# Introduction

> Begin at the beginning, and go until you come to the end: then stop.
> (*Alice's Adventures in Wonderland*, Lewis Carroll)

The production function is undoubtedly one of the most widely used concepts in economics. Students of economics are normally introduced to the theory of production at an early stage of their studies. Introductory microeconomics textbooks outline the production function, isoquants, the conditions for cost minimisation, the demand for factors of production (based upon the marginal product theory of factor pricing) and so on. At the same time, the production function is extended seamlessly in first-year macroeconomic textbooks to encompass individual industries or, indeed, the whole economy. There is, however, little, or more usually no, discussion about the conditions under which it is legitimate to sum micro-production functions to give a well-defined aggregate production function. That this should be considered is not simply for mere intellectual curiosity. Indeed, since the 1940s, economists such as Leontief, Klein, or Nataf, among others, studied the aggregation problem, and for very good reasons. The same functional form is often assumed to hold irrespective of whether the production function refers to an individual plant, firm, industry, or to the whole economy. This, it turns out, has little or no theoretical justification. Sato and Fisher clarified and extended the work on aggregation during the 1960s and 1970s and obtained very damaging conclusions for the plausibility of aggregates such as output and capital.

The Cobb–Douglas production function is usually the first specific functional form that students encounter, partly because of its mathematical simplicity and the pedagogical advantage that this brings. But it is not merely useful for teaching purposes. It is also used in many theoretical and empirical research papers, as a perusal of the recent issues of any mainstream economics journal will confirm. Clearly, it is widely considered that the Cobb–Douglas is more than just a convenient teaching concept, but does indeed represent the actual production conditions of an industry or economy, albeit only as an approximation. Other more flexible functional forms, such as the constant elasticity of substitution (CES) or translogarithmic (translog) production functions, are also used. However, in many, but not all, cases it seems that relatively little is to be gained in the use of these more complex production functions. Moreover, the fact that factor

shares are roughly constant is seen to provide an empirical justification for the use of the Cobb–Douglas production function. For example, Hoover (2012, p. 330) states in his intermediate macroeconomics textbook:

> The striking fact that, while there is some variation, the variation [in factor shares] is small and there is no trend. The approximate constancy of the labor share confirms the prediction of our model and provides a good reason to take the Cobb–Douglas production function as a reasonable approximation of aggregate supply in the U.S. economy.

Similar sentiments are expressed in Mankiw's (2010, pp. 56–9) more introductory macroeconomics textbook, where the Cobb–Douglas production function is uncritically introduced. As Kuhn (1962 [1970]) has shown, textbooks are important in that they inculcate the student into the prevailing paradigm and implicitly set the legitimate questions to be examined, through, for example, worked examples and the questions at the end of the chapters. These set the agenda for what are seen as the appropriate models and methodology for work at the frontiers of paradigm. Consequently, the erroneous impression that the aggregate production function is a useful approximation to the technological conditions of, say, the whole economy, is perpetuated.

The more flexible production functions suffer from other problems. For example, the CES production function is a non-linear form and its econometric estimation is more difficult. And the translog often suffers from severe multicollinearity. Indeed, the ubiquity of the Cobb–Douglas production function makes it a toss-up as to whether the names 'Cobb and Douglas' or 'Keynes' have been mentioned more frequently in the economics literature over the last few decades.

In fact, in spite of the criticisms that Cobb and Douglas's original empirical work received (Cobb and Douglas, 1928), so hostile that Douglas momentarily considered abandoning all further work on the production function, their article has subsequently been recognised as one of the top 20 papers published during the last hundred years in the *American Economic Review* (Arrow et al., 2011). The citation to their work reads:

> The cliché surely applies here: this paper needs no introduction. The convenience and success of the constant-elasticity Cobb–Douglas function has spread its use from representing production possibilities, which of course was its original use, to representing utility functions and to much else throughout empirical and theoretical economics. Cobb and Douglas explored elementary properties and implications of the functional form and pointed to the approximate constancy of the relative shares of labour and capital in total income as the validating empirical fact. (p. 2)

Ever since Solow's (1956, 1957) two seminal papers on growth theory, the aggregate production function has become the *sine qua non* of neo-

classical growth models. The more recent developments in endogenous growth theory that began in the mid-1980s depend equally on the validity of the concept of the aggregate production function. Indeed, it is possible to go so far as to say that the core of neoclassical macroeconomics relies on the aggregate production function in one form or another, including, for example, real business-cycle theory and short-run models of unemployment. If we were compelled to dispense with the aggregate production function, then it is fair to say that little would remain of either short- or long-run neoclassical macroeconomic models. This would be a disconcerting prospect for many economists, to be resisted at any cost.

Nevertheless, notwithstanding its widespread use, there are a number of severe methodological problems facing the aggregate production function that make its use problematical. Most notably, there are the problems posed by both the Cambridge capital theory controversies and what may be generically termed the 'aggregation problems' that are to be found in the somewhat broader aggregation literature. While we discuss these in more detail in Chapter 1, it is useful to consider them briefly here.

The Cambridge capital theory controversies, as the name suggests, were concerned with the theoretical problems of aggregating heterogeneous individual capital goods into a single index that could be taken as a measure of 'capital' as a factor input. The debate started in earnest in the 1950s, and went through much of the 1960s and up to the early 1970s, although its origins can be traced back to the Classical economists. The outcome was that it was generally agreed that no such index could be constructed (Harcourt, 1972; Cohen and Harcourt, 2003, 2005). The debate further showed that, when comparing steady-state economies, there is no necessary inverse monotonic relationship between the rate of profit and the capital–labour ratio, as in the neoclassical schema, outside of the restrictive one-sector model.

However, there was a good deal more to the debate than a clash of ideologies (or paradigms, to use a less emotive word), as Solow, for example, retrospectively views it.[1] Even some neoclassical economists were disturbed by the conclusions of the controversies. Commenting on Brown's (1980) comprehensive survey of both the capital controversies and the aggregation problems, Burmeister (1980, p. 423) concluded, 'I agree fully with Brown's stated conclusion that "the neoclassical parable and its implications are generally untenable". . . . Freak cases such as Samuelson's surrogate production function example are of little comfort'.

[1] 'The whole episode now seems to me to have been a waste of time, a playing-out of ideological games in the language of analytical economics' (Solow, 1988, p. 309).

He even made the radical suggestion that 'for the purpose of answering many macroeconomic questions – particularly about inflation and unemployment – we should disregard the concept of a production function at the microeconomic level' (pp. 427–8). If we follow this advice, then, of course, the concept of the production function at the macroeconomic level is also vitiated.

A second criticism is the 'aggregation problem'. This shows that the conditions under which it is possible to sum micro-production functions to give an aggregate relationship are so restrictive as to make the concept of the aggregate production function untenable (Brown, 1980; Fisher, 1992; Felipe and Fisher, 2003). It should be noted that this problem occurs in spite of the implausible assumption that there exist well-defined production functions at the firm level, where the inputs are all used optimally.

The technical literature on this is quite complicated and we review it briefly in the next chapter, but the problem is intuitively very straightforward. Consider, say, the manufacturing sector. This consists of such diverse industries as (to take as random examples) SIC 204, Grain Mill Products, and SIC 281, Industrial Organic Chemicals. Does it make any sense to combine the values of each of the outputs and the inputs of the two industries and estimate a production function that purportedly represents the underlying combined technology of these two industries? How do we even interpret the 'average' elasticity of substitution? In fact, the actual position is even worse than this, as estimating an aggregate production function for, say, manufacturing, combines many more disparate industries, and for the total economy, an even greater number.

Consider, for example, a less developed country such as the Philippines where, in Manila, a modern international banking system, complete with the latest information technology, coexists with small back-street enterprises, such as food stalls, located literally only a few streets away. Again, does it make sense to combine these activities in terms of both their outputs and inputs, as is implicitly done when an aggregate production function is estimated for the whole economy? Do we expect all these industries to be technically efficient, which is one of the necessary conditions for aggregation? As Leibenstein (1966) has shown empirically, producer or X-efficiency can differ greatly between firms making identical products.

Are workers in the informal or in the rural sectors in developing countries paid their marginal products and are they fully employed with no disguised unemployment? How do we measure the output of a marginal worker in the service sector, when the national accounts often use the deflated value of the remuneration of the inputs (especially labour) in these sectors as a measure of the real value of the output, with possibly some arbitrary allowance for productivity growth? These, to our way of

thinking, are largely rhetorical questions, yet many studies uncritically use the aggregate production function, whether in a growth-accounting context (see, for example, the survey by Maddison, 1987) or in econometric analysis (for example, Mankiw et al., 1992), using data for both the advanced and the developing countries.

Fisher (2005, p. 490), who over the years has done more than most to determine the technical conditions under which one can aggregate micro-production functions into an aggregate production function, has summarised the conclusion to be drawn from this literature as follows: 'the conditions for aggregation are so very stringent as to make the existence of aggregate production functions in real economies a non-event'. He further argues that the conditions are such that aggregate production functions cannot even be regarded as *approximations*, as Solow (1957), for example, regarded them.

Yet, it is ironical that a consideration of these serious problems has all but totally disappeared from the textbooks, and the capital theory controversies have been relegated to the history of economic thought, which few economists bother with. Consequently, a whole new generation of economists uncritically use the aggregate production function with no appreciation of how tenuous its foundations are (Sylos Labini, 1995). It is indicative that Cohen and Harcourt felt compelled to write a reminder for the profession in the 2003 issue of the *Journal of Economic Perspectives* in the 'Retrospectives' section entitled 'Whatever Happened to the Cambridge Capital Theory Controversies?' and that Birner's 2002 volume, *The Cambridge Controversies in Capital Theory*, is part of the Routledge Studies in the History of Economics.[2] The aggregation problem has fared little better. In spite of Fisher's persistent warnings of its damaging implications for the aggregate production function, virtually none of the plethora of recent applied and theoretical papers on, for example, economic growth, pays even lip-service to the aggregation problem.

It is instructive to look at how the Cambridge capital theory controversies and the aggregation problem have been covered in the textbooks and survey articles on economic growth over the last 30 years, or so. We take 1971 as the starting year. This was chosen because by that date the main conclusions and implications of the Cambridge capital theory controversies had become established. Harcourt's (1969) accessible critique of the aggregate production function had been available for a couple of years.

[2] Birner's book, while predominantly examining the Cambridge controversies from a methodological perspective, also contains a clear exposition of some of the developments in capital theory subsequent to Harcourt's (1972) survey.

The damaging problems for the aggregate production function posed by the required aggregation conditions should also have been widely appreciated by this time. Fisher (1992, p. xiii), for example, indicates that as far back as 1970 he had already called 'into question the use of aggregate production functions in macroeconomic applications such as Solow's famous 1957 paper'.

The standard textbooks on economic growth at this time, namely, Wan (1971), Jones (1975) and Hacche (1979), and the survey article by Nadiri (1970), all mentioned the capital controversies. Wan, Jones and Nadiri also mentioned the aggregation problem.

Wan (1971) was, for its time, a highly mathematical postgraduate textbook that comprehensively covered the state of neoclassical growth theory at that date: the Solow model, vintage capital goods growth models, optimal growth models and so on. Nevertheless, it also found space to include a chapter on the Robinson and Kaldor growth models. Chapter 4 of Wan's book presents a concise introduction to both the Cambridge controversies and the aggregation problems, and the damaging implications are clearly set out on page 110 of the volume. Indeed, it is ironical that Wan notes on that page that 'Mrs Robinson originally was not pessimistic enough. She still maintained the hope that techniques can generally be ranked by their "real" capital/labour ratio'. Jones (1975) and Hacche (1979) were popular and clearly written third-year undergraduate and/or postgraduate textbooks. Both authors dealt with the Cambridge controversies, but only the former with the aggregation problem. Both spent a considerable portion of their books elaborating the Kaldorian or neo-Keynesian theories of economic growth, which have now entirely disappeared from the more recent growth textbooks. Nadiri's (1970, p. 1146) article was a survey of the more applied aspects of growth theory, including the growth-accounting approach, but ended with the warning that 'the aggregate production function does not have a conceptual reality of its own'. Regarding total factor productivity (TFP), he added: 'without proper aggregation we cannot interpret the properties of an aggregate production function, which rules the behavior of total factor productivity' (p. 1144).

But by the 1990s all mention of these problems had disappeared from the growth theory textbooks, including Barro and Sala-i-Martin (1995 [2003]), Jones (1998 [2002]), Aghion and Howitt (1998, 2009), Weil (2005) and Acemoglu (2009). The survey on growth accounting by Maddison (1987) did not share any of Nadiri's reservations about the aggregate production function. However, to be fair, Temple (1999, p. 150) in his survey of the new growth theory evidence, notes briefly that 'arguably the aggregate production function is the least satisfactory element of macroeconomics,

yet many economists seem to regard this clumsy device as essential to an understanding of national income levels and growth rates'. Nevertheless, Temple is more concerned about the importance of structural change, which one-sector models tend to abstract from, than about the legitimacy of the concept of aggregate production. Temple (2006) presents a defence of the use of the aggregate production function which is not compelling, as we shall show in this book.

Valdés (1999, p. xii) in the preface to his textbook on growth mentions that he hated, for example, the 'exaggeratedly heated "capital controversies"', but there is no further elaboration. He also mentions the need to 'accept that an aggregate production function exists' (p. 63), but there is no justification for this position. And on pages 105 to 106 of his textbook, he presents a model that does not satisfy the aggregation conditions.

After the substantial literature on neoclassical growth theory generated by Solow's (1956, 1957) path-breaking articles, the late 1970s and early 1980s were a relatively barren period for the subject.[3] But this was not because of any reservations about the use of the aggregate production function. It was simply because the important Kuhnian theoretical puzzles seemed to have been solved and it was thought that there was only some marginal tidying up to be done – the Solow growth model had been generalised to two sectors; optimal growth models had been constructed using the calculus of variations or optimal control theory; the golden rule of accumulation had been examined; the role of money in growth theory modelled; the implications of increasing returns for steady-state growth, although with diminishing returns to each factor of production, had been analysed. Indeed, the classic survey of Hahn and Matthews, although written in 1964, remained on many student reading lists for a good many years after its year of publication (complemented by the 1972 survey of the applied aspects of technical change by Kennedy and Thirlwall).

All this changed after the publication of Romer's 1986 paper, which presented the first of a new generation of endogenous growth models that attempted to explain technical progress.[4] Solow (1956) had treated this as exogenous, not because he believed that technical change appeared like 'manna from heaven', but simply for want of a satisfactory explanation.

---

[3] The growth-accounting approach of Denison (1967), and subsequent studies, had largely confirmed the quantitative importance of TFP growth, or the Solow residual (often misleadingly referred to as the rate of technical progress) found by Solow (1957) (see Solow, 1988). The claim by Jorgenson and Griliches (1967) to have fully explained away the residual was shown to be erroneous (Denison, 1972a and 1972b).

[4] Early endogenous growth models include Kaldor's (1957) 'technical progress function', Frankel's (1962) 'development modifier' and Arrow's (1962) 'learning-by-doing' model.

This, together with the rapid development of large databases (such as Summers and Heston's (1991) Penn World Tables) led to an explosion of both theoretical and applied neoclassical studies on economic growth. Consequently, there were new puzzles to solve (how to endogenise technical change and so on) and old puzzles became relevant again (Mankiw et al., 1992).

Given the normal lag between research publications and the inclusion of simplified versions of these models in textbooks, it was not until the mid-1990s that a new generation of growth textbooks became available. By now, neoclassical growth theory and, as we have seen, the use of the aggregate production functions were treated as uncontroversial and seen as useful for understanding the determinants of economic growth, even though at a high level of aggregation. This is not to say that there were (and still are) no disagreements of how best to solve the neoclassical growth 'puzzles', with the rehabilitation of Solow's approach by Mankiw et al. (1992) and the different approaches taken to endogenise technical change (Romer, 1994). Moreover, questions regarding the best econometric specifications and best statistical methods to be used in testing or estimating economic growth models remained. But the Cambridge capital theory controversies, aggregation problems and the alternative growth models of Joan Robinson and Nicholas Kaldor had been banished to the nether regions. Not all mention of the Cambridge controversies disappeared from the recent literature, but references were few and far between. Pasinetti (1994, p. 357), for example, felt compelled to remind the participants at a major IEA conference on economic growth:

> This result [that there is no unambiguous relationship between the rate of profit and the capital–labour ratio], however uncomfortable it may be for orthodox theory, still stands. Surprisingly, it is not mentioned. In almost all 'new growth theory' models, a neoclassical production function, which by itself implies a monotonic inverse relationship between the rate of profits and quantity of capital per man, is simply *assumed*. (Emphasis in the original)

Bernanke (1987, p. 203, emphasis in the original), commenting on the new endogenous growth models, also aired a similar concern: 'It would be useful, for example, to think a bit about the meaning of those artificial constructs "output", "capital" and "labor" when they are measured over such long time periods (*the Cambridge–Cambridge debate and all that*)'.

The aggregation problem, in contrast, has never been discussed in any great depth at the textbook level, and while neoclassical economists working on constructing capital stocks have inevitably encountered, and accepted, the various problems, it has never been seen as insurmountable

in either theoretical or applied work.[5] Notable exceptions, noted above, are Brown (1980) and Burmeister (1980) and, of course, the extensive work of Fisher (1992, 2005).

The short-run aggregate production function, holding capital constant, has also been widely used in macroeconomics, especially since the development of the aggregate supply–aggregate demand (AS/AD) model in the neoclassical synthesis. A key tenet of this neoclassical theory is that unemployment is a consequence of real wage rigidity. The model assumes the existence of an inverse relationship between employment and the wage rate, namely, the labour demand function, in turn derived from the aggregate production function. The more recent New Classical real business-cycle models also depend on the aggregate production function and productivity shocks to explain fluctuations in employment.

These arguments show that the theoretical foundations of the aggregate production function are so flawed that there is little justification for using it, *even as an approximation*. Moreover, these problems first became apparent decades ago. Yet, Walters (1963a), for example, who had written one of the early definitive studies on cost and production functions that included a discussion of the aggregation problem, and is still worth reading today, could not avoid the temptation of estimating aggregate production functions (Walters, 1963b). As he put it: 'the theoretical foundations of the aggregate production functions give one grounds for doubting whether the concept is at all useful. Nevertheless, the temptation to discuss movements in indices of input and output in terms of such a function is difficult to resist. And there is no doubt that it is useful to rationalize the data along these lines' (Walters, 1963a, p. 425). It is somewhat difficult to reconcile the last sentence with the conclusions of his survey of production and cost functions (both published the same year), to say the least. Today, economists seem to be largely unaware of the seriousness of the aggregation problem.

Solow (1957, p. 312) argued that the aggregate production function is merely a (heroic) simplification and like any model will have unrealistic assumptions. As he put it: 'it takes something more than the usual "willing suspension of disbelief" to talk seriously of the aggregate production function', but, even so, he is willing to suspend disbelief. At the end of

[5] For example, Hulten (1980, p. 124) accepts that 'capital aggregation must therefore be regarded as an approximate, or as applying in exact form only under exceptional circumstances. Applied economists can either accept this unfortunate situation or try to work directly with a disaggregated form of their model'. But he then cites Fisher (1965) as saying that the problem may, in fact, be insoluble. Nevertheless, Hulten, *inter alios*, is one of the leading exponents of the growth-accounting approach, which assumes the existence of an aggregate production function together with the usual neoclassical conditions.

the day, the question is whether or not the aggregate production function provides a reasonable approximation to the underlying technology of an economy, notwithstanding all its underlying problems; and whether it provides useful insights into, say, the growth process. This does raise the question as to how we are to judge whether or not the insights that it supposedly provides have any verisimilitude. A standard defence of the aggregate production function, for example, compares capital reswitching to the anomalous case of the Giffen good in consumer theory, the existence of which has not led to the abandonment of the law of demand. This, however, largely begs the question as it is not clear whether capital reswitching is the rule or the exception. Simulation exercises suggest that perhaps it is the latter, but such results depend upon the exact structure of the simulation models used, and it is doubtful if they fully capture the complex production process of a modern economy. Moreover, others such as Sraffa, take this to be irrelevant – the problem is that one cannot work with a construct, such as the aggregate production function, that is *logically* flawed. The Giffen good is not a logical inconsistency in consumer theory.

The answer to why the production function continues to be widely used today seems to be that its estimation, ever since Douglas's work in the 1920s with Cobb and subsequently in the 1930s with other colleagues, generally, but not always, gives good statistical fits. Furthermore, the estimated output elasticities obtained by Douglas using cross-sectional data were often very close to the factor shares obtained from the national accounts, as predicted by the aggregate marginal productivity theory of factor pricing. As Solow once remarked to Fisher, 'had Douglas found labor's share to be 25 per cent and capital's 75 per cent instead of the other way around, we would not now be discussing aggregate production function' (cited by Fisher, 1971b, p. 305).

The good statistical fit that the aggregate production function can give was forcibly brought home to one of the authors (McCombie), who, while estimating the Verdoorn law in the 1970s at Cambridge, UK, constructed estimates of regional capital stocks for the US.[6] Almost as an afterthought, he used these to estimate a conventional Cobb–Douglas production function for the two-digit SIC manufacturing industries for the US states' cross-regional data. Given the prevailing view at Cambridge, UK, at that time (namely, that it had been conclusively proved that the concept of aggregate production function was logically untenable), it came as quite

[6] The Verdoorn law is the relationship between the growth of industrial productivity and output and came to prominence in Kaldor's (1966) inaugural lecture as an explanation of the UK's slow rate of economic growth in the early postwar period. See McCombie et al. (2002).

a shock to find estimates of the output elasticities of labour and capital usually around 0.75 and 0.25, and $R^2$s of over 0.9. It immediately led to a careful check to see if an error in the estimation or the punching of the data on computer cards had been made; it had not. This was a puzzle at the time, as, given all the problems associated with the aggregate production function, these results seemed too good to be true. It was not until much later that he found the beginnings of a convincing answer to this conundrum, almost by serendipity, in the form of articles by Phelps Brown (1957) and Shaikh (1974, 1980).

But we are getting ahead of ourselves. In retrospect, McCombie's results merely confirmed the earlier cross-sectional results of Douglas (1948) and those of Hildebrand and Liu (1965). At about the same time, Moroney (1972) published a detailed neoclassical study estimating the production function using US state data that found similar good fits. In the early 1990s, something similar happened to Felipe, trying to estimate endogenous growth models using cointegration methods.

Time-series data do not always give good statistical fits to the aggregate production function, although adjusting the capital stock for the level of capacity utilisation generally improves the results and gives putatively plausible results. Douglas (1976, p. 914), in reviewing his studies on the aggregate production function commented, 'a considerable body of independent work tends to corroborate the original Cobb–Douglas formula, but, more important, the approximate coincidence of the estimated coefficients with the actual shares received also strengthens the competitive theory of distribution and disproves the Marxian'.

Consequently, the defence of the use of the aggregate production function rests largely on a methodological instrumental argument. All models involve unrealistic assumptions; after all, as Joan Robinson once remarked, a map on a scale of one to one is of no use to anyone. What matters is the explanatory power of the model, which is taken to be synonymous with its predictive power – the symmetry thesis (Friedman, 1953). Wan (1971, p. 71), for example, views the aggregate production function as an empirical law in its own right which is capable of statistical refutation, a view shared by Solow (1974). Ferguson (1969, p. xvii) explicitly made this instrumental defence with respect to the criticism about the measurement of capital as a single index in Cambridge capital theory controversies:

> Its validity is unquestionable, *but its importance is an empirical or an econometric matter* that depends upon the amount of substitution there is in the system. Until the econometricians have the answer for us, placing reliance upon [aggregate] neoclassical economic theory is a matter of faith. I personally have faith. (Emphasis added)

But all this does not explain *why* aggregate production functions generally give such good statistical results, especially in the light of Fisher's (2005, p. 490) warning:

> One cannot escape the force of these results [of the aggregation literature] by arguing that aggregate production functions are only approximations. While, over some restricted range of the data, approximations may appear to fit, good approximations to the true underlying technical relations require close approximation to the stringent aggregation conditions, and this is not a sensible thing to suppose.

The answer for cross-sectional data is to be partly found in an article by Phelps Brown (1957) 'The Meaning of the Fitted Cobb–Douglas Production Function', which ironically was published the same year as Solow's (1957) influential paper entitled 'Technical Change and the Aggregate Production Function'. Buried in Phelps Brown's paper is the argument that the regression estimates are not capturing any aggregate technological parameters of the economy (which almost certainly do not exist), but are merely picking up an underlying identity, namely, *that value added is, by definition, equal to the wage bill plus the total remuneration of capital*.

Theoretically, the aggregate production function represents a technological relationship and as such is a relationship between the output and inputs measured in *physical* terms. However, because of the problems of the heterogeneity of output and inputs, notably capital (but also labour, although it is often treated as being homogeneous), constant-price value measures have to be used. And therein lies the explanation of the good statistical fits. (Studies that actually use physical data, the so-called 'engineering production functions', are few and far between. See Wibe, 1984.)

There is an underlying accounting identity that holds for the *i*th firm and which is given by $V_i \equiv W_i + \Pi_i$, where $V$ is constant-price value added, $W$ is the total wage bill, and $\Pi$ denotes total profits. This identity can also be written as $V_i \equiv w_i L_i + r_i J_i$, where $w$ is the wage rate, $L$ is the employment, $r$ is the *ex post* or earned rate of profit and $J$ is the constant-price value of the capital stock, usually calculated by the perpetual inventory method. The identity also holds for gross output, where the value of output also includes the cost of materials. Furthermore, it holds at any level of aggregation, that is, for a sector or for the national economy. In fact, the National Income and Product Accounts (NIPA) show how the economy's total output is divided between wages and profits (the operating surplus). There is no assumption or theory (for example, Euler's theorem) behind this identity. It is important to emphasise that throughout the book we use $V$ and $J$ to denote the constant-price value measures of output (value

added) and the capital stock; while $Q$ and $K$ are the homogeneous physical measures of these variables.

This identity holds regardless of the state of competition, whether or not constant returns to scale prevail, and whether or not factors are paid their marginal products. In fact, it holds even if there is no well-defined production function at either the micro or aggregate level. One of Kaldor's (1961) stylised facts is that factor shares are constant over time. It is termed a stylised fact because while it is always possible to find exceptions to it, especially in the short run, these are rare. Constant shares can arise because firms pursue a constant mark-up pricing policy, for which there is a good deal of empirical evidence (Lee, 1998). They do not necessarily require an underlying Cobb–Douglas technology in physical terms, even if such a well-behaved production function actually exists. If we sum the individual firms' output arithmetically and, given that wages and the rate of profit are approximately constant across firms, we obtain for an industry the definition for value added that $V \equiv wL + rJ$, where $V = \sum V_i$, and so on. The aggregate factor shares are also likely to be roughly constant. (Solow (1958) has demonstrated that the aggregate factor shares may well be more stable than the individual sector shares.) It may be shown (see Chapter 3) that purely for arithmetical reasons, a close approximation to the linear accounting identity is given by:

$$V \equiv AL^{a}J^{(1-a)}, \tag{I.1}$$

where $a$ and $(1-a)$ are the labour and capital shares in output, respectively, that is, $a = wL/V$ and $(1-a) = rJ/V$; and $A$ equals $Bw^{a}r^{(1-a)}$, which is a constant provided that there is no variation in the wage rate or the profit rate across industries (or regions if we use spatial data). If equation (I.1) is estimated using cross-sectional or regional data with the coefficients unrestricted, then we are bound to get a near perfect statistical fit, and with the estimates of the coefficients equal to the factor shares. It is readily apparent that the equation is *formally identical* to the Cobb–Douglas 'aggregate production function' with constant returns to scale, and the 'output elasticities' equal to the observed factor shares, but it is not a production function. (If wages and the rate of profit show some variation, then this may bias the estimated parameters, although in practice this bias is likely to be small.) Thus the putative aggregate Cobb–Douglas production function will give a very close fit to the data, even though, for example, the aggregate production function may not exist, markets are not competitive and increasing returns to scale prevail.

The implications of this critique are far reaching. If good statistical fits to a functional form that resembles the Cobb–Douglas production

function (or, indeed, a more flexible production function) can be obtained using aggregate data that merely track the underlying identity, then it is not possible to interpret the statistical evidence as supporting the view that the relationship that has been estimated represents the technical conditions of production (such as the aggregate elasticity of substitution).

However, even though Phelps Brown's article was published in a leading journal, namely, the *Quarterly Journal of Economics*, it had almost no impact on the economics profession. Simon (with Levy) six years later published a formalisation of what could be taken to be Phelps Brown's argument. Nevertheless, Simon and Levy (1963) themselves were not entirely sure whether or not this was the case, as Phelps Brown's argument was admittedly somewhat obscure. Later, Simon (1979b) generalised the argument to explain why estimations of aggregate production functions using time-series data also give such good results and he also showed that the critique holds for other production functions, such as the CES. He thought these criticisms sufficiently important to mention them explicitly in his Nobel prize lecture (Simon, 1979a), but the message still fell on deaf ears. Simon was deeply sceptical of the marginal productivity theory of factor pricing as his correspondence in the early 1970s with Solow, recently unearthed by Carter (2011b), shows. In this correspondence, Simon pointed out to Solow the damaging implications of the accounting identity. (See also Felipe and McCombie, 2011–12.) To the best of our knowledge, there have been only three textbooks that have considered the argument, and only in so far as it relates to the cross-sectional (regional) data. These are Cramer (1969), Intriligator (1978) and Wallis (1979), but even here the full implications of the critique seem to have escaped these authors, who were perhaps more concerned with technical econometric issues. Intriligator, for example, merely notes that the identity will bias the estimates towards constant returns to scale, but not that it totally undermines the justification of the estimation of the aggregate production function in the first place.

Independently, Shaikh (1974) published an important short note similarly generalising the argument for the Cobb–Douglas to time-series data. This was, unfortunately and erroneously, dismissed by Solow (1974) in a one-page rejoinder, which began 'Mr Shaikh is wrong pure and simple'. This probably explains why little notice was ever paid to the paper. Shaikh's (1980) convincing rejoinder was eventually published in a book and not in the original journal, the *Review of Economics and Statistics*, which is why it is probably generally overlooked. The 2005 symposium on the aggregate production function published by the *Eastern Economic Journal* clarifies many of these issues.

The argument concerning the accounting identity is deceptively simple,

but these arguments made in the above articles are not the whole story. The criticism has also been subject to a number of serious misunderstandings and erroneous objections including Solow (1974, 1987) and Temple (2006, 2010). In this book, the critique is examined in some length, given its undoubted importance, and new arguments and evidence are presented that provide additional support for it. While in many cases attention is confined to the Cobb–Douglas for expositional ease, it is also shown that the critique applies to *all* aggregate production functions. *Moreover, it should also be stressed that even if there were no aggregation problems and output and capital could be accurately measured in value terms, the critique would still apply*. The only solution is to use physical magnitudes, and even then some insurmountable problems still remain, namely the correct specification of TFP and the level and rate of growth of technology.

Of course, macroeconomics abounds with identities, but these are explicitly recognised for what they are; namely, definitionally true relationships. Take the simple national expenditure identity, $Y \equiv C + I + G + Z$, where $Y$ is national income, $C$ is consumption, $I$ is investment, $G$ is government expenditure and $Z$ is net exports. No one would regress the growth of income on the growth of these variables, find a remarkably close statistical fit with the estimate coefficients having highly significant $t$-ratios (which would depend solely on the degree of stability of the share of the relevant variable in income) and contend that these results confirm the Keynesian theory of the importance of the role of the growth of demand in determining the growth of income.

Solow (1957, p. 312) comments that 'the aggregate production function is only a little less legitimate a concept than, say, the aggregate consumption function'. But this overstates the case. Let us take the simplest specification of the consumption function, $C = C_0 + b_1 Y^d$, where $C_0$ is autonomous consumption and $Y^d$ is private disposable income. However, there is also an underlying identity that relates consumption to savings and income, namely, $C \equiv S + Y^d$, where $S$ is private savings. It would be pointless to estimate either this or the transformation $\ln C = b_2 \ln S + b_3 \ln Y^d$. One reason for estimating the consumption function is to determine the value and the degree of stability of the marginal propensity to consume. This is based on the assumption that autonomous consumption is roughly constant (or grows at roughly a constant rate when time-series data are used). But it is usually fully appreciated that there is an underlying identity.

There are, of course, aggregation problems in constructing the data by summing over the individuals' income and expenditure. But these are much less severe than aggregating the diverse and complex production processes implicit in the aggregate production function. There is also

the important difference that theoretically the consumption function is a relationship between the deflated values of consumption and income. We are interested, for example, in estimating the increase in expenditure (in money terms) on consumption goods when disposable income increases by a certain value. However, in estimating the aggregate production function, as we have stressed and will discuss more fully in the book, we are using these value measures as a proxy for physical magnitudes. The parameters of the aggregate production function arise theoretically from engineering relationships, and the use of value data vitiates this interpretation of the estimates. Moreover, the aggregate production function has a number of implications, such as the marginal productivity theory of factor pricing and the distinction between the contribution to output growth of the growth of factor inputs and the rate of technical change, that are absent from the consumption function.

The content of this volume is as follows. Chapter 1, 'Some problems with the concept of the aggregate production function', summarises the problems underlying the concept of the aggregate production function, namely the aggregation problem and the Cambridge capital theory controversies. For reasons of space, these are dealt with only briefly.

Chapter 2, 'The aggregate production function: behavioural relationship or accounting identity?', outlines the central tenet of the book, namely that the aggregate production function is best regarded as nothing more than the mathematical transformation of an identity. This is the accounting identity that defines value added in terms of total wages and profits, or gross output when the value of intermediate inputs is taken into account. The question posed in the title of the chapter is, therefore, largely rhetorical. As we have mentioned, the basic tenet is deceptively simple, and in this chapter we set out the theoretical arguments in some detail. It is shown that it is not only the underlying identity that poses problems, but also the fact that constant-price value data are almost invariably used in estimating production functions. The problem is that while neoclassical production theory explicitly refers to a technological relationship between *physical* units of output, labour and capital (strictly speaking the flow of labour and capital services), applied work almost invariably relies on value measures which pose the insuperable problem. Because of the underlying identity, it can be shown theoretically that the best statistical fit to a supposed aggregate production function will be given when there are constant returns to scale and the 'output elasticities' equal the respective factor shares. Of course, many statistical estimates of supposed aggregate production functions do not give perfect statistical fits. However, from the accounting identity and the Kaldorian stylised facts we can show, a priori, why this is the case. Moreover, in some circumstances, we can determine

the direction of bias of the estimated coefficients before a single regression has been run. It is also possible to find the transformation of the identity that will give a perfect fit to the data.

In this chapter we also deal briefly with some common objections that have been made to us (both in seminars and in some referees' reports) and show that they are all based on fundamental misunderstandings of the argument. Temple (2006, 2010) is the only person who has considered the argument in print in any detail. While he sees some merit in the argument, he does not find it convincing. However, his comments are not compelling but, nevertheless, are instructive to the extent that if they are implicitly shared by other economists, they go a long way to explain why the critique has not had the impact it should have had. We reflect briefly on his comments in this chapter, but save a detailed consideration until Chapter 12. In the appendix to Chapter 2, we present an example using regression analysis that illustrates empirically the problems posed by the accounting identity. We also show explicitly that the critique applies to more flexible functional forms including the CES and the translog. Temple erroneously maintains that it applies only to the Cobb–Douglas production function, and hence that the argument has to rely on the *ad hoc* assumption (actually a stylised fact), *inter alia*, that factor shares are constant.

One of the problems is that the researcher can, for the vast majority of the estimations of production functions, only use value data and hence has no idea of the true underlying technological relationships. No one would deny that production functions exist in the sense that the volume of physical output is determined by the inputs of materials, labour of various skills and the vast number of different types of capital. (There are, of course, other problems associated with those large sectors of the economy, such as finance, services, government and local authority services, where there is no measure of output totally independent of the inputs. But we ignore this complication for the moment.)

Production relationships are likely to be very complex and differ from firm to firm, even those making the same product or producing the same service. However, as we have already noted, the problem is that the researcher simply does not have these physical data. One way out of this impasse is to use simulated data where, by construct, we do know the hypothetical underlying technology. This has the advantage of allowing us to demonstrate explicitly the extent of the problem. This is what Chapter 3, 'Simulation studies, the aggregate production function and the accounting identity', does. We start with some simulations of our own. To begin with, we assume that there are 'true' underlying Cobb–Douglas micro-production functions expressed in physical terms, but where the output elasticities of labour and capital are constructed to be 0.25 and 0.75,

respectively. In other words, they differ from labour's and capital's factor shares (0.75 and 0.25, respectively). However, if firms follow a mark-up pricing where the mark-up is 1.333, the statistical results produce erroneous estimates of the factor shares equal to the observed output elasticities. This is even true when the underlying micro-production functions exhibit increasing returns to scale or where there is no well-defined relationship between output, labour and capital. The chapter also considers a number of other simulation studies where the hypothetical data give good statistical fits to the data, even though the underlying micro-production functions are nowhere near being of the Cobb–Douglas form.

Chapter 4, '"Are there laws of production?" The work of Cobb–Douglas and its early reception', is a step back in time and as the title suggests looks at the early reception of Cobb and Douglas's initial work. Today, it is often forgotten just how critical was this reception of their early studies, on both econometric and other grounds. This reception was so hostile that Douglas admitted that he almost lost heart and nearly gave up entirely his work estimating aggregate production functions. This chapter, though, is more than just an exercise in the history of economic thought, as it is shown that the accounting identity critique goes back, albeit in a rudimentary form, many years.

Chapter 5, 'Solow's 'Technical change and the aggregate production function' and the accounting identity', and Chapter 6, 'What does total factor productivity actually measure? Further observations on the Solow model', continues this theme. Solow's (1957) paper, along with his companion theoretical paper (Solow, 1956),[7] proved to be immensely influential in the subsequent development of neoclassical growth theory. But it is not generally realised how shaky are the foundations of Solow's model. We discuss Shaikh's (1974) provocative Humbug critique, where he shows that the method Solow used to 'correct' the production function for technical change is essentially tautological. As such, the resulting specification cannot but give a near perfect fit to the data. Shaikh showed that this was the case even using a hypothetical dataset where the scattergram of productivity on the capital–labour ratio spells out the word 'HUMBUG'. Perhaps more importantly, Shaikh also shows that Solow's model is subject to the accounting identity critique using time-series data. Chapter 6 shows how the identity is responsible for the evidence that has been used by a leading growth theory textbook to justify the empirical relevance of the Solow model. This chapter shows that the high explanatory power of Solow's model is very misleading.

---

[7] Swan (1956) also independently developed a similar model.

The next five chapters provide empirical examples of where the accounting identity is largely, or entirely, responsible for generating the results of the estimation of the theoretical model. The Mankiw–Romer–Weil model (1992) was an influential extension of the Solow model. Chapter 7, 'Why are some countries richer than others? A sceptical view of Mankiw–Romer–Weil's test of the neoclassical growth model' shows why. It will come as little surprise to learn that all that the statistical fits of the world aggregate production function are capturing are the underlying accounting identity and the Kaldorian stylised facts. Chapter 8, 'Some problems with the neoclassical dual-sector growth model' demonstrates how the accounting identity, together with the national expenditure identity, is responsible for the empirical results that suggest that there are substantial externalities to the growth of exports and/or government expenditure. In fact, the regression results, because of the accounting identity, can shed no light on the existence, or otherwise, of externalities.

Oulton and O'Mahony (1994) use a large database of UK manufacturing industries to test whether or not capital is special. By 'special', they mean whether the output elasticity of the capital stock exceeds its factor share. If this is the case, they argue that it lends credence to the endogenous growth model where the growth of capital has a substantial externality effect. They find it doesn't but this should come as no surprise, as Chapter 9, 'Is capital special? The role of the growth of capital and its externality effect in economic growth' shows, given the existence of the accounting identity and the fact that they are using value data. In Chapter 10, 'Problems posed by the accounting identity for the estimation of the degree of market power and the mark-up', we consider Hall's work, where he attempts to estimate the degree of market power using the aggregate production function and the Solow residual. Hall finds that the estimate of labour's output elasticity exceeds its observed factor share, which he interprets as evidence of market power. We know, a priori, that because of the accounting identity, Hall should not have found this discrepancy, and the intriguing question is why? The answer is that he assumes that technical progress occurs at a constant rate. From the identity we know that the rate of technical progress is nothing more than the growth of the weighted wage rate and the rate of profit, which has a strong cyclical component. This causes the estimate of the mark-up to exceed unity. The last example we look at is in Chapter 11, 'Are estimates of labour demand functions mere statistical artefacts?', which considers various estimations of the neoclassical labour demand functions. It is shown that the negative relationship between the logarithm of employment and the logarithm of the real wage is likewise driven by the underlying identity and has obvious policy implications.

These examples are drawn from previously published articles of the authors and we have condensed and somewhat simplified the various arguments for this volume. Consequently, the reader is invited to consult the originals for a more detailed analysis.

We can liken these chapters to the game of 'Where's Waldo?'. This is a children's game where the character Waldo in his distinctive red and white shirt is hidden in a picture among a large number of other colourful characters and the task is to find him. Our examples are, of necessity, eclectic, and we leave it to the reader, having read the book, to see if he/she can spot Waldo (the accounting identity that drives the estimates), in other papers that use the aggregate production function. We close the book with Chapter 12, 'Why have the criticisms of the aggregate production function generally been ignored? On further misunderstandings and misinterpretations of the implications of the accounting identity'. The chapter also includes a detailed discussion of the ancillary issue of the persuasiveness of those few criticisms of our arguments that have been voiced. We have yet to find any such critiques, including those of Temple (2006, 2010), in the least bit compelling – but we are, of course, content to let the reader make up his or her own mind.

We have decided not to add a chapter on what to put in place of the aggregate production function. We have discussed this question in a number of seminars and presentations. Answering this sixty-four thousand dollar question would in itself take another book or, rather, several books. Other approaches to growth that do not rely on a production function do exist, including case studies and the insights provided by economic historians such as Landes (1998). The aim of this volume is much more limited – it is merely to show that the need for an alternative approach is long overdue. In other words, the above discussion is what Lawson (2004) terms 'an exercise in under-labouring'; or as John Locke (1690) put it: 'removing some of the rubbish that lies in the way of knowledge' (cited by Lawson, 2004, p. 317). But as Kuhn (1962 [1970]) pointed out, one paradigm, no matter how logically or empirically flawed, is only abandoned if it is replaced by another paradigm. Nevertheless, as we have mentioned above, it is beyond the scope of this volume to discuss alternative approaches. We merely hope to have made the case for serious consideration to be given to alternative approaches.