home market which the country could afford.

The subject of trade and growth, and models of export-led growth, is discussed in detail in Chapter 15.

Smith's model of development is driven by capital accumulation generated by profits from industry; and the stimulus to invest, as in all classical models, comes from the rate of profit. If the rate of profit falls, the desire to invest diminishes. Smith was somewhat ambiguous about what happens to the rate of profit as development proceeds. On the one hand, he recognized that as the economy's capital stock grows, the profit rate will tend to fall due to competition between capitalists and rising wages. On the other hand, new investment opportunities raise the rate of return. Thus, the rate of profit may rise or fall in the course of development depending on whether investment is in old or new technology. If there is any tendency towards a stationary state, in which the rate of profit falls to zero so that there is no further incentive to invest, it is a long way off in Smith's model, in contrast to the models of Malthus, Ricardo and Marx, in which a fall in the rate of profit is seen as inevitable.

Before turning to these models, which focus on some of the more depressing features inherent in the development process, it needs to be said that Smith's vision of development as a cumulative interactive process based on the division of labour and increasing returns in industry lay effectively dormant until Allyn Young, an American economist, revived it in a profound but neglected article in 1928 entitled 'Increasing Returns and Economic Progress'. As Young (1928) observed:

> Adam Smith's famous theorem [that the division of labour depends on the extent of the market and the extent of the market depends on the division of labour] amounts to saying that the division of labour depends in large part on the division of labour. [But] this is more than mere tautology. It means that the counter forces which are continually defeating the forces which make for equilibrium are more pervasive and more deeply rooted than we commonly realise . . . Change becomes progressive and propagates itself in a cumulative way.

For Young, increasing returns are not simply confined to factors that raise productivity *within* individual industries, but are related to the output of *all* industries, which, he argued, must be viewed as an interrelated whole: what are now sometimes called **macroeconomies of scale**. For example, a larger market for product X may make it profitable to use more machinery in its production, which reduces the cost of X *and* the cost of the machinery, which then makes the use of machinery more profitable in other industries and so on. Under certain conditions, change will become progressive and propagate itself in a cumulative way; the precise conditions being increasing returns and an elastic demand for products so that as their relative price falls, proportionately more is bought. Take the example of steel and textiles, both of which are subject to increasing returns and are price elastic. As the supply of steel increases, its relative price (or exchange value) falls. If demand is price elastic, textile producers demand proportionately more steel, and offer proportionately more textiles in exchange. Textile production increases and its exchange value falls. If demand is price elastic, steel producers demand proportionately more textiles and so on. As Young said: 'under these circumstances there are no limits to the process of expansion except the limits beyond which demand is not elastic and returns do not increase'.

The process described above could not occur with diminishing returns activities with an inelastic price demand, which characterizes most primary products. No wonder rapid development tends to be associated with the process of industrialization. It is true to say, however, that Young's vision was also lost until the 1950s, when economists, such as Gunnar Myrdal, Albert Hirschman and Nicholas Kaldor, started to challenge equilibrium theory and develop non-equilibrium models of the growth and development process, in books such as *Economic Theory and Underdeveloped Regions* (Myrdal, 1957), *Strategy of Economic Development* (Hirschman, 1958), *Strategic Factors in Economic Development* (Kaldor, 1967) and *Economics without Equilibrium* (Kaldor, 1985). Kaldor used to joke that economics went wrong after Chapter 4, Book I of Smith's *Wealth of Nations*, when Smith abandoned the assumption of increasing returns in favour of constant returns, and the foundations for neoclassical general equilibrium theory were laid. In contrast, it is now Smith and Young's emphasis on increasing returns that lies at the heart of the new endogenous growth theory.

The classical pessimists

The prevailing classical view after Smith was very pessimistic about the process of economic development, focusing on the problems of rapid population growth and the effect exerted on the rate of profit in industry by rising food prices owing to diminishing returns and rising costs in agriculture. One of the foremost pessimists was **Thomas Malthus**, and it might be said that the ghost of Malthus still haunts many developing countries today with respect to his views on population. But there are two strands to Malthus's writing: his theory of population, and his focus on the importance for development of maintaining 'effective demand' – a concept later borrowed by Keynes, who acknowledged a debt to Malthus. In fact, Malthus was the only classical economist to emphasize the importance of demand for the determination of output – all others adhered to **Say's law**: that supply creates its *own* demand, so that the level and growth of output are a function of the supply of physical inputs alone. For Malthus, effective demand must grow in line with productive potential if profitability as the stimulus to investment is to be maintained, but there is nothing to guarantee this. Malthus focused on the savings of landlords and the possible imbalance between the supply of saving and the planned investment of capitalists, which might impede development. If landlord saving exceeded the amount that capitalists wished to borrow, Malthus suggested the taxation of landlords as one solution.

Malthus is best known, however, for *An Essay on the Principle of Population* ([1798]1983), in which he claimed that there is a 'constant tendency in all animated life to increase beyond the nourishment prepared for it'. According to Malthus, 'population goes on doubling itself every twenty five years, or increases in a geometrical ratio', whereas 'it may be fairly said … that the means of subsistence increase in an arithmetical ratio'. Taking the world as a whole, therefore, Malthus ([1798]1983) concluded:

> the human species would increase (if unchecked) as the numbers 1, 2, 4, 8, 16, 32, 64, 128, 256, and subsistence as 1, 2, 3, 4, 5, 6, 7, 8, 9. [This would mean that] in two centuries the population would be to the means of subsistence as 256 to 9; in three centuries as 4096 to 13, and in two thousand years the difference would be incalculable.

If food production only grows at an arithmetic rate, this implies, of course, diminishing returns to agriculture. The imbalance between population growth and growth of the food supply would lead to the per capita income of countries oscillating around the subsistence level, or being caught in what is now sometimes called a **low-level equilibrium trap** (see Chapter 11). Any increases in per capita income brought about by technical progress lead to more births, which then reduce

Thomas Malthus

Born 1766, Surrey, England. Died 1834. Professor of History and Political Economy at the East India Company College. Famous for *An Essay on the Principle of Population* (1798) predicting that population growth will outstrip food supply because of diminishing returns in agriculture. Some communities in developing countries still have Malthusian characteristics, and some 'environmentalists' predict a Malthusian world in the future.

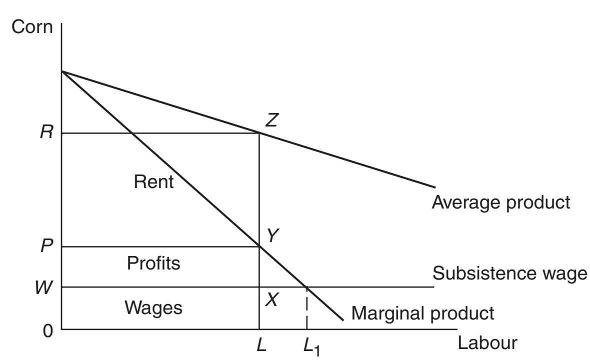
per capita income back to subsistence level. Early development models of the 'big push' were designed to lift economies from this trap. Malthus recognized certain checks to the process, which he divided into 'preventative' and 'positive' checks, some of which still operate today in certain countries. Preventive checks include sexual abstinence or the use of contraception, although Malthus was opposed to the latter. Where preventive checks are weak, positive checks take over in the form of pestilence, disease and famine. Malthus's solution to population growth was the 'postponement of marriage in a viceless society'.

While Malthusian economics may still have relevance in certain parts of Africa and Asia, Malthus's gloomy prognostications have not materialized for the world as a whole, because preventive checks have become stronger and food production has grown not at an arithmetic rate but at a rate faster than the growth of population (see Chapter 5). Technical progress in agriculture has offset diminishing returns. The underestimation of technical progress in agriculture has confounded all the classical pessimists.

David Ricardo was another of the great classical pessimists. In 1817 he published his *Principles of Political Economy and Taxation* (Ricardo, [1817]1992), in which he predicted that capitalist economies would end up in a stationary state, with no growth, also owing to diminishing returns in agriculture. In Ricardo's model, like Smith's, growth and development are a function of capital accumulation, and capital accumulation depends on reinvested profits. However, profits are squeezed between subsistence wages and the payment of rent to landlords, which increases as the price of food rises owing to diminishing returns to land and rising marginal costs. Ricardo thought of the economy as 'one big farm' in which food (or corn) and manufactures are consumed in fixed proportions, so that corn can be used as the unit of account. Figure 4.1 illustrates the model.

With the employment of *L* amount of labour, the total output is 0*RZL*. Rent is determined by the difference between the average and marginal product of labour working on the land and is given by the area *PRZY*. Wages are equal to 0*WXL*, and profit is the difference between rent and wages, equal to *WPYX*. As output increases and the marginal product of labour falls to the subsistence wage (L_1), profits disappear. In equilibrium, the rate of profit in agriculture must equal the rate of profit in industry. As the profit rate in agriculture falls, capital will shift to industry, causing the rate of profit to decline there. Profits are also squeezed because wages rise in terms of food. But for Ricardo, unlike Malthus, there was no problem of effective demand. Ricardo saw no limit to the amount of capital that could be employed because he accepted Say's law that supply creates its own demand. The villain of the piece is wages. Ricardo ([1817]1992) wrote: 'there is no limit to demand – no limit to the employment of capital while it yields any profit, and that

Figure 4.1 Ricardo's model of the economy



however abundant capital may become, there is no other adequate reason for a fall in profit but a rise in wages'. As profits fall to zero, capital accumulation ceases, heralding the stationary state. As Ricardo ([1817]1992) put it:

> a real rise of wages is necessarily followed by a real fall in profits, and, therefore, when the land of a country is brought to the highest state of cultivation, when more labour employed upon it will not yield in return more food than what is necessary to support the labourer so employed, that country has come to the limit of its increase both of capital and population.

As discussed in Chapter 5, Arthur Lewis's well-known development model, 'economic development with unlimited supplies of labour', is a classical Ricardian model, but wages are assumed to stay the same until disguised unemployment on the land is absorbed.

Given the central importance of capital accumulation in Ricardo's vision of economic progress, anything that reduces capital accumulation (including rises in wages) will slow economic growth. Ricardo was thus opposed to all forms of taxes, levies and tariffs on inputs into the productive system, including tariffs on imported food. Indeed, he believed that the importation of cheap food might delay the predicted stationary state indefinitely by holding down wages measured in terms of food:

> a country could go on for an indefinite time increasing its wealth and population, for the only obstacle to this increase would be the scarcity, and consequent high value, of food and other raw produce. Let these be supplied from abroad in exchange for manufactured goods, and it is difficult to say where the limit is at which you would cease to accumulate wealth and to derive profits from its employment (Ricardo [1817]1992).

It was for this reason that Ricardo campaigned for the abolition of the Corn Laws in Britain in the nineteenth century. These laws were eventually repealed in 1846 – to the benefit of industrialists but the detriment of domestic farmers. In developing countries today, governments often attempt to keep the price of agricultural goods artificially low in order to keep wages low (measured in terms of food). Doing this, however, raises another problem of reducing the incentive of farmers to produce. Determining the equilibrium terms of trade between agriculture and industry, which maximizes the output growth of both sectors, is a difficult empirical issue (see Chapter 5).

David Ricardo

Born 1772, London, England. Died 1823. Led a colourful life as politician, industrialist, speculator and economist. *Principles of Political Economy and Taxation* (1817) made him the foremost classical economist of the first half of the nineteenth century. Predicted that economies will end up in a stationary state through diminishing returns in agriculture reducing the rate of profit in industry. Campaigned for the repeal of the Corn Laws in England to make food cheaper. Most famous for his formulation of the doctrine of **comparative advantage**; one of the few non-trivial theorems in economics.

Finally, we turn to **Karl Marx**, famous for *Das Kapital* ([1867]1967), and his prediction of the collapse of capitalism. All members of the classical school agreed that the rate of profit on capital would fall as the economy grew, but they differed as to the reason for the fall. Adam Smith saw the decline in profits as the result of competition among capitalists. Ricardo saw the fall as the result of diminishing returns to land, and profits being squeezed between rent and wages, leading to a stationary state. For Marx, the economy does not grow forever, but the end comes not from a stationary state but from 'crises' associated with overproduction and social upheaval. But Marx's model bears many similarities to the other classical economists. The capitalist surplus is the source of capital accumulation and the principal mainspring of growth. Population growth responds to wages in Malthusian fashion, keeping wages down, and the rate of profit has a long-run tendency to fall.

Let us briefly consider Marx's model and his prediction of crisis. Gross output consists of three elements:

- Variable capital or the wage bill (v)
- Constant capital (c), that is, plant and machinery and the raw materials used in production
- Surplus value or profit (s).

The wages of labour are determined by the minimum subsistence level (what Marx called the cost of reproducing the working class), and surplus value (which only labour can create) is the difference between output per worker and the minimum wage per worker. The rate of surplus value, or what Marx called the 'degree of exploitation', is given by s/v. The rate of profit is given by the ratio of surplus value to total capital; that is:

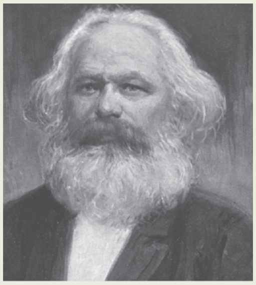
$$s/(v + c) = (s/v)/(1 + [c/v]) \tag{4.1}$$

where the ratio of constant capital to variable capital (c/v) is defined as the **organic composition of capital**. As techniques of production become more capital intensive, the organic composition rises through time, and as it does so, the rate of profit falls unless surplus value rises. While there is no limit to the rise in c/v, however, there is a limit to s/v. Marx foresaw no major problem as long as surplus labour exists to keep wages down, but he predicted that as capital accumulation continues, the **reserve army of labour**, as he called it, would disappear, driving wages up and profits down. The capitalist's response is then either to attempt to keep wages down, leading to the **immiserization of workers** and social conflict, or to substitute more capital for labour, which would worsen the problem by raising c/v.

For Marx ([1867]1967), the desire and necessity to invest are inherent in the psychological makeup of the capitalist: 'To accumulate, is to conquer the world of social wealth, to increase the mass of human beings exploited by him, and then to extend both the direct and indirect sway of the capitalist.' Thus the capitalist's motto is 'Accumulate, accumulate! That is Moses and the Prophets.' But as capital is substituted for labour, there is another problem: labour cannot consume all the goods produced, and a **realization crisis** is caused by the failure of effective demand. Capitalism eventually collapses through its own 'inner contradictions', and power passes to the working classes because fewer and fewer people benefit from capitalism. Capitalism is replaced by socialism, whereby workers own the means of production, distribution and exchange, and ultimately the state withers away.

Marx's analysis contains valuable insights into the functioning of capitalism, but his predictions, like those of his predecessors, have not materialized. There seem to be two basic reasons for this. The first is that there is a confusion in Marx's work between *money wages* and *real wages*. A rise in money wages as surplus labour disappears does not necessarily mean a rise in

Karl Marx



Born 1818, Trier, Germany. Died 1883. Settled in England in 1848 and supported by Friedrich Engels. *Das Kapital* (vol. 1, 1867, and three further volumes published posthumously), his major work, has inspired generations of left-wing thinkers in their critiques of the inequalities of capitalism. His prediction that capitalism would collapse through a decline in the rate of profit and the immiserization of workers leading to social revolution has not materialized, because in a growing economy with technical progress, there is no clash between real wages and the rate of profit.

real wages; but in any case, any rise in real wages could be offset by a rise in labour productivity, leaving the rate of profit unchanged. The second and related reason is that just as the other classical economists underestimated the rate of technical progress in agriculture as an offset to diminishing returns, so Marx underestimated the effect of technical progress in industry on the productivity of labour. It can be seen from equation (4.1) that even if c/v is rising, the rate of profit can remain unchanged if technical progress exceeds the rate of wage growth by the same amount. Technical progress also means there is no necessary clash between real wages and the rate of profit. Both can rise.

For nearly 60 years after Marx's death in 1883, growth and development theory lay effectively dormant, as economics came to be dominated by static neoclassical value theory under the influence of Alfred Marshall's *Principles of Economics* (1890). Marshall treated growth and development as more or less a 'natural' phenomenon; an evolutionary process akin to biological developments in the natural world. Modern growth theory started with the classic article by British economist Roy Harrod, 'An Essay in Dynamic Theory' (1939), and American economist Evsey Domar arrived at Harrod's result independently in 1947 (Domar, 1947), which led to the development of what is now called the Harrod–Domar growth model. The model has played a major part in thinking about development issues since the Second World War, and is still widely used in development planning (see Chapter 9).

The Harrod–Domar growth model

Harrod's original model is a dynamic extension of Keynes' static equilibrium analysis. In Keynes' *The General Theory of Employment, Interest and Money* (1936), the condition for income and output to be in equilibrium (in the closed economy) is that plans to invest equal plans to save (or injections into the circular flow of income should equal leakages). The question Harrod asked is: If changes in income induce investment, what must be the *rate of growth of income* for plans to invest to equal plans to save in order to ensure a *moving* equilibrium in a growing economy through time? Moreover, is there any guarantee that this required rate of growth will prevail? If not, what will happen? In static Keynesian theory, if equilibrium between saving and investment is disturbed, the economy corrects itself and a new equilibrium is achieved via the multiplier process. If growth equilibrium is disturbed, will it be self-correcting or self-aggravating? Moreover, will this equilibrium rate be equal to the maximum rate of growth that the economy is able to sustain, given the rate of growth of productive capacity? If not, what will happen? These are fundamental questions for the understanding of the growth performance of any country, be it developed or underdeveloped, and Harrod's place in the history of economic thought was guaranteed by the insight and simplicity with which he answered them.

To consider the questions posed, Harrod distinguished three different growth rates: what he called the **actual growth rate** (g), the **warranted growth rate** (g_w) and the **natural growth rate** (g_n). The actual growth rate is defined as:

$$g = s/c \tag{4.2}$$

where s is the ratio of savings to national income (S/Y) and c is the *actual* incremental capital–output ratio, that is, the ratio of extra capital accumulation or investment to the flow of output ($\Delta K/\Delta Y = I/\Delta Y$). This expression for the actual growth rate ($g = s/c$) is, by definition, true since it expresses the accounting identity that savings equals investment. We can see this if we substitute the expressions for s and c into equation (4.2) – that is, $s/c = (S/Y)/(I/\Delta Y) = \Delta Y/Y$, given $S = I$, where $\Delta Y/Y$ measures the growth of output.

We need more than a definitional equation, however, to know whether the actual growth rate will provide the basis for steady advance in the future, in the sense that it keeps plans to invest and plans to save in line with one another at full employment. This is where the concepts of the warranted rate of growth and the natural rate of growth become important.

Harrod (1939) defined the warranted rate of growth as:

> that rate of growth which, if it occurs, will leave all parties satisfied that they have produced neither more nor less than the right amount. Or, to state matters otherwise, it will put them into a frame of mind which will cause them to give such orders as will maintain the same rate of growth.

In other words, the warranted growth rate is the rate that induces just enough investment to match planned saving and therefore keeps capital fully employed (that is, there is no undercapacity or overcapacity), so that manufacturers are willing to carry on investment in the future at the same rate as in the past. How is this rate determined? Plans to save at any point in time are given by the Keynesian savings function:

$$S = sY \tag{4.3}$$

where s is the propensity to save. This gives the potential supply of investment goods. The demand for investment is given by the **acceleration principle** (or what Harrod calls 'the relation'), where c_r is the accelerator coefficient measured as the *required* amount of extra capital or investment to produce a unit flow of output at a given rate of interest, determined by technological conditions. Thus:

$$c_r = \Delta K_r/\Delta Y = I/\Delta Y \tag{4.4}$$

The demand for investment, given by the accelerator principle, is then:

$$I = c_r \Delta Y \tag{4.5}$$

For planned saving to equal planned investment, therefore, we have:

$$sY = c_r \Delta Y \tag{4.6}$$

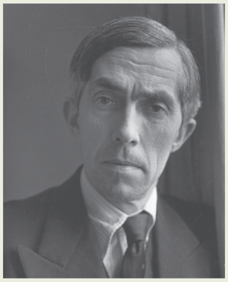
and the required rate of growth for a moving equilibrium through time is:

$$\Delta Y/Y = s/c_r = g_w \tag{4.7}$$

This is the warranted rate of growth, g_w. For dynamic equilibrium, output must grow at this rate. At this rate, expenditure on consumption goods will equal the production of consumption goods, and this is the only rate at which entrepreneurs will be satisfied with what they are doing, so that they do not revise their investment plans.

Now, suppose there is a departure from this equilibrium rate. What happens? The condition for equilibrium is that $g = g_w$ or, from equations (4.2) and (4.7), that $gc = g_w c_r$. First, suppose that the actual growth rate exceeds the warranted rate. It is easily seen that if $g > g_w$, then $c < c_r$, which means that actual investment falls below the level required to meet the increase in output. There will be a shortage of equipment, a depletion of stocks and an incentive to invest more. The actual growth rate will then depart even further from the warranted rate. Conversely, if the actual growth rate is less than the warranted rate, $g < g_w$, then $c > c_r$, and there will be a surplus of capital goods and investment will be discouraged, causing the actual growth rate to fall even further below the equilibrium rate. Thus, as Harrod points out, we have in the dynamic field a condition opposite to that in the static field. A departure from equilibrium, instead of being self-righting, will be self-aggravating. This is the short-term trade cycle problem in Harrod's growth model.

Roy Harrod



Born 1900, Norfolk, England. Died 1978. Spent all his academic career in Christ Church College, Oxford. One of Keynes' inner circle; in the late 1930s, made static Keynesian theory dynamic, thus pioneering modern growth theory. One of the most original economists of the twentieth century; made important contributions to the theory of the firm, international economics and economic dynamics. Authored a book on inductive logic as well as the first biography of Keynes.

American economist **Evsey Domar**, working independently of Harrod, arrived at Harrod's central conclusion, although by a slightly different route. Domar recognized that investment is a double-edged sword: it both increases demand via the multiplier, and increases supply via its effect on expanding capacity. The question Domar asked, therefore, was: What rate of growth of investment must prevail in order for supply to grow in line with demand (at full employment)? The crucial rate of growth of investment can be derived in the following way. A change in the level of investment increases demand by:

$$\Delta Y_d = \Delta I/s \tag{4.8}$$

and investment itself increases supply by:

$$\Delta Y_s = I\sigma \tag{4.9}$$

where σ is the productivity of capital or the flow of output per unit of investment ($\Delta Y/I$). For $\Delta Y_d = \Delta Y_s$, we must have:

$$\Delta I/s = I\sigma \tag{4.10}$$

or:

$$\Delta I/I = s\sigma \tag{4.11}$$

In other words, investment must grow at a rate equal to the product of the savings ratio and the productivity of capital. With a constant savings–investment ratio, this also implies output growth at the rate $s\sigma$. If $\sigma = 1/c_r$ (at full employment), then the Harrod–Domar result for equilibrium growth is the same.

But even if growth proceeds at the rate required for full utilization of the capital stock and a moving equilibrium through time, this still does not guarantee the full employment of labour, which depends on the natural rate of growth. The **natural growth rate** is derived from the identity $Y^* = L^*(Y/L)^*$, where Y^* is the potential level of output, L^* is the potential labour force and $(Y/L)^*$ is the potential level of labour productivity. Taking rates of change of the variables gives $y^* = l + \dot{q}$. The natural rate of growth (g_n) is therefore made up of two components: the growth of the potential labour force (l) and the growth of potential labour productivity ($\dot{q}$) (what Harrod called the rate of growth of the labour force in efficiency units) – both exogenously determined in the Harrod model.[1] The natural rate of growth plays an important role in Harrod's growth model in two ways. First, it defines the rate of growth of productive capacity or the long-run full employment equilibrium growth rate. Second, it sets the upper limit to the actual growth rate, which brings cumulative expansion in the Harrod (trade cycle) model to a sticky end. If $g > g_w$, g can continue to diverge from g_w only until it hits g_n, when all available labour has been completely absorbed: g cannot be greater than g_n in the long run. The long-run question for an economy, then, is the relation between g_w and g_n; that is, the relation between the growth of capital and the growth of the labour force (measured in efficiency units). With fixed coefficients of production, the full employment of labour clearly requires $g = g_n$. The full employment of labour *and* capital requires:

$$g = g_w = g_n \tag{4.12}$$

a state of affairs that the well-known Cambridge economist Joan Robinson once called a 'golden age' to emphasize its mythical nature, because there is nothing in the Harrod model that would automatically generate this happy coincidence.

Let us now consider what happens if the warranted growth rate diverges from the natural rate. If $g_w > g_n$, there will be a chronic tendency towards depression because the actual rate of growth will never be sufficient to stimulate investment demand to match the amount of saving at full employment equilibrium. There is too much capital and too much saving. This was the worry that economists had in the 1930s, particularly when it was predicted that the size of the population would fall in developed countries because the net reproduction rate had fallen below one (that is, females were not replacing themselves). If $g_w < g_n$, there will be a tendency towards demand inflation because there will be a tendency for the actual rate of growth to exceed that necessary to induce investment to match saving. Inflationary pressure, however, will be accompanied by growing unemployment of the structural variety, because the growth of capital falls short of the growth of the effective labour force and there is no change in the techniques of production.

Where do developing countries fit into this picture? In most developing countries, the natural growth rate exceeds the warranted rate. If the population growth is, say, 2% and labour productivity is growing at 3%, this gives a rate of growth of the labour force in efficiency units of 5%. If the net savings ratio is, say, 9% and the required incremental capital–output ratio is 3, this gives a warranted growth rate of 3%. This has two main consequences. First, it means that the effective labour force is growing faster than capital accumulation, which is part of the explanation for growing unemployment in developing countries. Second, it implies plans to invest greater than plans to save, and therefore inflationary pressure. If $g_n = 5\%$ and $c_r = 3$, there will be profitable investment opportunities for 15% saving, whereas actual saving is only 9%.

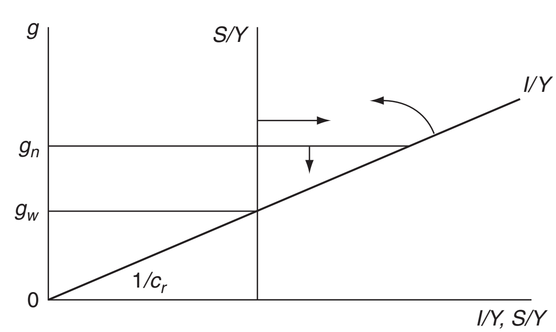
The simultaneous existence of inflation and high unemployment in developing countries is therefore not a paradox. It can easily be explained within the framework and assumptions of the Harrod growth model, as can a great deal of development policy. Given the inequality $g_n \neq g_w$, or $l + \dot{q} \neq s/c_r$, it can be seen that there are basically four ways in which g_n and g_w might be reconciled:

1. If the problem is $g_n > g_w$, the rate of growth of the labour force could be reduced. Measures to control population growth can be justified on these grounds, as a contribution to solving the problem of structural unemployment.
2. A reduction in the rate of growth of labour productivity would help, but this would reduce the growth of living standards of those in work. There is a clash here between employment and efficiency.
3. A rise in the savings ratio could narrow the gap. This is at the heart of monetary and fiscal reform in developing countries (see Chapter 13).
4. The natural and warranted growth rates might be brought into line by a reduction in the required capital–output ratio through the use of more labour-intensive techniques.

There is an active debate in developing countries over the appropriate choice of techniques, and whether developing countries could move towards the use of more labour-intensive techniques without impairing output and sacrificing saving (see Chapter 6). All these adjustment mechanisms are illustrated in Figure 4.2.

Growth is measured on the vertical axis, and the investment and savings ratios on the horizontal axis. Growth and the investment ratio are related through c_r – the required incremental capital–output ratio. The savings ratio is independent of the growth rate. Figure 4.2 depicts a situation in which the natural growth rate (g_n) exceeds the warranted growth rate (g_w). To equalize g_n and g_w, we can bring down g_n to g_w by measures to curb labour force growth; we can shift rightwards the S/Y curve through monetary and fiscal policies (and also by foreign borrowing) to raise g_w to g_n, or we can pivot the I/Y curve inwards by reducing c_r through the use of more labour-intensive techniques of production.

Figure 4.2 Adjustment of g_w and g_n



The Harrod framework is not only useful for understanding some of the development difficulties of developing countries, it is also useful for planning purposes. If a country sets a target rate of growth of, say, 5% per annum and the required capital–output ratio is 3, it knows it must save and invest 15% of GDP if the target growth rate is to be achieved. If domestic saving is less than 15% of GDP, there is an investment–savings gap to fill, which might be done by foreign borrowing (see Chapter 14).

At a theoretical level, there has been a great deal of discussion in the literature of whether *automatic* adjustment mechanisms might not come into play to reconcile the divergence between g_n and g_w. In the Harrod model, the parameters and variables that make up the model, l, $\dot{q}$, s and c_r, are all independently determined. Harrod himself recognized that, in the long run the savings ratio may not be fixed, but will adjust. Specifically, in periods of recession, savings may fall, and in periods of demand inflation, savings may rise. One way this may come about is through a change in the functional distribution of income between wages and profits. This is a possible mechanism of adjustment emphasized by the **post-Keynesian economists of Cambridge, England**, represented by Joan Robinson, Nicholas Kaldor, Richard Kahn, Luigi Pasinetti and others. If $g_w > g_n$, and there is a tendency towards depression, this will tend to reduce the share of profits in national income and increase the share of wages, so that if the propensity to save out of profits is higher than the propensity to save out of wages, this change in the distribution of income will lower the overall savings ratio and reduce g_w towards g_n. There is a limit, however, to which the share of profits can fall, given by the minimum rate of profit acceptable to entrepreneurs. Likewise, if $g_w < g_n$, and there is a tendency towards demand inflation, the share of profits in national income will tend to rise, increasing the overall savings ratio and raising g_w towards g_n. There is also a limit, however, to the rise in the profit share, set by the degree to which workers are willing to see their real wages reduced – what Joan Robinson called 'the inflation barrier' (see Chapter 13).

At the same time (the 1950s), in contrast to the Cambridge, England school of post-Keynesian economists, a formidable group of **economists in Cambridge, Massachusetts**, represented by Robert Solow, Paul Samuelson, Franco Modigliani and others, developed the so-called **neoclassical model of growth** as an attack on Harrod and the post-Keynesian school. They pointed out that the gloomy conclusions of Harrod concerning the possibility of achieving steady growth with full employment assume fixed coefficients of production, and that if the capital-labour ratio is allowed to vary, there is the possibility of equilibrium growth at the natural rate. In other words, if capital grows faster than labour ($g_w > g_n$), economies will move smoothly via the price mechanism to more capital-intensive techniques, and growth in the long run will proceed at the exogenously given natural rate. Conversely, if labour grows faster than capital ($g_n > g_w$), the wage rate will fall relative to the price of capital, economies will adopt more labour-intensive techniques, and again growth will proceed at the natural rate.

One central feature of this neoclassical model, which has come under sustained attack in recent years from new growth theory, is that investment does not matter for long-run growth. Any increase in the savings or investment ratio is offset by an increase in the capital–output ratio, because of diminishing returns to capital, leaving the long-run growth rate (at the natural rate) unchanged. The argument depends, however, on the productivity of capital falling (or c_r rising) as the capital–labour ratio rises. This is disputed by the new growth theorists. If there are mechanisms to prevent the productivity of capital from falling as investment increases, then investment does matter for long-run growth and growth is *endogenous* in this sense. But before we turn to new growth theory and the important new studies of the macrodeterminants of growth, we need to consider the assumptions and predictions of neoclassical growth theory, and see how it has been used empirically for understanding the sources of growth in developed and developing countries.

Neoclassical growth theory

There are three basic propositions of neoclassical growth theory:

1. In the long-run steady state, the growth of output is determined by the *rate of growth of the labour force in efficiency units*, that is, by the rate of growth of the labour force plus the rate of growth of labour productivity (exogenously given as in Harrod's natural rate of growth), and is independent of the ratio of saving and investment to GDP. This is so because a higher savings or investment ratio is offset by a higher capital–output ratio or lower productivity of capital, because of the neoclassical assumption of **diminishing returns to capital**.
2. The *level* of per capita income (PCY), however, *does* depend on the ratio of saving and investment to GDP. The level of PCY varies positively with the savings–investment ratio and negatively with the rate of growth of the population.
3. If there is an inverse relation across countries between the capital–labour ratio and the productivity of capital, and tastes (i.e. savings behaviour) and technology are the same across countries, poor countries with a small amount of capital per head should grow faster than rich countries with a lot of capital per head, leading to the *convergence* of per capita incomes and living standards across the world.

Let us now consider how these fundamental propositions are arrived at. The basic **neoclassical growth model** was first developed independently by Robert Solow and Trevor Swan in 1956 and has been very influential in the analysis of growth ever since – particularly the use of the aggregate production function, as we shall see. The model is based on three key assumptions (ignoring technical progress for the moment):

1. The labour force grows at a constant exogenous rate, l.
2. Output is a function of capital and labour: $Y = F(K, L)$; the production function relating output to inputs exhibits constant returns to scale, diminishing returns to individual factors of production, and has a unitary elasticity of substitution between factors (see later).
3. All saving is invested: $S = I = sY$; there is no independent investment function.

What the basic neoclassical growth model is designed to show is that an economy will tend towards a long-run equilibrium capital–labour ratio (k^*) at which output (or income) per head (q^*) is also in equilibrium, so that output, capital and labour all grow at the same rate, l. The model therefore predicts long-run growth equilibrium at the natural rate.

Robert Solow

Born 1924, New York City, USA. Spent all his academic career at the Massachusetts Institute of Technology. Famous for his pioneering work on the theory of economic growth and technical change with his 1956 paper 'A Contribution to the Theory of Economic Growth', which challenged the rigid Harrod model of long-run disequilibrium growth. Also made important contributions to mathematical economics, capital theory and macroeconomics. Received the Nobel Prize for Economics, 1987.

The most commonly used neoclassical production function with constant returns to scale is the so-called **Cobb–Douglas production function**:

$$Y = bK^{\alpha}L^{1-\alpha} \tag{4.13}$$

where α is the elasticity of output with respect to capital, $1-\alpha$ is the elasticity of output with respect to labour, and $\alpha + (1 - \alpha) = 1$, that is, a 1% increase in K and L will lead to a 1% increase in Y, which is what is meant by output exhibiting constant returns to scale.

Equation (4.13) can also be written in 'labour-intensive' form by dividing both sides of the equation by L to give output per head as a function of capital per head:

$$\frac{Y}{L} = \frac{bK^{\alpha}L^{1-\alpha}}{L} = b\left(\frac{K}{L}\right)^{\alpha} \tag{4.14}$$

or, for short:

$$q = b\,(k)^{\alpha} \tag{4.15}$$

This is the 'labour-intensive' form of the neoclassical production function, and can be drawn as in Figure 4.3. The diminishing slope of the function represents the diminishing marginal product of capital.

Now impose a ray from the origin along which the rate of growth of capital is equal to the rate of growth of labour, so that the capital–labour ratio is constant. This is given by:

$$q = (l/s)k \qquad (4.16)^{2}$$

where s is the savings ratio. This straight line from the origin with slope l/s shows the level of q that will keep capital per head constant, and the level of k that will keep output per head constant – given the rate of growth of the labour force, l. Superimposing equation (4.16) on Figure 4.3 gives Figure 4.4.

The slope of the ray from the origin to any point on the production function determines the capital–output ratio at that point. It is clear from Figure 4.4 that only where these two lines cross is an equilibrium capital–labour ratio (k^*) and output per head (q^*) defined. To the left of k^* (at k_1), where $q > (l/s)k$, q is greater than necessary to maintain k constant; that is, there is too much saving and capital accumulation relative to the growth of the labour force, and steady growth requires more capital-intensive techniques. There will be a movement from k_1 towards k^*. The capital–output ratio adjusts to bring the rate of growth of capital and labour (or the warranted and natural growth rates) into line. Similarly, to the right of k^* (at k_2), where $q < (l/s)k$, q is less

Figure 4.3 The 'labour-intensive' form of the neoclassical production function

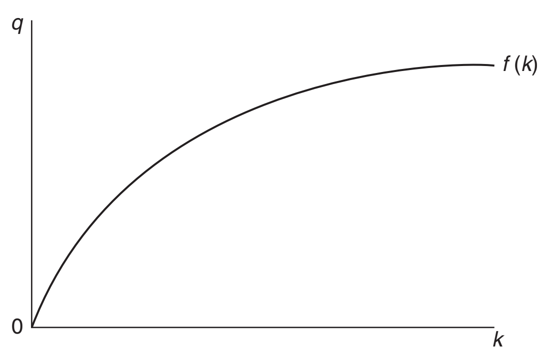
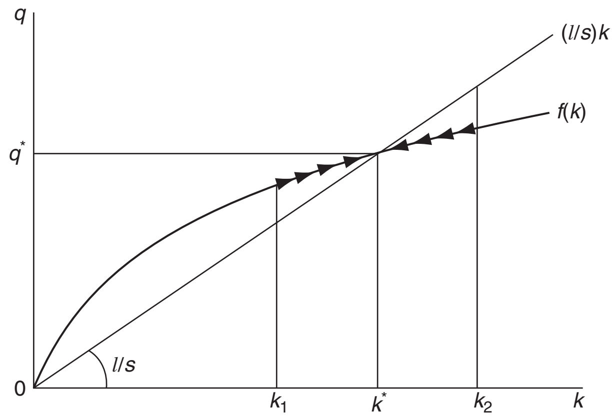


Figure 4.4 Equilibrium capital–labour ratio and output per head



than necessary to maintain k constant; there is too little saving and capital accumulation to keep pace with the rate of growth of the labour force, and steady growth requires more labour-intensive techniques. There will be a movement from k_2 towards k^*. Again, the capital–output ratio adjusts if there is a spectrum of techniques to choose from.

When k reaches an equilibrium, q also reaches an equilibrium, so output must be growing as fast as labour. Thus, output, labour and capital must all be growing at the same rate, l, the natural rate of growth, with the capital–output ratio constant. This is the neoclassical story.

We can now see what happens if there is an increase in the ratio of savings and investment to national income (s). If s rises, this lowers the slope of the l/s line in Figure 4.4, *which increases the equilibrium level of per capita income and the capital–labour ratio, but leaves the equilibrium growth rate unchanged*. This demonstrates formally the first two basic propositions of neoclassical growth theory. The reason a higher savings or investment ratio does not affect the long-run equilibrium growth rate is that a higher savings–investment ratio is ultimately offset by a higher capital–output ratio. The capital–output ratio adjusts 'passively' to keep the growth of capital in line with the growth of the labour force.

None of these conclusions is altered if technical progress is introduced into the model. If technical progress augments the productivity of labour only (so-called Harrod neutral technical progress, which leaves the capital–output ratio unchanged), the *effective* labour force now grows at the rate $l + \dot{q}$, where $\dot{q}$ is the rate of growth of labour productivity. Equilibrium is now defined in terms of output per effective worker and capital per effective worker. Equilibrium capital per worker requires $\Delta K/K = l + \dot{q}$, and equilibrium output per effective worker requires $\Delta Y/Y = l + \dot{q}$, so that capital per head and output per head grow at the same rate $\dot{q}$; that is, by the rate of Harrod neutral technical progress:[3]

$$\Delta Y/Y - l = \Delta K/K - l = \dot{q} \tag{4.17}$$

This is, of course, consistent with what we observe in the real world – output and capital grow faster than the rate of growth of the labour force. But a rise in the ratio of savings and investment to GDP still has no effect on the equilibrium growth of output, unless, of course, a higher level of investment raises the rate of growth of labour-augmenting technical progress, but this is ruled out by assumption in the neoclassical model, because technical progress is assumed to be exogenously determined.

It now only remains to demonstrate the third basic proposition of neoclassical growth theory: that poor countries should grow faster than rich countries, leading to the convergence of per capita incomes because poor countries with a low ratio of capital to labour will have a higher productivity of capital (or lower capital–output ratio). The capital–output ratio may be written as:

$$\frac{K}{Y} = \frac{K}{L} \cdot \frac{L}{Y} \tag{4.18}$$

Given diminishing returns to capital (so that Y/L does not rise in the same proportion as K/L), it can be seen that a higher K/L ratio will be associated with a higher K/Y ratio. This means that if the ratio of savings and investment to GDP is the same across countries, capital-rich countries should grow slower than capital-poor countries. Note, however, that if there are *not* diminishing returns to capital, but, say, constant returns to capital, a higher capital–labour ratio will be exactly offset by a higher output–labour ratio, and the capital–output ratio will not be higher in capital-rich countries than in capital-poor countries, so convergence is not to be expected. If there are not diminishing returns to capital, this also means that the capital–output ratio will not rise as more investment takes place, and therefore **the ratio of saving and investment to GDP does matter for growth. In this sense, growth is endogenously determined**; it is not simply exogenously determined by the rate of growth of the labour force and technical progress. This is the starting point for the new (endogenous) growth theory, which seeks to explain why, in practice, living standards in the world economy have not converged (see Chapter 2), contrary to the predictions of neoclassical theory. The explanation offered by the new growth theory is that there are forces at work that prevent the marginal product of capital from falling (and the capital–output ratio from rising) as more investment takes place as countries get richer. Before turning to the new growth theory, however, let us first consider how the neoclassical production function can be used to analyse the sources of growth. This requires us to look more closely at the concept of the production function and the properties of the Cobb–Douglas production function, which is still widely used in the analysis of growth in developed and developing countries.

The production function approach to the analysis of growth

We have already seen that there are several ways in which the growth of income or output of a country may be expressed, but frequently they consist of identities that tell us very little about the causes or sources of growth. For example, in the Harrod–Domar model, growth can be expressed as the product of the ratio of investment to GDP and the productivity of investment, so that, by definition, slow growth is the product of a low investment ratio and/or a low productivity of capital. By itself, however, this does not further much our understanding of the growth process in different countries. Why do some countries save and invest more than others, and why does the productivity of capital differ? Likewise, we have seen that the growth of output can be expressed as the sum of the rate of growth of the labour force and the rate of growth of labour productivity. By definition, slow growth is attributable to a slow rate of growth of the labour force and/or a slow rate of growth of labour productivity. Again, why does growth in labour productivity differ between countries? Is it because of differences in capital accumulation, or differences in technical progress, broadly defined to include such factors as improvements in the quality of labour, improvements in the quality of capital, economies of scale, advances in knowledge, a better organization of capital and labour in the productive process and so on? Growth identities cannot distinguish between such competing hypotheses.

The production function approach to the analysis of growth is a response to this challenge. It takes the concept of the aggregate production function and attempts to disaggregate the sources of growth into the contribution of labour, capital, technical progress and any other variable included in the production function that is thought to influence the growth process. In this sense, it is a very versatile approach. It is, however, a **supply-oriented** approach. It does not tell us *why* the growth of capital, labour, technical progress and so on differs over time or between countries. The sources of growth are treated as *exogenous*. In practice, however, the supply of most resources to an economic system is endogenous, responding to the demand for them. Capital is a produced means of production and comes from the growth of output itself; labour is very elastic in supply from internal and external sources (migration), and technical progress is itself partly dependent on the growth of output arising from static and dynamic returns to scale.

Thus, while the production function approach can disaggregate any measured growth rate into various constituent growth-inducing sources, and can 'explain' growth rate differences in terms of these sources, it cannot answer the more fundamental question of why labour supply, capital accumulation and technical progress grow at different rates in different countries. The answer to this question must lie in differences in the strength of *demand* for countries' products, which, in the early stages of development, depends largely on the prosperity of agriculture (see Chapter 5), and in the later stages of development depends largely on the country's export performance relative to its import propensity (see Chapter 16).

Having said this, the production function approach can provide a useful **growth accounting** exercise, which is in fact widely used. Apart from deciding which determinants of growth to specify in the production function, and accurately measuring the independent variables, the main problem is a methodological one of fitting the appropriate production function to the data; that is, specifying the function relating output to inputs.

The production function

A desirable property of any macroeconomic hypothesis, apart from being consistent with the observed facts, is that it should be consistent with and derivable from microeconomic theory. What we call the **production function approach to the analysis of growth** in the aggregate possesses, in part, this desirable property, in that it borrows the concept of the production function from the theory of the firm. Just as it can be said that, for a firm, output is a function of the factors of production – land, labour, capital and the level of technology (or factor efficiency) – so aggregate output can be written as a function of factor inputs and the prevailing technology:

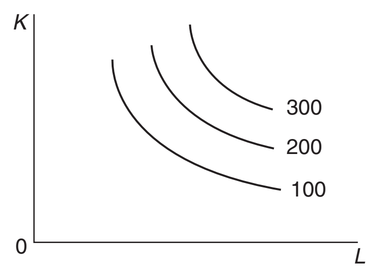
$$Y = f(R, K, L, T) \tag{4.19}$$

where R is land, K is capital, L is labour and T is technology.[4]

The question is how to separate empirically the contribution to output growth of the growth of factor inputs from other factors that can lead to higher output, included in T, such as economies of scale (due to technical change and increases in factor supplies), improvements in the quality of factor inputs, advances in knowledge, better organization of factors and so on. The task is to fit an appropriate, correctly specified production function that, if possible, will not only separate the contribution of factor inputs to growth from the contribution of increases in output per unit of inputs (increases in 'total' factor productivity), but will also distinguish between some of the factors that may contribute to increases in the productivity of factors, such as education, improvements in the quality of capital, and economies of scale.

Before going on to discuss the types of function that may be employed, let us examine in a little more detail the properties of a production function. We have established so far that the aggregate production function expresses the functional relation between aggregate output and the stock of inputs. If land is subsumed into capital, and technology is held constant, we are left with two factors, and the production function may be drawn on a two-dimensional diagram, as in Figure 4.5. Capital (K) is measured on the vertical axis and labour (L) on the horizontal axis, and each function represents a given level of output that can be produced with different combinations of capital and labour. The functions slope negatively from left to right on the assumption that marginal additions of either factor will increase total output – that is, factors have positive marginal products – and they are drawn convex to the origin on the assumption that factors have a diminishing marginal productivity as their supply increases, so that if one unit is withdrawn, it needs to be substituted by more and more of the other factor to keep output constant. The position of the functions broadly reflects the level of technology. The more 'advanced' the technology, the greater the level of output per unit of total inputs, and the closer to the origin will be the production function representing a *given* output.

Figure 4.5 The production function



From the simple production function diagram, it is easy to see how output may increase. First, there may be a physical increase in factor inputs, L and K, permitting a higher level of production. Either or both factors may increase. If only one factor increases, the movement to a higher production function will involve a change in the combination of factors, and output will not be able to increase forever, because ultimately the marginal product of the variable factor will fall to zero. This is illustrated in Figure 4.6, where, with a given stock of capital $0K_1$, output cannot increase beyond 300 with increases in the supply of labour ($0L_1$, $0L_2$ and so on) beyond the limit indicated. The diminishing productivity of the variable factor, labour, with

Figure 4.6 Production function diagram

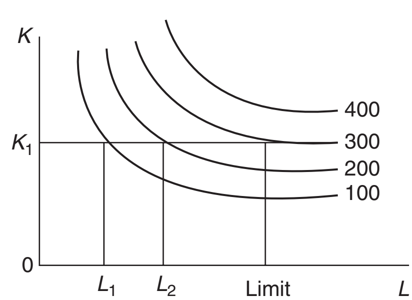
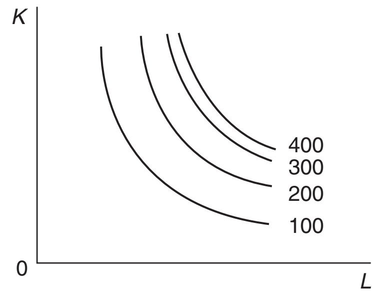


Figure 4.7 The effect of increasing returns



capital fixed, is shown by the flatter and flatter slope of the production functions at successive points, L_1, L_2, until, at the limit, the production function is horizontal and the marginal product of labour is zero.

If both factors increase in supply, however, there is no reason why output should not go on increasing indefinitely. In fact, if both factors increase in supply, there is a possibility that production may be subject to increasing returns, such that output rises more than proportionately to the increase in combined inputs. If this is the case, output per unit of total inputs will increase and the production functions representing equal additional amounts of production, for example 100, 200, 300 and so on, must be drawn closer and closer together, as in Figure 4.7.

In the opposite case of decreasing returns, the functions would be drawn further and further apart. Finally, in the case of production subject to constant returns, the functions would be drawn equidistant from one another.

Increasing returns may also result from advances in technology, irrespective of increases in factor supplies. These are called **technological economies of scale**. In this case, increases in output per unit of input would have to be represented on a production function diagram either by relabelling the functions or relabelling the axes. That is, either the same amount of factor inputs, measured on the axes, would have to be shown to be producing a higher output than before, or the same output could be shown to be produced by lesser amounts of inputs. If the functions are relabelled and not the axes, this is tantamount to a shift in all the production functions towards the origin. Shifts in the production function towards the origin are implied by all forms of technical progress or any factor that increases the productivity of the physical inputs.

In short, three broad sources of growth can be distinguished using the production function framework:

1. Increases in factor supplies
2. Increasing returns
3. Technical progress, interpreted in the wide sense of anything that increases the productivity of factors other than increasing returns.

The Cobb–Douglas production function

The production function most commonly fitted to aggregate data to distinguish empirically between these three broad sources of growth has been the unconstrained form of the **Cobb–Douglas production function**, named after its two American originators, Charles Cobb (a mathematician) and Paul Douglas (an economist), who pioneered research in the area of applied economic growth in the 1920s and 1930s (Cobb and Douglas, 1928). The Cobb–Douglas function may be written as:

$$Y_t = T_t K_t^{\alpha} L_t^{\beta} \tag{4.20}$$

where Y_t is real output at time, t, T_t is an index of technology, or 'total' productivity, K_t is an index of the capital stock, or capital services, at constant prices, L_t is an index of labour input (preferably man-hours), α is the partial elasticity (responsiveness) of output with respect to capital (holding labour constant), and β is the partial elasticity of output with respect to labour (holding capital constant).

It is assumed that changes in technology are exogenous and independent of changes in factor inputs, and that the effect of technical progress is neutral on the factor intensity of production (see Chapter 6 for a definition of neutral technical progress). T_t, α and β are constants to be estimated empirically if the function is unconstrained. If α and β are assigned values in advance of the use of the function for estimating purposes, the function is said to be constrained. Normally, α and β will be less than unity on the assumption of diminishing marginal productivity of factors. The sum of the partial elasticities of output with respect to the factors of production gives the scale of returns, or the degree of homogeneity, of the function: $\alpha + \beta = 1$ represents constant returns, $\alpha + \beta > 1$ represents increasing returns, and $\alpha + \beta < 1$ represents decreasing returns, and the function is said to be homogeneous of degree one, greater than one, and less than one, respectively.

If α and β are not estimated empirically but are assumed to sum to unity, in which case the function will be constrained to constant returns, then increasing or decreasing returns will be reflected in the value of T_t, which is the index of total factor productivity. The existence of increasing returns would bias the value of T_t upwards, and decreasing returns would bias the value of T_t downwards. These points are made because, in practice, the Cobb–Douglas function is often employed in this constrained form with the sum of α and β put equal to unity. Then values are assigned to α and β according to the share of capital and labour in the national income. The underlying assumption is the perfectly competitive one that if production is subject to constant returns and factors are paid the value of their marginal products, then factor shares will reflect the elasticity of output with respect to each factor.[5]

In order to use equation (4.20) to separate the influence of the three broad sources of growth mentioned earlier, we must first make it operational by transforming it into *rate of growth* form. This can be done by taking logarithms of the variables and differentiating with respect to time, which gives:[6]

$$\frac{d \log Y_t}{dt} = \frac{d \log T_t}{dt} + \alpha \frac{d \log K_t}{dt} + \beta \frac{d \log L_t}{dt} \tag{4.21}$$

or:

$$\frac{dY}{dt} \times \frac{1}{Y} = \left(\frac{dT}{dt} \times \frac{1}{T}\right) + \alpha\left(\frac{dK}{dt} \times \frac{1}{K}\right) + \beta\left(\frac{dL}{dt} \times \frac{1}{L}\right)$$

The above equations are in continuous time. The discrete approximation, taking annual rates of change of the variables, may be written as:

$$r_Y = r_T + \alpha r_K + \beta r_L \tag{4.22}$$

where r_Y is the annual rate of growth of output per time period, r_T is the annual rate of growth of total productivity, or technical progress, r_K is the annual rate of growth of capital, r_L is the annual rate of growth of labour, and α and β are the partial elasticities of output with respect to capital and labour, respectively, as before.

In other words, equation (4.22) says that the rate of growth of output is equal to the sum of the rate of growth of 'total' factor productivity, the rate of growth of capital weighted by the partial elasticity of output with respect to capital and the rate of growth of labour weighted by the partial elasticity of output with respect to labour. With knowledge of r_Y, r_K, r_L, α and β, it becomes possible as a first step to separate out the contribution of factor inputs to growth from increases in output per unit of inputs represented by r_T. Now let us give an illustrative example. Suppose $r_Y = 5\%$ per annum, $r_K = 5\%$ per annum, $r_L = 1\%$ per annum, $\alpha = 0.25$ and $\beta = 0.75$ (decided on the basis of factor shares). Substituting in equation (4.22), we have:

$$5.0 = r_T + 0.25\,(5.0) + 0.75\,(1.0) \tag{4.23}$$

The contribution of capital to measured growth is 0.25 (5.0) = 1.25 percentage points; the contribution of labour is 0.75 (1.0) = 0.75 percentage points; and r_T is left as a residual with a contribution of 3.0 percentage points. If α and β were estimated empirically and there happened to be increasing returns ($\alpha + \beta > 1$), the significance of the factor contribution would be enhanced and r_T would be smaller.

On the assumption of constant returns to scale, the production function can also be estimated in its so-called labour-intensive form to analyse the **growth of output per head** (see equation (4.14)). If we subtract r_L from both sides of equation (4.22) and assume $\alpha + \beta = 1$, so that $\beta = 1 - \alpha$, we get:

$$r_Y - r_L = r_T + \alpha\,(r_K - r_L) \tag{4.24}$$

which means that the rate of growth of output per head (or labour productivity) is equal to the sum of the rate of growth of total productivity plus the rate of growth of capital per head times the elasticity of output with respect to capital. Taking the illustrative figures above, if $r_Y = 5\%$ and $r_L = 1\%$, then the rate of labour productivity growth is 4%. Therefore:

$$4.0 = r_T + 0.25\,(5.0 - 1.0) \tag{4.25}$$

The contribution of capital per head (capital deepening) to productivity growth is 1 percentage point, leaving r_T with a contribution of 3 percentage points (as before).

Although r_T has been variously called 'technical progress', 'advances in knowledge' and so on, it is, by definition, that portion of the growth of output not attributable to increases in the factors of production, and includes the effects not only of the multifarious factors that increase the productivity of labour and capital but also any measurement errors in the capital and labour input series. Perhaps r_T is best described as a residual, or, perhaps more appropriately still, a 'coefficient of ignorance' if the analysis proceeds no further.

One important component of r_T, which can be considered the result of measurement errors, is likely to be the effect of **resource shifts** from less productive to more productive activities. Since the analysis is aggregative, there is bound to be a confounding of changes in actual output with changes in the composition of output unless the weights used for aggregating inputs are continually revised. Resource shifts from agriculture to industry can be expected to figure prominently in any production function study of developing countries, as they do for studies of many advanced economies.

Limitations of the Cobb–Douglas function

Before considering some of the results of applying the Cobb–Douglas function to empirical data, we must briefly mention some of its limitations. Its use has come under attack on four main counts. First, since only one combination of factor inputs can be observed at any one time, there is an identification problem in attempting to distinguish shifts in the function (technical progress) from movements along the function (changes in factor intensity), unless the assumption of neutral technical progress is made. But technical progress may not be neutral and therefore the effects of technical progress and changing factor intensity become confused, biasing the results of the contribution of factor inputs and technical progress to growth.

Second, the assumption that technical progress is independent of increases in factor inputs can been questioned. This is not a specification error of the function itself, however, and the Cobb–Douglas function can be used, making technical progress a function of the rate of growth of inputs – so-called 'endogenous models of technical progress'.

Third, the Cobb–Douglas function possesses the restrictive property of constant unitary elasticity of substitution between factors, whatever the factor intensity.[7] The assumption of constant elasticity means that the function cannot represent a change in the ease of substitution between capital and labour. The assumption of unitary elasticity may be serious if the elasticity of substitution of factors differs significantly from unity and there are wide discrepancies in the growth rate of factors. For example, if the elasticity of substitution between capital and labour is significantly less than unity, and capital grows faster than labour, this will result in an overestimate of the contribution of capital to growth and an underestimate of the role of other factors. The intuitive explanation of this bias is that the smaller the elasticity of substitution, the more difficult it is in practice to obtain increased output just by increasing one factor, because diminishing returns set in strongly. By assuming the elasticity is higher than it is, the importance of the fastest growing factor is exaggerated. If elasticity is high, diminishing returns are not a problem, and if both capital and labour expand at the same rate, growth is obviously independent of the elasticity of substitution.[8]

A final criticism relates to the measurement of output and inputs. What, argue some, is the meaning of a function that aggregates so many heterogeneous items; in particular, what is the meaning of an aggregation of capital goods built at different times, at different costs and with varying productivities? How are such capital goods to be equated and added in an aggregate measure of capital?

By and large, most of the above-mentioned criticisms are theoretical worries, the practical significance of which is hard to determine. Studies of the nature of technical progress, at least in advanced countries, suggest that the assumption of neutrality is a fair working hypothesis. The fact that technical progress may be dependent on factor accumulation can be accommodated within the Cobb–Douglas framework. Capital and labour would have to grow at very different rates for the value of the elasticity of substitution to matter very much, but, in any case, studies show that it is quite close to unity. Finally, although the aggregation of heterogeneous outputs and inputs can present severe problems, especially the aggregation of capital, which cannot be measured directly in physical units, there are techniques of aggregation available that various studies have used with some success.

Application of the Cobb–Douglas function

What have been the results of applying the Cobb–Douglas function to empirical data? First, let us consider its application in developed countries and consider the conclusions that emerge. We can start with the pioneering work of Cobb and Douglas themselves. Ironically, the Cobb–Douglas function, as first conceived, was not intended as a device for distinguishing the sources of growth but as a test of neoclassical marginal productivity theory; that is, to see whether elasticities of output with respect to labour and capital corresponded to the shares of factors of production in national income. Douglas had observed that the output curve for US manufacturing industry for the period 1899–1922 lay consistently between the two curves for the factors of production, and he suggested to his mathematician friend Cobb that they should seek to develop a formula that could measure the relative effect of labour and capital on the growth of output over the period in question. This story is described by Douglas (1948 p. 20) in his fascinating review article 'Are There Laws of Production?' As an insight into the inductive method, the relevant passage is worth quoting in full:

> Having computed indexes for American manufacturing of the number of workers employed by years from 1899 to 1922 as well as indexes of the amounts of fixed capital in manufacturing deflated to dollars of approximately constant purchasing power, and then plotting these on a log scale, together with the Day index of physical production for manufacturing, I observed that the product curve lay consistently between the two curves for the factors of production and tended to be approximately one-quarter of the relative distance between the curve of the index for labour, which showed the least increase in the period, and that of the index of capital which showed the most. I suggested to my friend Charles Cobb that we seek to develop a formula which could measure the relative effect of labour and capital upon product during this period. At his suggestion the sum of the exponents was tentatively made equal to unity in the formula $Y = TK^{\alpha}L^{1-\alpha}$ [our notation] . . . The fact that on the basis of fairly wide studies there is an appreciable degree of uniformity, and that the sum of the exponents approximates to unity, fairly clearly suggests that there are laws of production which can be approximated by inductive studies and that we are at least approaching them.

The estimated function derived was $Y = 1.01K^{0.25}L^{0.75}$, which lent support to the neoclassical model of constant returns and marginal product pricing. There was no discussion of the relative importance of factors of production and the T variable in accounting for measured growth. It was not until Abramovitz (1956) and Solow (1957) showed that 80–90% of the growth of output per head in the US economy in the first half of the twentieth century could not be accounted for by increases in capital per head that the production function started to be used in earnest as a technique in the applied economics of growth. Abramovitz (1956, p. 11) remarked that:

> This result is surprising in the lop-sided importance which it appears to give to productivity increase and it should be, in a sense, sobering, if not discouraging to students of economic growth. Since we know little about the causes of productivity increase, the indicated importance of this element may be taken to be some sort of measure of our ignorance about the causes of economic growth in the USA, and some sort of indication of where we need to concentrate our attention.

Abramovitz's findings were supported by Solow, who found, when examining the data for the non-farm sector of the US economy for the period 1919–57, that approximately 90% of the growth of output per head could not be accounted for by increases in capital per head; that is, using the notation in equation (4.24):

$$r_T/(r_Y - r_L) = 0.90 \tag{4.26}$$

The findings of Abramovitz and Solow disturbed economists brought up in the belief that investment and capital accumulation played a crucial role in the growth process. Even allowing for the statistical difficulties of computing a series of the capital stock, and the limitations of the function applied to the data (for example, the assumption of constant returns and neutral technical progress, plus the high degree of aggregation), it was difficult to escape from the conclusion that the growth of the capital stock was of relatively minor importance in accounting for the growth of total output.

It would not be misleading to say that much of the subsequent research effort in this field of growth was designed (even before the advent of new growth theory) to reverse this conclusion, or rather to 'assign back' to the factors of production sources of growth that make up the residual factor but are interrelated with, or dependent on, the growth of factor inputs. Work has proceeded on two fronts. On the one hand, attempts have been made to disaggregate the residual factor, measuring factor inputs in the conventional way; on the other hand, attempts have been made to adjust the labour and capital input series for such things as changes in the *quality* of factors and their composition, so that much more measured growth is seen to be attributable to increases in factor inputs in the first place. For example, the labour input series has been adjusted for improvements in its quality due to the growth of education, and for changes in its composition due to age/sex shifts. Likewise, the capital stock series has been adjusted to reflect changes in its composition and, more importantly, to allow for the fact that new additions to the capital stock in any line of production are likely to be more productive than the existing capital stock as a result of technical advance. This is the notion of **embodied** or **endogenous technical change** as opposed to the exogenous technical change assumption of the original Cobb–Douglas function, which assumes that all vintages of capital share equally in technical progress.

A distinction is made, therefore, between embodied and disembodied technical progress – 'embodied technical progress' refers to technical improvements that can only be introduced into the productive system by new investment, and 'disembodied technical progress' is exogenous and not dependent on capital accumulation. There are several ways in which embodied technical progress can be isolated from the residual factor by appropriate adjustments to the capital stock series to reflect the greater productivity of the latest investments. The net result is to enhance the role of capital accumulation in the growth process.

Efforts have also been made to overcome one of the problems associated with the aggregation of outputs by taking explicit account of shifts of labour and capital from low-productivity to high-productivity sectors. This, too, reduces the significance of the residual factor and makes the role of labour and capital in the growth process look correspondingly more important.

Since Abramovitz and Solow reported their findings in 1956 and 1957, a substantial body of empirical evidence relating to the sources of growth has accumulated, experimenting with different specifications of the aggregate production function. Unfortunately, it is not systematic. The time periods taken, the data used, the sectors of the economy examined, and the methodology employed all vary within and between countries.

Until recently, most of the evidence available pertained to fairly advanced economies and it is largely from this evidence, wisely or not, that conclusions have been drawn on development strategy for developing countries. Research in developing countries has been hampered by a shortage of reliable empirical data and perhaps an even greater suspicion of the aggregate production function, and its implicit assumptions, than in developed countries. The assumption that factor shares can be used as weights to measure the relative contribution of labour and capital to growth is probably more dubious in developing countries than in developed countries. First, the price of labour almost certainly exceeds its marginal product, while the price of capital falls short of it so that the share of income going to labour exceeds the elasticity of output with respect to labour and the share of income going to capital understates the elasticity of output with respect to capital. Second, the aggregation of inputs and outputs is generally more difficult, and there are greater problems of resource underutilization to contend with. The recent past, however, has witnessed a number of production function studies for developing countries.

Production function studies of developing countries

Two of the early production function studies of the sources of growth in developing countries are by Maddison (1970) and Robinson (1971) (surveyed, with others, by Nadiri, 1972). More recent studies include the World Bank (1991), Young (1995), Hu and Khan (1997), Felipe (1999), Senhadji (2000), and Sala-í-Martin (1997) who surveys other studies. Let us consider these studies and bring out their major conclusions, especially any important contrasts with the conclusions from studies of developed countries.

The major conclusions of the early production function studies of developing countries were:

- Capital accumulation is more important as a source of growth than total productivity growth, and more important than in developed countries.
- Improvements in the *quality* of labour are important through better health, nutrition and education.
- Resource shifts are not as important as might have been expected, perhaps due to the general surplus of labour in developing countries and the low capacity to absorb labour into productive employment in the industrial sector.

Now let us turn to the more recent studies. Table 4.1 provides the results of a World Bank study for the period 1960–87, showing the contribution of factor inputs and total productivity growth to the growth of output in various continents. It is clear that the major source of growth is not productivity growth, but the growth of inputs themselves.

Young (1995) has used the production function model to debunk the idea that there has been a 'growth miracle' in the four East Asian countries of Hong Kong, Singapore, South Korea and Taiwan (the so-called 'tigers' or 'little dragons'). Young uses the production function approach and shows that while the growth of output was spectacular over the 1966–90 period, most of the growth can be accounted for by the rapid growth of factor inputs and there was nothing abnormal about the growth of total factor productivity. Table 4.2 presents the figures. Young describes such calculations as 'the tyranny of numbers', by which he means that there is nothing special to explain. On the basis of Young's calculations, Krugman (1994) has described the 'Asian miracle'

Table 4.1 Contribution of factor inputs and total productivity growth to economic growth in 68 developing countries, 1960–87

	GDP growth (% p.a.)	Contribution of labour	Contribution of capital	Total factor productivity
Africa	3.3	1.0	2.3	0.0
East Asia	6.8	1.1	3.8	1.9
Europe, Middle East and North Africa	5.0	0.7	2.9	1.4
Latin America	3.6	1.2	2.4	0.0
South Asia	4.4	0.9	2.9	0.6
68 economies	4.2	1.0	2.6	0.6

Source: World Bank, 1991.

Table 4.2 Growth of output and total factor productivity in the East Asian 'tigers', 1966–90 (%)

	Output growth	Total factor productivity growth
Hong Kong	7.3	2.3
Singapore	8.7	0.2
South Korea	8.5	1.7
Taiwan	8.5	2.1

Source: Compiled from Young, 1995.

as a myth. The spectacular growth of inputs, however, does need explaining. The rapid growth of capital and labour is a function of an internal dynamism fuelled by the relentless and successful drive for export markets, partly engineered by deliberate government intervention. East Asia is not the bastion of free-market enterprise that is often portrayed. The growth of factor inputs may decelerate in the future, but the performance of these four economies up to 1990 was indeed remarkable, notwithstanding the relatively low rate of growth of total factor productivity.

Hu and Khan (1997) use the production function[9] approach to understand the sources of fast growth in China over the period 1953–94, and the acceleration of growth after the economic reforms and 'open door' policy were introduced in 1978. From 1953 to 1978, GDP grew at 5.8% per annum, and then accelerated to 9.3% per annum from 1979 to 1994. Why was this? To estimate the contribution of labour, capital and total factor productivity (TFP) to measured growth over the periods, factor shares of GDP are taken as the elasticities of output with respect to labour and capital, with labour's elasticity approximately 0.4, and capital's elasticity approximately 0.6. The results are shown in Table 4.3.

To give an example, in the pre-reform period 1953–78, the growth of capital was 6.2% per annum. Multiplying 6.2 by 0.6 (capital's elasticity) gives a contribution of capital to growth of 3.72 percentage points, which is approximately 65% of the total growth of output of 5.8%. Capital accumulation was by far the most important contributor to growth in this period. In the post-reform period 1979–94, however, it can be seen that the contribution of productivity growth increases considerably to almost equal importance with capital. The rate of growth of TFP more than triples, from 1.1% per annum to 3.9%, contributing over 40% to measured growth. According to Hu and Khan (1997), the process of reform stimulated productivity growth in a number of ways, including the transfer of resources from agriculture to industry, a reallocation of resources from the public to the private sector, the encouragement of foreign direct investment (FDI), and a faster growth of exports.

Table 4.3 Sources of growth in China, 1953–94 (%)

	1953–94	1953–78	1979–94
Output growth	7.2	5.8	9.3
Capital input growth	6.8	6.2	7.7
Labour input growth	2.6	2.5	2.7
TFP growth	2.1	1.1	3.9
Contribution of capital	55.6	65.2	45.6
Contribution of labour	14.9	16.8	12.8
Contribution of productivity growth	29.5	18.0	41.6

Source: Hu and Khan, 1997.

Felipe (1999) surveys the studies done of TFP growth in the whole of East Asia, most of which use the production function approach. He is critical of many of them, and shows how estimates of TFP can vary according to the time period taken, the estimates made of the growth of factor inputs, and the assumed elasticities of output with respect to labour and capital. Remember that TFP is obtained as a residual after the contribution of the factor inputs has been calculated. The various methodological and conceptual problems associated with the use of production functions discussed earlier are also emphasized, particularly the assumption that technical progress and factor inputs are exogenous and not interrelated.

The most comprehensive recent study of the sources of growth using the aggregate production function comes from Senhadji (2000) at the IMF. He estimates production functions for 66 countries over the period 1960–94 (including 46 developing countries) of the form: $Y = TK^{\alpha}(LH)^{1-\alpha}$, where T is TFP, K is the stock of capital, L is the active population and H is an index of human capital. The function is estimated using both levels of the variables (measured in logarithms) and taking first differences of the log level (i.e. in rate of growth form – see equation (4.21)). The estimates of the elasticity of output with respect to capital (α) vary considerably across countries (and regions) and also according to whether levels or first differences of the variables are used (which is another problem). Using levels, the estimates of (α) range from 0.43 in sub-Saharan Africa to 0.63 in the Middle East and North Africa. Using first differences, the estimates of α range from 0.30 in East Asia to 0.62 in Latin America. Using the mean value of α from the equations estimated in levels gives the sources of growth in different regions shown in Table 4.4.

It can be seen again from Table 4.4 that capital accumulation is by far the most important contributor to measured growth in all the regions. The small contribution of TFP in the fastest growing region of East Asia confirms the conclusions of Young (1995). Notice also, the *negative* contribution of TFP in Africa and Latin America. Human capital formation makes a positive contribution to growth in all regions, but a relatively minor one.

It is satisfying that the conclusions from a wide range of studies using different techniques and dubious data should all point in roughly the same direction:

1. The major source of growth in developing countries is increased factor inputs, aided by improvements in the quality of labour through health improvement and education.
2. The growth of 'total' factor productivity in developing countries is relatively slow compared with that in developed countries, which may be partly a reflection of the different stage of development reached.
3. Resource transfers from agriculture to industry are quite important as a source of growth, but not as important as one might have expected. They will become more important as the ability of the industrial sector to absorb surplus labour increases.

Table 4.4 Sources of growth by region of the world, 1960–94

Region	Output growth (%)	Contribution (percentage points) of:			TFP
		Capital	Labour	Human capital	
East Asia	6.49	4.50	1.27	0.44	0.28
South Asia	4.66	2.87	0.99	0.25	0.55
Sub-Saharan Africa	2.83	1.79	1.39	0.22	−0.56
Middle East & North Africa	5.05	3.99	0.84	0.25	−0.03
Latin America	3.42	2.31	1.22	0.28	−0.39

Source: Senhadji, 2000.

Before ending, it should be said again that the aggregate models that produced the above results are rough tools. They do, however, give an important idea of the forces at work and a rough idea of the likely quantitative significance of different factors. The production function approach is also a versatile tool of analysis. Sala-í-Martin (1997) has surveyed a number of production function studies and found that researchers have included at least 62 different variables in the production function to explain growth, in addition to the growth of capital and labour.

'New' (endogenous) growth theory and the macrodeterminants of growth[10]

Since the mid-1980s there has been an outpouring of literature and research on the applied economics of growth, attempting to understand and explain the differences in the rates of output growth and per capita income growth across the world, many inspired by the so-called 'new' growth theory, or endogenous growth theory. This spate of cross-sectional studies seems to have been prompted by a number of factors:

- Increased concern with the economic performance of the poorer regions of the world, and particularly the striking differences between countries and continents.
- The increased availability of standardized data (e.g. Summers and Heston, 1991; and the World Bank *World Development Indicators*), enabling more reliable econometric work.
- Pioneering studies (e.g. Baumol, 1986) that could find no convergence of per capita incomes in the world economy, contrary to the prediction of neoclassical growth theory based on the assumption of diminishing returns to capital, which, given identical preferences and technology across countries, should lead to faster growth in poor countries than in rich ones.

It is the latter finding (although hardly new, as outlined in Chapter 2) that has been the major inspiration behind the development of the 'new' growth theory, which relaxes the assumption of diminishing returns to capital and shows that, with constant or increasing returns, there can be no presumption of the convergence of per capita incomes across the world, or of individual countries reaching a long-run steady-state growth equilibrium at the natural rate. If there are not diminishing returns to capital, investment is important for long-run growth and growth is endogenous in this sense. In these 'new' models of endogenous growth, pioneered by Robert Lucas (1988) and Paul Romer (1986, 1990), there are assumed to be positive externalities associated with human capital formation (e.g. education and training) and research and development (R&D) that prevent the marginal product of capital from falling and the capital–output ratio from rising. We have a production function in capital of:

$$Y = AK^{\alpha} \tag{4.27}$$

where K is a composite measure of capital (i.e. physical capital plus other types of reproducible capital), and $\alpha = 1$. This is the so-called **AK model** of new growth theory. As Barro and Sala-í-Martin (2003) put it: 'the global absence of diminishing returns may seem unrealistic, but the idea becomes more plausible if we think of K in a broad sense to include [for example] human capital'. It can be seen from the expression for the capital–output ratio, that is:

$$\frac{K}{Y} = \frac{K}{L} \cdot \frac{L}{Y} \tag{4.28}$$

that anything that raises the productivity of labour (Y/L) in the same proportion as K/L will keep the capital–output ratio constant. Learning by doing and embodied technical progress in the spirit of Arrow (1962) and Kaldor (1957), as well as technological spillovers from trade (Grossman and Helpman, 1990, 1991) and FDI (de Mello, 1996), are other possibilities in addition to education and R&D.

The first crude test of the new growth theory is to see whether or not poor countries do grow faster than rich ones, or, in other words, to see whether there is an inverse relation between the growth of output (or output per head) and the *initial* level of per capita income. If there is, this would provide support for the neoclassical model. If there is not, this would support the new growth theory's assertion that the marginal product of capital does not decline. The equation to be estimated is:

$$g_i = a + b_1 (PCY)_i \qquad (4.29)$$

where g_i is the average growth of output per head of country i over a number of years and PCY_i is its initial level of per capita income. A significantly negative estimate of b_1 would be evidence of **unconditional convergence**, or **beta (β) convergence** as it is called in the literature; that is, poor countries growing faster than rich without allowing for any other economic, social or political differences between countries. As we saw in Chapter 2, none of the studies taking large samples of developed and developing countries has been able to find evidence of unconditional convergence. The estimate of b_1 is not significantly negative; in fact, it is invariably positive, indicating divergence.[11]

Before jumping to the conclusion that this is a rejection of the neoclassical model, it must be remembered that the neoclassical prediction of convergence assumes that the savings or investment ratio, population growth, technology and all factors that affect the productivity of labour are the same across countries. Since these assumptions are manifestly false, there can never be the presumption of unconditional convergence (even if there are diminishing returns to capital), only **conditional convergence**, holding constant all other factors that influence the growth of per capita income, including population growth (p), the investment ratio (I/Y) and variables that affect the productivity of labour, for example education (ED), research and development expenditure ($R + D$), trade (T) and even non-economic variables such as political stability measured by the number of revolutions and coups (PS). The equation to be estimated is therefore:

$$g_i = a + b_1(PCY)_i + b_2(p)_i + b_3(I/Y)_i + b_4(ED)_i + b_5(R + D)_i + b_6(T)_i + b_7(PS)_i + \ldots \qquad (4.30)$$

and the question to be asked is: What happens to the sign of the initial per capita income variable (PCY) when these other variables are introduced into the equation? If the sign turns negative ($b_1 < 0$) when allowance is made for these other factors, this is supposed to represent a rehabilitation of the neoclassical model (see Barro, 1991); that is, there *would be* convergence if it were not for differences between rich and poor countries in all these other important variables in the growth process. New growth theory would be supported by finding that education, R&D expenditure and so on matter, and it is these factors that keep the marginal product of capital from falling, producing actual divergence in the world economy.

Note here that if the model of new growth theory is represented by the AK model, as in equation (4.27), this can be shown to be equivalent to the Harrod–Domar growth equation. Assuming $\alpha = 1$, totally differentiate equation (4.27) and divide by Y. This gives:

$$dY/Y = A(dK/Y) = A(I/Y) \qquad (4.31)$$

where dY/Y is the growth rate, I/Y is the investment ratio, and A is the productivity of capital (dY/I), which is the reciprocal of the incremental capital–output ratio. This is the same as the Harrod growth equation $g = s/c$, where s is the savings ratio and c is the incremental capital–output ratio, or the Domar equation $g = s\sigma$, where σ is the productivity of capital.

If the productivity of capital was the same across countries, there would be a perfect correlation between the growth rate of countries and the investment ratio where the slope of the relationship is the reciprocal of the incremental capital–output ratio (c). If there is not a perfect correlation, then, by definition, the productivity of capital, or the capital–output ratio, must differ between countries. New growth theory equations that attempt to explain growth rate differences between countries (such as equation (4.30) – and see empirical studies later) are really asking the question (and hopefully answering it): Why does the productivity of capital differ between countries (assuming I/Y is included in the equations)? (See Nell and Thirlwall, 2016.)

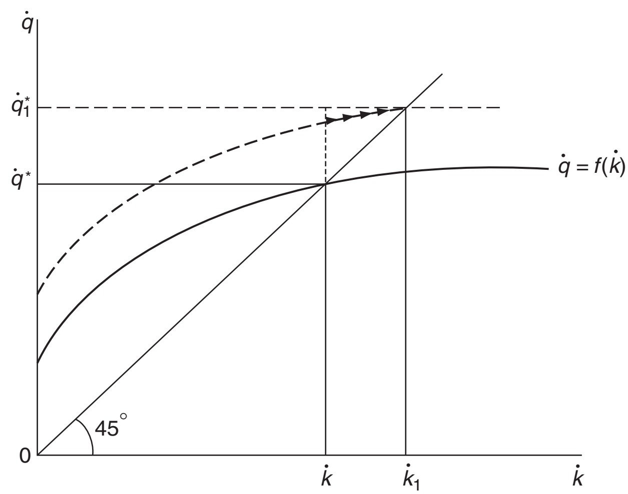
We said above that evidence of conditional convergence delights the neoclassical economists because it is interpreted as a rehabilitation of the neoclassical growth model with diminishing returns, but this may be a hasty judgement. Outside the neoclassical paradigm, there is another distinct body of literature that argues that economic growth *should be* inversely related to the initial level of per capita income because the more backward the country, the greater the scope for **catch-up**; that is, for absorbing a backlog of technology (see Gomulka, 1971, 1990; Abramovitz, 1986; Dowrick and Nguyen, 1989; Dowrick and Gemmell, 1991; Amable, 1993). Thus, the negative sign on the per capita income variable could be picking up the effect of catch-up, and the notion of catch-up is conceptually distinct from the *shape* of the production function and whether or not there are diminishing returns to capital. How are the two effects to be distinguished? As Benhabib and Spiegel (1994) remark in their paper on the role of human capital in development: ‘a negative coefficient estimate on initial income levels may not be a sign of convergence due to diminishing returns, but of catch up from adoption of technology from abroad. These two forces may be observationally equivalent in simple cross-section growth accounting exercises’. Also, output growth will be a function of the stage of development because of sectoral differences in the productivity growth rates of agriculture, industry and services, so that convergence may also be partly ‘structural’, independent of both diminishing returns and catch-up (see Cornwall and Cornwall, 1994). This adds further complications to the interpretation of the coefficient relating country growth rates to the initial level of per capita income.

Now let us turn to the question of the capital–output ratio. Non-diminishing returns to capital, or constancy of the capital–output ratio, lie at the heart of new growth theory, as pioneered by Lucas and Romer, who emphasize externalities to education and research. For the historical record, however, it should be mentioned that, many years ago, Cambridge economist Nicholas Kaldor pointed out the fact that despite continued capital accumulation and increases in capital per head through time, the capital–output ratio remains broadly the same, implying some form of externalities or constant returns to capital. It is worth quoting Kaldor (1961) in full:

> As regards the process of economic change and development in capitalist societies, I suggest the following ‘stylised facts’ as a starting point for the construction of theoretical models . . . (4) steady capital–output ratios over long periods; at least there are no clear long-term trends, either rising or falling, if differences in the degree of capital utilisation are allowed for. This implies, or reflects, the near identity in the percentage rate of growth of production and of the capital stock i.e. for the economy as a whole, and over long periods, income and capital tend to grow at the same rate.

Kaldor’s explanation (as a critique of the neoclassical production function) lay in his innovation of the **technical progress function**, which relates the rate of growth of output per worker $\dot{q}$ to the rate of growth of capital per worker $\dot{k}$, as depicted in Figure 4.8.

Figure 4.8 Kaldor's technical progress function



The position of the function depends on the exogenous rate of technical progress, and the slope of the function depends on the extent to which technical progress is embodied in capital. Along the 45^0 line the capital–output ratio is constant, and the equilibrium growth of output per head will be at $\dot{q}$. An upward shift of the technical progress function – associated, for example, with new discoveries, a technological breakthrough or more education – will shift the curve upwards, causing the growth of output to exceed the growth of capital, raising the rate of profit and inducing more investment to give a new equilibrium growth of output per worker at $\dot{q}_1^*$. An increase in capital accumulation to $\dot{k}_1$ *without* an associated upward shift in the schedule will cause the capital–output ratio to rise. New growth theory is precisely anticipated. Kaldor's technical progress function is the true progenitor of endogenous growth theory.

What applies to countries through time applies, *pari passu*, to different countries at a point in time, with differences in growth rates at the same capital–output ratio being associated with different technical progress functions. To quote Kaldor (1972, emphasis added) again:

> A lower capital–labour ratio does not necessarily imply a lower capital–output ratio – indeed, the reverse is often the case. The countries with the most highly mechanised industries, such as the USA, do not require a higher ratio of capital to output. The capital–output ratio in the USA has been falling over the past 50 years whilst the capital–labour ratio has been steadily rising; *and it is lower in the United States today than in the manufacturing industries of many underdeveloped countries.*

In other words, rich and poor countries are simply not on the same production function.

Empirical studies

In this section we survey six pioneer studies of intercountry growth rate differences that have been inspired by new growth theory. A summary of the studies is given in Table 4.5. Before turning to the individual studies, it may be said from the outset that only four variables seem to be

Table 4.5 The macrodeterminants of growth

Study	Dependent variable	Convergence	Savings–investment ratio	Population growth	Education	Government consumption distortions	Political instability	Monetary and fiscal variables	Trade variables	Inflation
Barro (1991)	Growth of per capita income	Conditional	Significant (+)	Not considered	Significant (+)	Significant (−)	Not considered	Not considered	Not considered	Not considered
Mankiw et al. (1992)	Level and growth of per capita income	Conditional	Significant (+)	Significant (−)	Significant (+)	Not considered	Not considered	Not considered	Not considered	Not considered
Knight et al. (1993)	Growth of output per worker	Conditional	Significant (+)	Significant (−)	Significant (+)	Not considered	Not considered	Not considered	Significant (+)	Not considered
Barro and Lee (1993)	Growth of per capita income	Conditional	Significant (+)	Not considered	Significant (+)	Significant (−)	Significant (−)	Not considered	Not considered	Not considered
Levine and Renelt (1992)	Growth of per capita income	Conditional	Significant (+)	Not robust	Significant (+)	Not robust	Not robust	Not robust	Not robust	Not robust
Levine and Zervos (1993)	Growth of per capita income	Conditional	Not considered	Not considered	Significant (+)	Not considered	Significant (−)	Weak	Weak	Not significant

Note: Barro (1991), Mankiw et al. (1992), Knight et al. (1993) and Levine and Zervos (1993) – 98 countries, 1960–85; Barro and Lee (1993) – 116 countries, 1965–85; and Levine and Renelt (1992) – 119 countries, 1960–89.

robust in the sense that they remain statistically significant regardless of what other variables are included in the equation. Consider an equation of the form:

$$Y = b_i I + b_m M + b_z Z + \mu \tag{4.32}$$

where *I* is a set of variables always in the regression, *M* is the variable of interest, and *Z* is a subset of variables added to the regression. As a first step, estimate the regression with the *I* variables (for example, *PCY*) and the variable of interest (say, investment). Then add up to three other variables and observe the significance of the variable of interest. If the variable remains significant without changing its sign, the variable is regarded as robust; otherwise, it is 'fragile'. The only robust variables found in the majority of studies are the ratio of savings and investment to GDP, population growth, the initial level of per capita income, and investment in human capital measured by the secondary school enrolment rate. All other variables are fragile.

The six studies surveyed are Barro (1991), Mankiw et al. (1992), Levine and Renelt (1992), Levine and Zervos (1993), Barro and Lee (1993) and Knight et al. (1993).

Robert Barro has been one of the major investigators of new growth theory. He examines the growth of per capita income across 98 countries over the period 1960–85. He is interested in testing the neoclassical growth model augmented by human capital formation. There is no significant relation between the initial level of *PCY* and the growth rate of *PCY*, which, on the surface, contradicts the neoclassical model and supports the new models of endogenous growth, which assume non-diminishing returns to capital. In the first instance, however, he does not allow for differences in investment ratios and population growth. Instead, he augments the model by allowing for differences in human capital formation, proxied by school enrolment ratios. With this additional variable, *PCY* growth is found to be negatively related to initial levels of *PCY*, which, he argues, supports the neoclassical (conditional) convergence hypothesis.

An interesting difference between 'continents' is apparent. The Pacific Rim countries in 1960 had higher human capital formation than predicted by the level of *PCY* and grew rapidly, while Africa had lower human capital formation than predicted by *PCY* and grew slowly. Countries with high ratios of human capital formation also seem to have lower fertility rates and higher ratios of physical investment to GDP, which means that the human capital variable is likely to be picking up differences in population growth and investment ratios.

Mankiw et al. (1992) take three samples of countries over the period 1960–85: 98 non-oil-producing countries, 76 developing countries (excluding small countries and those where data are doubtful), and 22 OECD countries with a population of more than 1 million. First, they take the *level* of *PCY* as the dependent variable and find that differences in savings rates and population growth account for over 50% of income differences in the large sample of countries, which is support for the second basic proposition of neoclassical growth theory. However, the cross-section regression implies a much higher elasticity of output with respect to capital than capital's share of national income, so that the empirical model overpredicts. The authors thus augment the model for differences in human capital formation, proxied by secondary school enrolment rates, and find that the augmented Solow model 'explains' 80% of differences in *PCY*, and human capital formation is a significant variable in all three samples of countries. Regressing the *growth* of *PCY* on initial *PCY* levels shows no tendency for convergence (except in the OECD sample), but there is evidence for conditional convergence in all three samples if differences in investment ratios and population growth are allowed for. It is therefore claimed by Mankiw et al. (1992) that the data give support to the Solow neoclassical model against the new endogenous growth models, which, because of the assumption of non-diminishing returns to capital, predict that differences in *PCY* between countries will persist indefinitely or even widen.

Knight et al. (1993) extend Mankiw et al.'s study in two ways. First, they use panel data (that is, pooled time-series and cross-section data) to look at country-specific effects. Second, they assume that the rate of technical progress is influenced by the 'outwardness' of trade policy and by the stock of infrastructure investment (proxied by the 'flow' variable, government fixed investment as a proportion of GDP). Trade is assumed to influence technical progress in two ways: through technological transfers, and through greater availability of foreign exchange, which enables countries to purchase technologically superior capital goods. Tests of the model, taking two samples (76 developing countries and 22 OECD countries), show that the growth of output per worker is positively related to the savings ratio, and negatively related to the growth of population and the initial level of *PCY*; that is, there is evidence of conditional convergence. Human capital investment is significant and raises the productivity of physical investment. The tests of trade 'openness', and the role of infrastructure investment, also show significant positive effects and enhance the coefficient on physical capital.

Barro and Lee (1993) analyse 116 countries over the period 1965–85 and find that five factors differentiate reasonably well slow-growing countries from fast-growing countries:

1. The initial level of *PCY* (relative to educational and health attainment), which has a negative effect (that is, there is evidence of conditional convergence).
2. The investment ratio (+).
3. The ratio of government consumption to GDP (−).
4. Market distortions measured by the black market rate of foreign exchange (−).
5. Political instability measured by the number of political 'revolutions' per year (−).

These five variables 'explain' 80% of the growth rate differences between countries. No trade variables are included in the analysis.

Levine and Renelt (1992) show that cross-country regression results are 'fragile' to model selection and datasets, but at least two 'robust' results stand out: the relation between investment and growth, and the relation between the investment ratio and the ratio of international trade to GDP. Levine and Renelt first take 119 countries over the period 1960–89 and use the growth of *PCY* as the dependent variable. The *I* (constant) variables used (see equation (4.32)) are the investment ratio, the initial level of *PCY*, the initial level of secondary school enrolment, and population growth. The pool of *Z* variables used includes government expenditure, exports, inflation, the variance of inflation, domestic credit expansion and its variance, and political instability. When the *Z* variables are added to the *I* variables, the investment ratio remains robust, the initial *PCY* variable remains robust (that is, there is evidence of conditional convergence), the secondary school enrolment rate is robust, but not population growth. None of the *Z* variables themselves are robust, however; they depend on the conditioning variables, that is, which other *Z* variables are introduced. Levine and Renelt repeat the Barro (1991) study and find only the investment ratio and the initial level of *PCY* to be robust. No fiscal or monetary indicators are robust, and no trade variables. Levine and Renelt (1992) suggest that the importance of trade probably works through investment (rather than through improved resource allocation).

Levine and Zervos (1993) report new evidence on the 'robustness' of variables, taking a different set of *I* and *Z* variables. The *I* (constant) variables used are the Barro (1991) variables of initial *PCY*, initial secondary school enrolment rate and the number of political revolutions and coups. The results largely support the earlier findings of Levine and Renelt (1992), but no investment variable is included. Levine and Zervos pay particular attention to financial variables and the role of inflation. Various indicators of financial deepening are robust (which may be standing as a proxy for investment), and apparently there are no *Z* variables that make growth and inflation negatively correlated. Levine and Zervos (1993) comment that: 'given the uncharacteristically unified view among economists and policy analysts that countries with high inflation rates should adopt policies that lower inflation in order to promote economic prosperity, the inability to find simple cross-country regressions supporting this contention is both surprising and troubling' (see Chapter 13 for a discussion of the relation between inflation and growth).

The above studies relate to countries as a whole. Rodrik (2013), however, has found that if just the manufacturing sector of countries is considered, there is evidence of unconditional convergence, taking a sample of over 100 developed and developing countries since 1990. This is, perhaps, not surprising, because manufacturing industries produce tradable goods subject to competitive pressure in a global environment where there is scope for technological transfer and the absorption of new knowledge. Traditional agriculture and non-traded goods do not share these characteristics. This suggests that lack of unconditional convergence across countries has to do with the structural characteristics of countries and particularly the relative shares of traded and non-traded goods.

What have we learned?

These studies (and many others not reported here) have revealed a lot about the sources of intercountry growth rate differences. Interestingly, the variables of significance turn out to be those that have traditionally been at the heart of mainstream growth and development theory, particularly the importance of investment and capital accumulation.

On the other hand, it is often the case that studies reach conflicting conclusions, and a large proportion of intercountry growth rate differences remain unexplained (as much as 40%). Why is this? One set of reasons relate to the availability and quality of data, and the econometric procedures used for testing. Often, data are weak and unreliable, and the econometric methodology used not only differs but is also questionable, because allowance has not been made for lags in the relationship between variables or intercorrelation between variables. A second set of reasons is that countries are much more heterogeneous in their structure and institutions than most studies allow for. As Kenny and Williams (2001) put it: 'it is because countries are so heterogeneous in their make-up and institutions that cross section studies reach contradictory results and produce a lack of robustness'. They argue: 'perhaps more energy should be directed towards understanding the complex and varied inner-workings of actual economies rather than trying to assimilate them into abstract universal models'.

A similar point is made by Putterman (2000), who argues that one of the important reasons why countries have grown at different rates over the past 60 years is that the *preconditions* for development were not equal in terms of institutional structure (the strength of government, for example), the tax system, the state of agriculture, the stock of knowledge and ideas and so on, and these factors are not well captured by the initial level of per capita income. Emphasis on preconditions, and why countries have responded differently to the possibilities of industrialization, goes back to Rostow's ideas of the preconditions for take-off. To put it another way: economic history matters. The question is how to measure the level of 'pre-modern' development. Putterman (2000) concentrates on the conditions prevailing in agriculture such as cultivatable land per head, population density, and the prevalence of irrigation. When these variables are included in regression equations, along with the investment ratio, population growth and education, there is an increase in the proportion of the variance in growth rates that is explained.

Another serious weakness of new growth theory is that many of the models are closed economy models, and there are no demand constraints. It is difficult to imagine how growth rate differences between countries can be explained without reference to trade, and without reference to the balance of payments position of countries, which in most developing economies constitutes a major constraint on the growth of demand and output (see Chapter 16). Where a trade variable is included in the models tested, it is invariably insignificant, or loses its significance when combined with other variables. All this is very puzzling, given the rich theoretical and empirical literature that exists on the relation between trade and growth (see Chapter 15). There are at least two possible explanations. First, it could be that trade works through investment. Indeed, in some studies that look directly at the determinants of investment, trade and exports are found to be very significant. Second, the measure of trade taken is a very static one, usually measured as the share of trade in GDP. This may pick up the static gains from trade but not the dynamic gains. In a growth model, the most obvious trade variable to focus on is the *growth* of exports, which will favourably influence growth from the demand side (particularly by relaxing a balance of payments constraint on domestic demand), and from the supply side by raising import capacity. Nell and Thirlwall (2016) show export growth to be the second most important variable after the investment ratio in explaining differences in the growth performance of 84 countries over the period 1980–2011.

Finally, a more fundamental issue is raised by Pritchett (2000), who argues that it is difficult to characterize the growth of many developing countries by a single time trend because growth is very volatile. Periods of rapid growth are often followed by plateaus and steep declines. Rapid and slow growth are, for the most part, transitory. Very few countries see their success or failure persist from decade to decade. Correlations of country growth rates across periods (e.g. 5–10 years) show very low correlations. Taking 111 countries over a 25-year period, Pritchett shows that in 55 of them, growth either accelerated or decelerated by more than 3 percentage points on at least one occasion over the period. In 40% of developing countries, trying to estimate a time trend for the growth of output gives a correlation coefficient of less than 0.5; and volatility around the trend is much higher for developing countries than developed countries. So Pritchett asks the question: What aspects of a country's growth is growth theory trying to explain when growth is so ephemeral and volatile? If growth is so volatile, it is no wonder that the variance of growth explained by traditional variables is relatively low. What is important is to analyse and explain the determinants of *shifts* in growth rates from one period to another.

Hausmann et al. (2005) try to do this. They take 106 countries over the period 1957–92, defining a 'growth acceleration' as an increase in PCY growth of 2% or more per annum over an eight-year period with a minimum growth rate of 3.5% per annum. Also the post-acceleration output level must exceed the pre-episode peak level of income (to rule out cases of pure recovery from deep depression). They find 83 episodes of growth accelerations, and 60 countries out of the 106 had at least one. The average growth acceleration is 4.7 percentage points. When they look at the causes of accelerations, however, they are struck by their unpredictability. There is only a weak link between conventional determinants of growth and growth accelerations. Investment, trade and real exchange rate depreciation are the strongest links, but there seems to be very little association between standard economic reform packages and growth accelerations. **Only 14% of accelerations were associated with economic liberalization**. Only 18% of reform episodes and 14% of political regime change were followed by growth accelerations. Hausmann et al. (2005) conclude that 'growth accelerations seem to be driven largely by idiosyncratic causes'. This may be because the 'binding constraints' on growth are 'idiosyncratic' and for certain time periods get relieved. This leads us to a brief discussion of the topic of growth diagnostics and binding constraints on growth pioneered by Hausmann et al. (2008).

'Growth diagnostics' and binding constraints on growth

A poor growth and development performance may be caused by a multitude of factors, but a sweeping programme of reforms (à la Washington Consensus), including financial and trade liberalization, the privatization of enterprises and reductions in public expenditure, may not be the solution. This is what Hausmann et al. (2008) call the 'spray gun' approach to economic policy-making, which may not hit hard enough the binding constraints on growth and development that really matter, and which are likely to vary from one country to another. Much better, they argue, is to undertake 'growth diagnostics', which locates the binding constraints on economic performance, and to target them directly, giving the most favourable outcomes from the resources expended. For an overview of the approach, see Rodrik (2010).

The framework of 'growth diagnostics' encompasses all major strategies of development. Its importance is that it clarifies *which* strategies are most likely to be effective. Since investment is the key to long-run growth, the obvious starting point for growth diagnostics is to answer the question: Why is investment low? Is it the high cost or lack of access to finance? Is it an intrinsically low social rate of return to investment, or is it that returns cannot easily be appropriated by private agents? If the problem is the cost and availability of financial resources, this is likely to be associated with low savings rates, high interest rates, and large balance of payments deficits. If the social return is low, this could be due to unfavourable geography, lack of infrastructure, poor education and health, and a lack of technological dynamism. If the ability to appropriate returns is difficult, this could be due to an uncertain economic and political climate, high taxes, absence of the rule of law, and weak property rights. Once the diagnosis is done, certain policy reforms follow and others can be ruled out, saving time and effort. Policy must then be targeted as close to the distortion and binding constraint as possible.

Hausmann et al. (2008) illustrate their methodology by comparing and contrasting three developing economies: Brazil, El Salvador and the Dominican Republic. According to the identification of binding constraints, policy recommendations differ. Brazil, for example, seems to be constrained by a shortage of finance, not by low returns on investment, so the policy message is to raise domestic saving and to attract foreign funds. By contrast, the binding constraint in El Salvador is not a shortage of investment funds, but a low social return associated with a lack of technological dynamism. It needs new (industrial) activities to invest in. The Dominican Republic tells a story of inability to cope with shocks. Here, institutional and political reforms are likely to yield the highest return. Institutions to deal with conflict management are very important to cope with the consequences of shocks and change (see Chapter 8).

The World Bank's Commission on Growth and Development, chaired by the Nobel laureate Michael Spence, and including Robert Solow, identified 13 countries that have grown at more than 7% per annum for at least 25 years since 1950. They are listed in Case example 4.1. The ingredients of their success are also highlighted: commitment to growth, combined with effective governance; high savings and investment rates; rapid export growth; macroeconomic stability; the import of knowledge and technology, and market-friendly policies.

Case example 4.1

Findings of the Commission on Growth and Development

The World Bank's Commission on Growth and Development 2008, headed by Nobel laureate Professor Michael Spence, identified 13 countries that have grown at more than 7% per annum for at least 25 years since 1950. They are listed below.

		Per capita income	
Economy	**Period of high growth**	**At start of growth period**	**2005**
Botswana	1960–2005	210	3,800
Brazil	1950–1980	960	4,000
China	1961–2005	105	1,400
Hong Kong SAR	1960–1997	3,100	29,900
Indonesia	1966–1997	200	900
Japan	1950–1983	3,500	39,600
Korea	1960–2001	1,100	13,200
Malaysia	1967–1997	790	4,400
Malta	1963–1994	1,100	9,600
Oman	1960–1999	950	9,000
Singapore	1967–2002	2,200	25,400
Taiwan (Province of China)	1965–2002	1,500	16,400
Thailand	1960–1997	330	2,400

The commission identified six major ingredients of success:

1. Commitment to growth, combined with effective governance
2. High savings and investment rates
3. Fast export growth
4. Macroeconomic stability
5. Import of knowledge and technology
6. Market-friendly policies.

Source: World Bank, 2008.

Summary

- All the great classical economists of the eighteenth and nineteenth centuries were development economists, in the sense that they were all concerned with the causes and consequences of economic growth during the Industrial Revolution in Europe at this time.
- Adam Smith was optimistic about the growth and development process based on increasing returns in industry.
- Malthus, Ricardo, Mill and Marx were all pessimistic about the development process because of diminishing returns in agriculture and a declining rate of profit in industry. Classical pessimism has been confounded by rapid technical progress in both agriculture (offsetting diminishing returns) and industry (allowing real wages to rise without the rate of profit falling).

- Modern growth theory originated with Harrod's 1939 paper 'An Essay in Dynamic Theory', in which he distinguishes three growth rates: the actual growth rate (g), the warranted growth rate (g_w) and the natural growth rate (g_n). Divergences between g and g_w cause short-run instability. Divergences between g_w and g_n cause secular stagnation if $g_w > g_n$, or growing structural unemployment with inflation if $g_n > g_w$ (which is the case for most developing economies). There were no mechanisms in the Harrod model for equalizing g, g_w and g_n.
- Solow's 1956 neoclassical growth model provided an equilibrating mechanism to bring g_w and g_n together so that all economies in the long run would grow at their natural rate of growth determined by the growth of the labour force and the growth of labour productivity. In Solow's original model, investment did not matter for long-run growth because of the assumption of diminishing returns to capital.
- Solow's model also predicted that poor countries should grow faster than rich countries, leading to a convergence of per capita incomes across the world, but we do not observe a convergence of living standards. 'New' growth theory, or endogenous growth theory, attempts to provide an answer. There are lots of factors that prevent the marginal product of capital from falling as countries get richer and invest more, such as education, R&D expenditure, learning by doing, trade and so on, so that investment does matter for long-run growth and is not simply exogenously determined by the natural rate of growth. The AK model is the simplest new growth theory model, which assumes constant returns to capital.
- The neoclassical Cobb–Douglas production function can be used to decompose the sources of growth into the contribution of labour input, capital input and total factor productivity growth. In fact, any variable can be included in a production function and its contribution to output measured.
- New growth theory uses many of the same variables as the production function approach in analysing the sources of growth, but with particular focus on education, R&D effort and institutional variables, and testing for conditional convergence by including in the estimating equation the initial level of per capita income.
- Cross-country analysis, however, is unable to explain the growth and development experience of individual countries, so growth diagnostics become important in identifying the binding constraints on growth in particular economies, and the factors that are associated with growth accelerations within countries.
- The major determinants of rapid growth and development are: high investment, fast export growth (to pay for imports), macroeconomic stability, high levels of human capital formation, and an institutional framework conducive to growth (e.g. secure property rights, the rule of law and political stability).

Chapter 4

Discussion questions

1. What did Adam Smith mean when he said that the 'division of labour is limited by the extent of the market' and 'the extent of the market is limited by the division of labour'? What is the economic significance of these propositions?
2. Why were the classical economists after Adam Smith pessimistic about the growth and development process?

Chapter 4 **Discussion questions – *continued***

3. How does Harrod define the warranted growth rate and the natural growth rate? What are the implications for a country if the natural growth rate exceeds the warranted rate?
4. What is the mechanism in neoclassical growth theory by which the warranted rate of growth adjusts to the natural rate? Do you think it is a realistic mechanism?
5. What are the essential assumptions and predictions of neoclassical growth theory, and how is the conclusion reached that investment does not matter for long-run growth?
6. What are the special properties of the Cobb–Douglas production function, and how might the function be used to calculate the sources of growth?
7. What is the difference between exogenous and endogenous technical progress?
8. What factors does the growth of 'total factor productivity' depend on?
9. What have been the major findings of production function studies of the sources of growth in developing countries?
10. Outline the essential propositions of new (endogenous) growth theory.
11. What have we learnt from the major studies of the macrodeterminants of growth in developing countries?
12. How useful is the exercise of growth diagnostics?

Notes

1. In practice, g_n may respond to g. This is the idea of the endogeneity of the natural rate of growth. See Leon-Ledesma and Thirlwall (2002) for a study of developed countries; studies by Vogel (2009) and Libanio (2009) for Latin American countries, and Dray and Thirlwall (2011) for a study of a selection of Southeast Asian countries. Typically, an increase in the actual growth rate of 1 percentage point above the estimated natural rate increases the natural rate by between 0.5 and 0.8 percentage points through increasing the rate of growth of the labour force and labour productivity growth.
2. This can be seen by rearranging the equation to $qs/k = l$, where $q = Y/L$; $s = S/Y = \Delta K/Y$ (since all saving leads to capital accumulation); $k = K/L$, and $l = \Delta L/L$. Therefore $(Y/L)(\Delta K/Y)(L/K) = \Delta L/L$, or $\Delta K/K = \Delta L/L$.
3. To illustrate the effect of a change in technical progress on the steady state, the axes in Figure 4.4 must be changed to 'output per *effective* worker' on the vertical axis and 'capital per *effective* worker' on the horizontal axis.
4. For the time being, the formidable problems associated with an aggregate measure of capital are ignored.
5. The proof is as follows. The elasticity of output with respect to capital, α, is $(dY/Y)/(dK/K) = (dYK)/(dKY)$. Now, if capital is paid its marginal product, then $dY/dK = r$, where r is the rental on capital. Hence, $\alpha = rK/Y$, where rK/Y is capital's share of total output. Thus, under perfectly

competitive assumptions, the elasticity of output with respect to any factor is equal to that factor's share of total output.

6. Alternatively, the total differential of equation (4.20) can be taken and the result divided by output, which will also convert the equation into rate of growth form.
7. The elasticity of substitution (σ) relates the proportional change in relative factor inputs to a proportional change in the marginal rate of substitution between labour and capital (MRS) (or the proportional change in the relative factor–price ratio on the basis of marginal productivity theory). The elasticity of substitution may therefore be written as

$$\sigma = \frac{\partial \log (L/K)}{\partial \log \text{MRS}}$$

The proof that $\sigma = 1$ is very simple:

$$\text{MRS} = \frac{\partial Y}{\partial K} \bigg| \frac{\partial Y}{\partial L} \bigg| = \frac{\alpha L}{\beta K}$$

$$\log \text{MRS} = \log \frac{\alpha}{\beta} + \log \frac{L}{K}$$

Differentiating with respect to log MRS gives

$$1 = \frac{\partial \, log \, (L/K)}{\partial \log MRS} = \sigma$$

8. To overcome the restrictive property of the Cobb–Douglas function when the growth rates of factors differ, it is possible to use the more general constant elasticity of substitution production function, of which the Cobb–Douglas is a special case. We cannot discuss the function here except to say that it, too, is not without its specification errors. There is still the assumption of constancy, which has the drawback that one may be ascribing changes in elasticity to changes in technology that are really due to changes in factor proportions. This limitation can be overcome only with a function possessing the property of variable elasticity of substitution.
9. The production function used here is the translog production function, which allows for the elasticity of substitution between inputs to vary.
10. For a discussion of the origins of endogenous growth theory, and its relevance to developing countries, see Romer (1994), Pack (1994), Ruttan (1998) and Temple (1999). For an advanced textbook treatment of the topic, see Barro and Sala-í-Martin (2003).
11. For an advanced theoretical discussion of convergence issues, see Durlauf (1996); see also Islam (2003).

Websites on growth theory

New School for Social Research (New York) www.newschool.edu/nssr/
Economic Growth Resources run by Jon Temple, Bristol University, UK www.bristol.ac.uk/efm/people/jon-r-temple/overview.html
Overseas Development Institute www.odi.org
Foundation for Advanced Studies on International Development www.fasid.or.jp/english
Institute of Developing Economies: Japan External Trade Organization www.ide.go.jp/English/
Vienna Institute for International Economic Studies www.wiiw.ac.at
Carnegie Endowment for International Peace http://carnegieendowment.org/

II

FACTORS IN THE DEVELOPMENT PROCESS