Aggregate production functions and growth economics*

Jonathan Temple+

Department of Economics, University of Bristol
8 Woodland Road, Bristol BS8 1TN, UK and CEPR

Rigorous approaches to aggregation indicate that aggregate production functions do not exist except in unlikely special cases. This paper considers the awkward implications for growth economics. It provides a conventional defence of growth theory in terms of 'parables' and then considers how empirical growth research might avoid the need for aggregate production functions.

Keywords: aggregation, production functions, growth econometrics

* My interest in this topic arose from the presentation of Jesus Felipe at the conference "Understanding economic growth: new directions in theory and policy", Cambridge, 1-3 September 2005. I am grateful to Luis Correia and an anonymous referee for insightful comments on a previous draft of the paper. I have also benefited from discussions on related topics with Amitava Dutt, Jesus Felipe, Geoff Harcourt and Ludger Woessmann. I would like to thank the Leverhulme Trust for financial support under the Philip Leverhulme Prize Fellowship scheme. The usual disclaimers apply.

+ Email jon.temple@bristol.ac.uk

Aggregate production functions and growth economics

1. Introduction

This paper aims to say something new about an old topic: the role of aggregate production functions in macroeconomics, and especially their role in growth economics. The formal theory for aggregation of economic quantities indicates that aggregate production functions do not exist, except in unlikely special cases. Is there anything more to be said?

I will argue that there is. Before doing so, section 2 of the paper will briefly review the issues raised by aggregation. This will concentrate on the technical aspects of aggregation of heterogeneous quantities - whether output, capital, or labour – that apply when taking mainstream theory at face value. The section also briefly discusses the broader critique of mainstream theory that was an important part of the "Cambridge controversies" of the 1960s.

Section 3 will then consider the role of aggregate production functions in economic theory. It offers a conventional defence that models are typically "parables" that can be useful even when their assumptions are unrealistic and cannot be formally justified. As long as the limits to theory are recognised, this suggests that the aggregation critique is often more serious for empirical research than for theoretical modelling.

Section 4 discusses some of the issues raised by estimation of production relationships, while section 5 reviews the specific implications for empirical growth

research. One argument is that growth econometrics, at least as usually carried out, does not necessarily rely on aggregate production functions. This section also emphasizes the scope for answering growth questions indirectly. These indirect methods sometimes avoid the need for aggregation, and have other benefits relative to regression-based approaches. Finally, section 6 concludes with a brief summary, and some speculations about the relevance of aggregation to other macroeconomic debates.

2. Aggregation

There is a long history of theoretical work on aggregation, not all of which is well-known among macroeconomists. The best-known result requires all firms to use the same production technology, face the same factor prices, use inputs efficiently, and work under conditions of perfect competition. Then the aggregate production function will just be a scaled-up version of the firm-level production functions. The simplicity of this 'representative firm' approach is appealing, and as a modelling device, it can be defended using the arguments advanced in section 3 below.

The representative firm approach is clearly not much use if we want to consider aggregates in economies that are less artificial. A useful way to see the problems that arise is to consider an economy with two sectors producing distinct goods (agriculture and non-agriculture, say). It is easy to show that if the two sectors

each have Cobb-Douglas production technologies, and if the exponents on inputs differ across sectors, there cannot be a Cobb-Douglas aggregate production function.[1] More generally, since the allocation of labour across sectors is endogenous, and the equations that define this allocation are typically non-linear and cannot be solved analytically, it will rarely be possible to write down any closed-form expression for an aggregate production technology.

In fact, the news gets worse and worse. An applied macroeconomist or growth economist might often want to write down a function that relates "output" to indices of "capital" and "labour", where these indices aggregate several different types of each input. The conditions under which this is possible are extremely restrictive, something established in a series of contributions by Fisher (collected in Fisher 1993) and developed further by Blackorby and Schworm (1988). Felipe and Fisher (2003) provide an especially accessible review. A reasonable interpretation of this work is that aggregate production functions - and other aggregate quantities such as capital, investment and "the" input of labour - cannot be meaningfully defined in any circumstances that might apply to a real-world economy. This must also be true of "the" marginal products and "the" elasticity of substitution. Put simply, they don't exist.

Before turning to a defence of growth theory, it is worth emphasizing how pervasive these problems are. Aggregation is inherently problematic regardless of whether or not mainstream theory is taken at face value, and even when there is only

[1] The way to see this is to write down the aggregate labour share as a weighted average of labour shares in the two sectors. If the structure of output changes, the weights and the aggregate labour share will also change, and hence there cannot be an aggregate Cobb-Douglas production function (which would imply a constant labour share at the aggregate level).

one type of capital good. In the hazy collective memory of current economists, aggregation problems are instead often associated with the "Cambridge controversies" of the 1960s, and believed to relate mainly to some awkward but surmountable technicalities in the measurement of capital and the determination of the real interest rate. Against this view, Cohen and Harcourt (2003) argue that the profession has shown a kind of willed amnesia about the outcome of the debates and their consequences for mainstream theory, especially aggregate models that relate factor payments to factor scarcity via marginal products. These debates range over issues much broader and deeper than those I consider here, and interested readers should instead consult the excellent overview in Cohen and Harcourt (2003).

3. Aggregation and the uses of error

I will start by setting out an extreme position. Since aggregate production functions do not exist except in unlikely special cases, any economic theory that makes use of them is of no scientific value. Any researcher willing to place a false premise at the heart of their analysis can draw no useful conclusions.

This argument makes me uncomfortable, and not just because some of my past research would have been a waste of everybody's time. My objection is more fundamental: I think the argument misunderstands the nature and purpose of economic theory and empirical research. The arguments put forward here are rather conventional, but worth setting out as a response to some critics of mainstream economic theory. In essence, the argument is that the critics ask too much of theory.

I think there would be fairly general agreement that some economic debates are unlikely ever to be definitively resolved. This is the case however many theoretical models are developed, however many models are calibrated, and however many regressions are run. For many economic questions - I suspect most - scientific standards of proof are not attainable. The most we can ask of a research paper is that it shifts the burden of proof away from one side of a debate and towards another. One can phrase this in different ways: a useful paper may offer new "insights" or provide a new and more sophisticated way of thinking about a problem. Some of the assumptions will be questionable or false, but readers will come away with a modified view of the world.

It is clear that many disciplines proceed on exactly this basis. The study of history might be a good example. If we imposed scientific standards of proof, all meaningful study of history would have to cease: historians would have to confine themselves to listing verifiable facts, and cease to organize or interpret those facts. It should be clear that this requirement would be too extreme. After all, even in a world without academic historians, people would continue to make decisions, or think about the world, in a way that embodies a particular view of the past. If rigorous observers were to impose the most exacting standards of proof on historians, effectively imposing silence, we would be left with debates and views in which nobody examined their own positions, and in which instincts and prejudices were left permanently unchallenged.

Academic research in many disciplines can only *aspire* to be scientific. Whether applied to history or economics, this argument is often regarded as dangerous, because it risks excusing careless or slapdash research - work that is not

genuinely serious. But there is nothing in the argument that suggests that researchers should take their work, and their conclusions, less than seriously. The analysis of history should proceed as rigorously as possible and, ideally, historians should seek out evidence that might falsify their interpretations or their view of the past. The point is that the very real possibility of error is not enough to make silence the best research strategy. Another implication is that researchers have to be keenly aware of the limits to their interpretations.

The reason for this lengthy detour is that it potentially justifies the use of aggregate production functions as a way of learning about the world. This is the 'parable' defence associated with Samuelson (1961-62) and especially Solow (1966). Since the above remarks have been very general, I will try to give a specific example of this defence, before expanding on it.[2] The example relates to a long-standing debate: can we explain economic underdevelopment simply in terms of low investment? In a classic paper Lucas (1990) showed that, under conventional assumptions about the extent of diminishing returns, the vast differences in labour productivity we observe across countries cannot be explained by differences in capital intensity, without a counterfactual implication. If differences in capital intensity account for underdevelopment, the returns to investment in poor countries would have to be many times the returns in rich countries - to a far greater extent than is usually thought plausible.

One response to the Lucas paper is to say that, because his conclusions are derived from an aggregate production function, it is of no value. I think that is

[2] Some additional discussion, to the effect that theoretical growth models are often best interpreted as artificial "laboratories" for thought experiments, can be found in Temple (2003).

clearly wrong: Lucas has shifted the burden of proof away from one side of the debate and towards another. Opinions on the size of this shift will differ, but anyone who believes that capital intensity is central will have to make some effort to engage with his argument.

It might well be argued that his result is an artifact of aggregation, neglecting heterogeneity in inputs, outputs and technologies. But it is not at all obvious that this heterogeneity is going to be enough, since the underlying mechanism - diminishing returns - is not eliminated by heterogeneity. To be genuinely convincing, that criticism would need to be expanded upon, with examples or simulations, before we really began to discount the points made by Lucas. Put differently, it is not enough to say that the model's simplicity nullifies its conclusions; instead, there also needs to be some attempt to explain why the underlying logic will not apply in more complex and realistic circumstances.[3]

In practice, his points have been answered in other ways. An implication of the analysis in Ventura (1997) and Robertson (1999) is that two-sector models will feature less variation in the returns to capital, over time and space, than a closed economy Solow model. This is related to a standard result in simple 2 x 2 trade theory models, since the marginal product of capital will typically be independent of factor endowments while the economy remains incompletely specialized. Another relevant consideration is that the relative price of capital goods differs substantially across countries - a form of heterogeneity that is surely important, as discussed in

[3] From the perspective of one side of the Cambridge controversies, however, the assumption that the marginal product of capital always moves inversely with the capital-labour ratio is problematic. See Cohen and Harcourt (2003, especially p. 206-207) for discussion and references.

Cohen and Soto (2002) and Hsieh and Klenow (2003). These responses all have in common that they, too, start from relatively basic models, ones that are surely too simple in their assumptions, and that implicitly require aggregation properties that would never be satisfied in the real world. But I think we have learned something about the role of investment in development from Lucas (1990) and the work that has followed.

This is not to claim that heterogeneity is unimportant in explaining real-world phenomena. Melitz (2003) is one example of a paper that makes elegant and insightful use of a form of heterogeneity, showing how differences in costs, and the possibility of reallocations of output among firms, might modify our understanding of the effects of trade. Models with the same basic structure can be applied to other issues, including many questions of interest to growth economists. It is natural to think that such models will cast new light on the effects of heterogeneity, and modify some previous conclusions, at the expense of some restrictive assumptions. But none of this implies that macroeconomics without heterogeneity is inevitably misleading.

Recall the view that, when aggregation is not justified, it should be avoided. This risks silence on important questions. In many cases the only alternative, to address those questions, would be to construct a more complicated model that captured more aspects of reality - including differences in technologies and outputs across sectors, many different types of input, and so on - and solve it using a computer.

This can be done, as in models in the computable general equilibrium tradition, and is sometimes very informative.[4] In principle if not in practice, it is possible to envisage a model that could mimic the real world perfectly. Such a model would not be useless, as is sometimes implied. Instead, given that it must embed an accurate depiction of the formation of expectations, it could be used to simulate the path of the economy under alternative policies and histories.

In practice, a model will always fall short of this benchmark. As soon as this happens, it is important to be able to understand the model: the researcher needs to know which are the key assumptions, which results might not be robust to minor modifications, which results are likely to be more general, and so on. Without a good understanding of the model, it will be difficult to understand the world. And even with the most complex and sophisticated model - or perhaps especially then - that understanding of the model is likely to derive at least partly from pencil-and-paper exercises, back-of-the-envelope calculations, simple theories, toy models - in short, everything that we might understand by Solow's use of the term "parables".

4. Humbug production functions

One of the traditional defenses of aggregate production functions is a pragmatic one: they may not exist, but empirically they 'seem to work'. For example, attempts to estimate Cobb-Douglas production functions, at various levels of

[4] Work on agent-based computational economics is extending the reach of computer-based models further; see Tesfatsion and Judd (2006).

aggregation, often find returns to scale that are approximately constant and estimates of the input elasticities that are close to factor shares.

There is a long tradition of criticising the estimation of production functions on econometric grounds. The level or growth rate of technical efficiency usually has to be omitted from the empirical model. And because input choices are likely to reflect technical efficiency, this omitted variable is almost certainly correlated with some of the explanatory variables in the regression, such as the growth rates of inputs. As a result, estimates of the technology parameters will be biased.

I will call this the 'statistical' critique of the approach. Classic references include Marschak and Andrews (1944) and the very useful review by Griliches and Mairesse (1998). The most detailed treatments for cross-country growth regressions can be found in Benhabib and Jovanovic (1991) and Benhabib and Spiegel (1994, Appendix A.3). These papers show that, when the growth rate of output is related to the growth rate of inputs, and the growth of technical efficiency is omitted, the effect of physical capital accumulation may easily be overstated. In the microeconometric literature, and to a lesser extent in the growth literature, one response has been to use panel data estimators with fixed effects and/or instrumental variables, but that is subject to the problems discussed in Griliches and Mairesse (1998) and Temple (1999).

Another set of criticisms - an 'economic' critique - is more sweeping. The argument is that, because of the underlying value added identity, a constant-returns Cobb-Douglas production function will often appear to explain time series data successfully. This can happen even if no genuine production relationship exists. I will briefly summarize the argument: value added is equal to

$$Y \equiv wL + rK$$

where w is the wage, L is the labour input, K is capital, and r is the rate of profit (or whatever multiple of K will ensure the identity is satisfied). Note that this is an identity, which simply says that value added, by creating revenue, necessarily generates income for either capital or labour. In particular, no assumption has been made that returns to scale are constant or that factors are paid their marginal products, and neither of those assumptions will be required for the main argument that follows.

This identity can be differentiated with respect to time to give

$$\frac{\dot{Y}}{Y} = \frac{\mathrm{wL}}{\mathrm{Y}}\frac{\dot{\mathrm{w}}}{\mathrm{w}} + \frac{\mathrm{rK}}{\mathrm{Y}}\frac{\dot{r}}{r} + \frac{\mathrm{wL}}{\mathrm{Y}}\frac{\dot{\mathrm{L}}}{\mathrm{L}} + \frac{\mathrm{rK}}{\mathrm{Y}}\frac{\dot{K}}{K}$$

If we write the capital share as α=rK/Y then we can write a measure of TFP growth as

$$(1) \qquad \frac{\dot{Y}}{Y} - \alpha(t)\frac{\dot{K}}{K} - (1-\alpha(t))\frac{\dot{L}}{L} = \alpha(t)\frac{\dot{r}}{r} + (1-\alpha(t))\frac{\dot{\mathrm{w}}}{\mathrm{w}}$$

This is just a simple illustration of the "dual" growth accounting results, namely that TFP growth can be calculated either from quantities, or from factor prices. Jorgenson and Griliches (1967) develop this in detail. Note that the

interpretation of these specific expressions as measuring technical change does require constant returns, perfect competition, and that factors are paid their marginal products.

The economic critique essentially asks the following question: if we assume there is no production function, or perhaps a relationship of complex and unknown form, what happens when we estimate a production function from time series data? Imagine that for some unspecified reason, factor shares are constant and hence $\alpha(t)=\alpha$. Also imagine that the weighted average of the wage and profit rate growth rates is constant, equal to λ say. Arguably the most likely case would arise if wages simply trend smoothly upwards, while profit rates are constant, but the weighted average could be constant under more general conditions. Then (1) implies that

$$\frac{\dot{Y}}{Y} = \lambda + \alpha\frac{\dot{K}}{K} + (1-\alpha)\frac{\dot{L}}{L}$$

and so the data could easily appear to have been generated by a Cobb-Douglas aggregate production function with exponent α on capital and 1-α on labour. This argument has appeared in various forms, and Shaikh (1974) memorably characterized the implications: estimated production functions can produce only "humbug".[5]

Some interpretations of this result become overenthusiastic and suggest that a Cobb-Douglas technology will always fit the data well, simply because of an identity.

[5] For a more complete treatment of these ideas, and additional references, see Felipe and Holz (2001) and Felipe and Fisher (2003).

This should make us pause: for example, if the underlying technology was translog, could we really expect a Cobb-Douglas to fit the data well? Given sufficient variation in the input ratios, movements in factor shares would immediately reveal that Cobb-Douglas is not the right specification.[6] The argument that Cobb-Douglas results are spurious uses not only the value added identity, but also some additional structure: namely constant factor shares and constancy of the weighted average of the wage and profit rate growth rates.

The need for this extra structure points to the heart of the problem in estimating production relationships. Estimation must usually treat the level or growth rate of technology (TFP) as unobservable. And it is this omitted variable that poses the fundamental difficulty. If the data were generated by a translog, and the researcher had identified a good proxy for TFP, a suitably specified regression would accurately recover the parameters of that translog production function, and reject the Cobb-Douglas specification given sufficient variation in the data. It is the inability to control for the TFP term that causes problems, and this means the 'statistical' and 'economic' critiques are closer together than is usually acknowledged.[7]

What the economic critique adds to the more conventional (statistical) arguments is the point that, if you are unable to control for TFP, you may get results

---

[6] It is worth noting in passing that, along the balanced growth path of a deterministic Solow model, factor shares would be constant even with a translog aggregate production function. But then there is no variation in factor input ratios either (at least in efficiency-unit terms) and therefore no hope of identifying any underlying production relationship.

[7] This point tends to be obscured because critics in the Shaikh (1974) tradition proceed on the assumption that there is no underlying production relationship. The argument in the text is that the statistical and economic critiques share some common ground: unless the researcher can find a way of controlling for differences in efficiency, the results from estimates of production relationships can be highly misleading. The economic critique goes further than this, however, pointing out that apparently impressive results can arise even when no underlying production relationship exists.

from estimating a production function that are deceptively promising, especially over time periods where factor shares are broadly constant for whatever reason. The argument can even be taken a step further. It seems highly unlikely that the elasticity of substitution between capital and labour is always and everywhere equal to unity. In consequence, any evidence that a constant-returns Cobb-Douglas fits well might sometimes be worrying, since it could reflect a combination of the underlying value added identity and unmodelled variation in TFP. Overall the critique has some force. It deserves to be more widely known among researchers estimating production relationships using time series or panel data, including researchers who never doubt the existence of a well-behaved underlying relationship.

This all raises the question of whether it is ever sensible to estimate production functions, using data at the national, regional, industrial or even firm or establishment level. The most common motivation for estimating these relationships is to obtain information about technology or TFP; but without that information to start with, and hence without the possibility of controlling for TFP, any effort at estimation is on dangerous ground. There is a sense in which the estimation of production technologies is required on precisely the occasions when it is guaranteed not to work. That is perhaps a little too strong, and section 5.4 below will discuss some estimation methods that can accommodate unobserved differences in technology.

The arguments made so far relate to time series and panel data estimation. It is tempting to think that the argument applying over time could also be applied over space. This could explain why some researchers find good results when using Cobb-Douglas technologies to explain international output differences, and a recent paper

by Felipe and McCombie (2005) is directed at this question. In particular, they seek to explain why the empirical models in Mankiw, Romer and Weil (1992) work as well as they do. The Felipe and McCombie argument has to proceed under restrictive assumptions, and overall it seems much harder to apply the 'humbug' argument in the context of cross-sections. The problem of unobserved technology differences remains central, however.

5. Aggregation and empirical growth research

In this section, I take as given that (1) aggregate production functions do not exist, and (2) most strategies for estimating production relationships, at whatever level of aggregation, are intrinsically problematic. The section asks the following question: how essential is the artificial device of the aggregate production function to empirical growth research? I first discuss growth models that disaggregate inputs in tractable ways (section 5.1) or that disaggregate output, as in dual economy models (section 5.2).

Section 5.3 considers whether production functions are essential to empirical growth research as usually conceived. The section argues that the methods used in practice do not rely on aggregation, and usually represent ad hoc formulations that are nevertheless informative. Section 5.4 emphasizes the potential for learning about the extent and sources of international productivity differences through indirect routes - for example, structural trade models - that do not rely on aggregate production relationships.

5.1 Disaggregation

If aggregation is not possible, the obvious solution must be to disaggregate. Many of the empirical frameworks used by growth researchers do not intrinsically require aggregation of different kinds of input, for example. In the case of growth accounting, there is nothing to stop the researcher writing down

$$Y = F(K_1, K_2, ..., K_M, L_1, L_2, ..., L_N)$$

where there are M different types of capital input and N different types of labour input. This approach has been developed and made operational by Jorgenson and co-authors in a series of contributions, some of which are collected in Jorgenson (1995). This makes clear an important point: production theory and growth accounting do not inherently require aggregation of different types of input, or for that matter, a single form of output. Instead, it is lack of data that will typically restrict the applied researcher to simpler methods.

In principle, growth regressions can also accommodate different types of input, including heterogeneous capital goods. For example, in the classic analysis of Mankiw, Romer and Weil (1992), human capital is modelled precisely as a second kind of capital good, even sharing the same rate of depreciation as physical capital. In their model, a unit of output can be transformed into a unit of physical capital or a unit of human capital, the only cost to this process being foregone consumption.

Temple (1998) applies the same idea to derive an empirical growth model with three different types of capital: equipment, structures, and human capital.

This simplicity makes the model tractable, but comes at a cost. In the absence of externalities, it would be natural to impose the restriction that returns to holding the different types of asset are equal, but this will tend to complicate the analysis. A closely related issue is that a genuinely convincing model would need to acknowledge variation in the relative prices of capital assets. As it stands, the analysis in Temple (1998) does not allow the relative prices of different assets to vary, whether over space or over time, and this is clearly an important omission.

All this suggests that it would be interesting to explore regression frameworks that disaggregate inputs in more sophisticated ways. Greenwood et al. (1997) introduce a tractable model that might be a suitable basis for this kind of work. Temple (2001a) considers empirical growth models in which the labour input is disaggregated into skilled and unskilled workers, where the former may have higher productivity.

5.2 Dual economies

Discussions of aggregation often risk implying that the key problem is the aggregation of capital. As Fisher (1993) repeatedly emphasizes, aggregation is also a problem for other inputs, including labour, and for output. This suggests that growth economists should be looking for ways to carry out their analyses at a lower level of aggregation.

At least for developing countries, one of the most interesting approaches is to disaggregate the economy into agricultural and non-agricultural sectors, each with distinct outputs, as in the dual economy tradition. Data on agriculture's share of output and employment are easy to obtain. Although these data have limitations, it is remarkable how rarely information on sectoral structure has been used to inform analyses of differences in output levels and growth rates. Among the flurry of empirical growth papers in the 1990s, Dowrick and Gemmell (1991) is a lonely exception. Ros (2000) and Temple (2005) review the implications of dual economy models for growth economics, and empirical research in particular.

One issue deserves special mention. Consider a small open economy with two sectors, in which both goods can be traded at world prices. Output in each sector is produced using capital and labour, with constant returns to scale. The inputs are mobile between sectors, so that in equilibrium, wages are equal in both sectors, as are returns to capital. The TFP parameter in agriculture is A_a, and in non-agriculture A_m.

It is then possible to derive growth accounting results for this dual economy as a special case of the results in Jorgenson and Griliches (1967). In particular, using Divisia indices for quantities and prices leads to the following expressions for the growth accounting residual:

$$(2) \quad \frac{\dot{A}}{A} = \frac{\dot{Y}}{Y} - \alpha(t)\frac{\dot{K}}{K} - (1-\alpha(t))\frac{\dot{L}}{L} = s(t)\frac{\dot{A}_a}{A_a} + (1-s(t))\frac{\dot{A}_m}{A_m}$$

where $\alpha(t)$ is capital's share of national income and s(t) is the ratio of agricultural output to total output (at domestic prices). Note that both of these shares depend on time. In particular, given the two-sector structure of the model, the aggregate factor shares will tend to vary across countries and over time, even if the sectoral production functions are both Cobb-Douglas. As noted earlier, this is because the aggregate factor shares will be weighted averages of the sectoral factor shares, with weights equal to the shares of each sector in total value added.

The final equality in (2) shows that the growth accounting residual is a weighted average of TFP growth in the two sectors, where the weights are equal to the time-varying shares of the sectors in total value added. This simple result arguably deserves more attention than it receives in the empirical growth literature.[8] To see its implications, recall a frequent assumption of the empirical growth literature, namely that efficiency growth is the same across countries. This is typically justified on the basis that technologies can be easily transferred across national borders. In a two-sector world, the constancy of efficiency growth no longer follows from the possibility of technology transfer, except in unlikely special cases. For similar reasons, "levels accounting" decompositions now acknowledge that differences in sectoral structure could play an important role in thinking about TFP differences. Temple (2005) provides more discussion and references.

More generally, models of small open economies with two sectors suggest other ways of generalizing standard models in a tractable way. Landon-Lane and Robertson (2003) and Temple (2005) discuss the empirical implications of combining

[8] It can be seen as a special case of the more general principles of Domar aggregation. See, for example, Jorgenson and Stiroh (2000). The result given here is discussed, and generalized to distorted economies, in Temple and Woessmann (2004).

the Solow model with a version of Lewis (1954), the aim being to develop an empirical growth model that could apply to developing countries. Temple (2001b) considers how dual economy models might be used to understand the Golden Age of rapid growth in post-war Europe, following in a long tradition that includes Cornwall (1977) and Kindleberger (1967). Another advantage of dual economy models is that they can be used to distinguish between different types of growth and their effects on inequality; see Temple (2005) for more discussion.

5.3 Growth econometrics without production functions

The reason aggregate production functions are so routinely used is that disaggregation is not always possible, and even where possible, may not lead to a tractable model. It is therefore worth considering whether aggregate production functions are genuinely essential to the empirical analysis of growth.

As is well known, a wide variety of growth models all lead to an expression of the form

$$\log y(t) - \log y(0) = \theta \log y^* - \theta \log y(0)$$

where y is output per worker in terms of efficiency units (Y/AL in standard notation), y* is the steady-state level of y, and θ is a parameter related to the speed at which an economy closes the gap between its present position and the steady-state growth path. In the most familiar growth model, the steady-state level of the growth path is a simple function of the saving rate (s), the population growth rate (n), the rate of efficiency growth (g) and the depreciation rate (δ):

$$(3) \qquad y^* = \left(\frac{s}{n+g+\delta} \right)^{\frac{\alpha}{1-\alpha}}$$

as in Mankiw, Romer and Weil (1992). This leads to a growth regression in which the explanatory variables are the initial level of output per worker, the logarithm of the saving rate, and the logarithm of (n+g+δ). A great strength of this approach is that growth can be modelled as reflecting a process of capital accumulation, without the need ever to measure the level of the capital stock. The initial level of output per worker can be seen as a proxy for the initial capital-output ratio. But there is a problem, clearly articulated in Islam (1995), and which should now be familiar: in the absence of implausible assumptions, the parameter estimates will be biased because of an omitted variable, technical efficiency.

More generally, most growth regressions can be interpreted as something close to:

$$(4) \qquad \log y(t) - \log y(0) = \theta \beta' x - \theta \log y(0) + \varepsilon$$

where β is a vector of parameters and $x=(x_1, x_2, ..., x_K)'$ is a vector of variables that are thought to influence the steady-state level of the growth path.

This helps to clarify a key point. The specific expression for y* given in (3) follows from an aggregate Cobb-Douglas production function combined with the assumptions of the Solow model. But this need not be true of the more general

empirical framework represented by (4). That regression still holds considerable interest in the absence of an aggregate production function. It essentially asks whether, for two countries starting from the same position in terms of output per worker, the country with a higher level of a given x_i grows more rapidly, all else constant.

At least implicitly, this allows the researcher to recover the extent to which the level of the long-run growth path depends on the different components of the vector x. In essence, all we have is a model that combines partial adjustment to a long-run equilibrium, with a specific model for that equilibrium. There is nothing here which intrinsically relies on aggregate production functions, except perhaps the assumed homogeneity of the θ parameter across countries – and even this assumption could be relaxed using interaction terms and other standard methods.

An alternative to the specification (4) is to replace the dependent variable with the logarithm of output per worker and then specify its determinants on the right-hand-side, with no role for initial output:

$$\log y(t) = \pi' x + u$$

The fundamental drawback of this approach is that many determinants of the steady-state level of income will themselves be endogenous to the level of income. As a result, instrumental variable methods must be used, as in Acemoglu et al. (2001), Frankel and Romer (1999), and the paper that introduced this approach to explaining

productivity differences, Hall and Jones (1999). Note that regression (4) avoids the problem to some extent, by conditioning on the initial level of output per worker.

This discussion suggests that the partial adjustment model (4) is an attractive framework for growth econometrics, and this is reflected in its popularity. Admittedly, if the regression is used as a reduced-form, there are clear losses in not providing a structural model of the steady-state growth path. For example, we can no longer interpret the regression coefficients as structural parameters, or relate the speed of conditional convergence to these same parameters - all useful aspects of the analysis in Mankiw, Romer and Weil (1992).

But there are also some advantages to moving beyond a specific theoretical structure. At least two reasons stand out. First, it is not clear that economic theory offers a truly satisfactory structural model of the level of the growth path - and certainly does not, if we rule out the use of aggregate production functions. Second, the structural models typically adopted are often framed in terms of variables that are themselves endogenous. A classic example is the use of the saving/investment rate to explain growth in the model of Mankiw, Romer and Weil. Their paper suggests that the investment rate is an important determinant of the level of the growth path, but leaves the cross-country variation of investment unexplained. In attempting to explain differences in growth outcomes, a good case could be made for omitting the investment rate from the regression, so that the researcher detects 'indirect' effects on growth that work through a proximate cause, higher investment. And then we are clearly back at an ad hoc partial adjustment model of the form (4).

The alternative to the reduced form is to introduce additional structural relationships: for example, the investment rate might be explicitly related to the

relative price of capital or the extent of distortions. In a very interesting paper, De Long (1996) investigates a model of this kind, and this is an approach that arguably deserves more research attention. In the meantime, the point remains that growth econometrics typically relies on modelling partial adjustment to a growth path, the level of which can be specified either in structural terms or as a more ad hoc reduced-form. The structural approach will almost always rely on aggregate production functions, but when such relationships are believed to be absent, it is possible to appeal instead to the reduced-form interpretation.

5.4 Indirect inference on technology

The preceding argument is that much can be learnt about growth in the absence of aggregate production functions. Nevertheless, many researchers have judged it important to think about differences in TFP (in levels and growth rates) across countries. At first glance, the measurement of these differences seems to rely on specifying very simple production technologies at the aggregate level.

An interesting question is the extent to which technology differences are better inferred by indirect means. In the context of microeconomic data, there is a possible solution to the problem of estimating production relationships when technical efficiency is unobserved. This is to use a structural model of the decision-maker's optimisation problem, together with observed input choices, to infer changes in technology, as in Olley and Pakes (1996) and Levinsohn and Petrin (2003). For analysing firm-level data this seems a highly promising approach, when technologies

are observed by firms but not by the econometrician.[9] For cross-country data the same trick is much less likely to be effective, partly because the fiction of a unitary decision maker is less appropriate, and partly because there are relatively few input choices at the aggregate level that could be used to infer the behaviour of TFP.[10]

Are there other, similarly indirect approaches? Some of our best information about cross-country technology differences is likely to come from the analysis of trade flows. Given specialization according to comparative advantage, a structural model of trade can be used to reveal the technology differences that are needed to explain observed flows, as in Trefler (1995). Related ideas can be used to recover information about returns to scale, as in the estimates of Antweiler and Trefler (2002). It could be argued that these specific implications of trade data deserve more attention in the growth literature than they have received to date, a point made by Klenow and Rodriguez-Clare (1997). One advantage of these approaches is that, at least in principle, technology differences might be quantified for individual sectors, without relying on the artificial construction of an aggregate production function.[11]

Hendricks (2002) makes further progress on productivity differences, by a more careful treatment of human capital that takes an indirect route. Making use of data on immigrant workers in the US labour market, he can infer unmeasured aspects of their human capital, reflecting cross-country differences in the quality of

[9] Another approach is to estimate production relationships using latent variable methods to model the unobserved technology indices. See Krusell et al. (2000) for a macroeconomic application of this idea.

[10] It is possible that, under the intertemporal approach to the current account, movements in investment and the current account could be used to infer TFP changes. See Obstfeld and Rogoff (1996), especially chapter 2, for relevant discussion. The analysis of Kalemli-Ozcan et al. (2005) links net capital flows between US states to TFP at the state level.

[11] More generally, where sector-level data are available, it is possible to take a much richer view of technology differences than is usual in the cross-country literature. For example, see Bernard and Jones (1996).

schooling, for example. This allows a more accurate assessment of the role of capital and human capital in explaining productivity differences, relative to unobserved technology.

One source of information that has been relatively unexploited by growth economists is the cross-country data on relative prices. The PPP-adjusted output measures in the Penn World Table are built on detailed price data, and under relatively simple assumptions, it is possible to use these data to infer properties of underlying production relationships. Hsieh and Klenow (2003) and Herrendorf and Valentinyi (2005) are examples of this approach. Even without such detailed data, calibrated structural models can be used to infer the extent of productivity differences, as in Graham and Temple (2001). All these papers rely on a combination of sectoral disaggregation and a set of equilibrium conditions, rather than an aggregate production function.

Overall it seems clear that, with sufficient ingenuity, it is possible to learn a great deal about why income levels differ across countries, without requiring the existence of an aggregate production function. Although these indirect methods are intrinsically harder to implement than growth regressions, they are also a great deal more informative.[12] They surely represent one of the best ways forward for growth economics.

6. Conclusions

[12] In this respect, it is worth noting that many summaries of the empirical growth literature, including Temple (1999), have not been sufficiently eclectic in attempting to derive stylised facts about technology differences. The use of a wide variety of evidence and arguments in Easterly and Levine (2001) is a rare exception.

This paper has reviewed various issues raised by aggregation in the context of growth economics. The problems associated with aggregation are more serious than is often realized; they do not relate simply to capital, but also to labour and output. And a succinct summary of rigorous work on aggregation, especially that collected in Fisher (1993), would be that aggregate production functions do not exist.

For the reasons explained above, I think this result is more problematic for empirical research than theoretical work. The paper therefore concentrates on how empirical researchers might acknowledge the difficulties of aggregation, and move towards richer models. Empirical models could be devised to accommodate different types of capital and labour. Work in the dual economy tradition may be an especially attractive way to move beyond the artificial simplicities of one-sector models, without sacrificing either understanding or tractable empirical frameworks. It is also worth emphasizing that, in practice, much of growth econometrics can be justified and interpreted without any reference to an aggregate production function. Perhaps the key point, for both theorists and applied researchers, is to recognise at the outset how difficult it will be to approximate the underlying structure.

I will end this brief discussion with a claim that may be rather more controversial. The critique of aggregate production functions is often, though not always, part of a much wider critique of the assumptions and methods of modern macroeconomics. By emphasizing the dangers of aggregation, it is possible that some critics of mainstream macroeconomics (or, more broadly, neoclassical economics) are not addressing their criticisms at the mainstream's weakest point. One of the central conceptual weaknesses in general equilibrium models has always been the

requirement that all trades take place without any frictions, only at equilibrium prices, at one instant in time - the useful but abstract device of the Walrasian auctioneer. Like many commentators before me, I suspect that macroeconomics without an auctioneer might look very different to the world of Arrow-Debreu.

This relates to a much broader attack on mainstream theory that formed one side of the Cambridge controversies. Cohen and Harcourt (2003) emphasize the breadth of the areas of dispute. One point, mentioned above, is that neoclassical general equilibrium theory cannot be used to justify a simple inverse relationship between a "capital stock" and the marginal product of capital, a point obscured by aggregation. But there are also more wide-ranging elements to this critique. Capital theory relates to processes observed over time. Once we allow trades to take place over time and at disequilibrium prices, the equilibrium is unlikely to be independent of the transition towards it. As soon as the auctioneer is made aware that he is a fictional character, and asked to leave, we have path dependence.

In the long term, these criticisms could be more important than disputes over the assumptions needed to carry out aggregation. In this respect, it is interesting to note that much recent work in macroeconomics emphasizes the absence of an auctioneer in the labour market, so that matching between workers and firms is an ongoing and imperfect process. Without an auctioneer, and given the existence of non-pecuniary externalities in the process of job creation and matching, there is no longer a presumption that the decentralized market equilibrium is Pareto efficient. Depending on the precise class of models, the market equilibrium may be inefficient for most parameter values. Variations on these models can look similar to 'heterodox' models in some respects. For example, Hall (2005) explains the dynamics

of unemployment in terms of social norms in wage setting. And all this can be done without any heterogeneity in technology.

This leads me to a perhaps controversial conclusion. For the reasons discussed above, I think any researcher who writes down or estimates an aggregate production function needs to be aware of how dangerous this can be, and aware of some of the alternatives. At the same time, if orthodox general equilibrium macroeconomics has an Achilles heel, it is perhaps more likely to be the auctioneer than the strict conditions needed for aggregation.

References

Acemoglu, D., Johnson, S. and Robinson, J. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review*, 91(5): 1369-1401.

Antweiler, W. and Trefler, D. (2002). Increasing Returns and All That: A View from Trade. *American Economic Review*, 92(1), 93-119.

Benhabib, J. and Jovanovic, B. (1991). Externalities and growth accounting. *American Economic Review*, 81(1), 82-113.

Benhabib, J. and Spiegel, M. M. (1994). The role of human capital in economic development: evidence from aggregate cross-country data. *Journal of Monetary Economics*, 34, 143-173.

Bernard, A. and Jones, C. I. (1996). Comparing apples to oranges: productivity convergence and measurement across industries and countries. *American Economic Review*, 86, 1216-1238.

Blackorby, C. and Schworm, W. (1988). The Existence of Input and Output Aggregates in Aggregate Production Functions. *Econometrica*, 56(3): 613-43.

Cohen, A. J. and Harcourt, G. C. (2003). Whatever happened to the Cambridge Capital Theory Controversies? *Journal of Economic Perspectives*, 17(1), 199-214.

Cohen, D. and Soto, M. (2002). Why are Poor Countries Poor? A Message of Hope which Involves the Resolution of a Becker/Lucas Paradox. CEPR Discussion Paper no. 3528.

Cornwall, J. (1977). The relevance of dual models for analyzing developed capitalist economies. *Kyklos*, 30(1), 51-73.

De Long, J. B. (1996). Cross-Country Variations in National Economic Growth Rates: The Role of "Technology", in Fuhrer, J. C. and Little, J. S. (eds.) *Technology and growth: Conference proceedings*, June 1996. Boston: Federal Reserve Bank of Boston.

Dowrick, S. and Gemmell, N. (1991). Industrialisation, catching up and economic growth: a comparative study across the world's capitalist economies. *Economic Journal*, 101(405), 263-275.

Easterly, W. and Levine, R. (2001). It's Not Factor Accumulation: Stylized Facts and Growth Models. *World Bank Economic Review*, 2001, 15(2): 177-219.

Felipe, J. and Fisher, F. M. (2003). Aggregation in Production Functions: What Applied Economists Should Know. *Metroeconomica*, 54(2-3): 208-62.

Felipe, J. and Holz, C. A. (2001). Why do aggregate production functions work? Fisher's simulations, Shaikh's identity and some new results. *International Review of Applied Economics*, 15(3), 261-285.

Felipe, J. and McCombie, J. S. L. (2005). Why are some countries richer than others? A skeptical view of Mankiw-Romer-Weil's test of the neoclassical growth model. *Metroeconomica*, 56(3), 360-392.

Fisher, F. M. (1993). *Aggregation: Aggregate production functions and related topics*. Cambridge, MA: MIT Press.

Frankel, J. A. and Romer, D. (1999). Does trade cause growth? *American Economic Review*, 89, 379-98.

Graham, B. S. and Temple, J. R. W. (2001). Rich nations, poor nations: how much can multiple equilibria explain? CEPR discussion paper no. 3046.

Greenwood, J., Hercowitz, Z. and Krusell, P. (1997). Long-Run Implications of Investment-Specific Technological Change. *American Economic Review*, 87(3): 342-62.

Griliches, Z. and Mairesse, J. (1998). Production functions: the search for identification. In *Econometrics and Economic Theory in the twentieth century: the Ragnar Frisch centennial symposium*, pp. 169-203. Cambridge: Cambridge University Press.

Hall, R. E. (2005). Employment Fluctuations with Equilibrium Wage Stickiness. *American Economic Review*, 95(1): 50-65.

Hall, R. E. and Jones, C. I. (1999). Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics*, 114(1), 83-116.

Hendricks, L. (2002). How important is human capital for development? Evidence from immigrant earnings. *American Economic Review*, 92(1), 198-219.

Herrendorf, B. and Valentinyi, A. (2005). What sectors make the poor countries so unproductive? CEPR discussion paper no. 5399.

Hsieh, C.-T. and Klenow, P. J. (2003). Relative Prices and Relative Prosperity. NBER Working Paper no. 9701.

Islam, N. (1995). Growth empirics: a panel data approach. *Quarterly Journal of Economics*, 110(4), 1127-1170.

Jorgenson, D. W. (1995). *Productivity. Volume 1. Postwar U.S. economic growth*. Cambridge and London: MIT Press.

Jorgenson, D. W. and Griliches, Z. (1967). The explanation of productivity change. *Review of Economic Studies*, 34(3), 249-283.

Jorgenson, D. W. and Stiroh, K. J. (2000). U. S. economic growth at the industry level. *American Economic Review*, 90(2), 161-167.

Kalemli-Ozcan, S., Reshef, A., Sørensen, B. E. and Yosha, O. (2005). Net capital flows and productivity: evidence from US states. Manuscript, University of Houston.

Kindleberger, C. (1967). *Europe's postwar growth: the role of labor supply*. Harvard University Press, Cambridge, MA.

Klenow, P. J. and Rodriguez-Clare, Andres (1997). Economic Growth: A Review Essay. *Journal of Monetary Economics*, 40(3): 597-617.

Krusell, P., Ohanian, L. E., Rios-Rull, J.-V. and Violante, G. L. (2000). Capital-skill complementarity and inequality: a macroeconomic analysis. *Econometrica*, 68(5), 1029-1053.

Landon-Lane, J. and Robertson, P. E. (2003) Accumulation and productivity growth in industrializing economies. Manuscript, UNSW, Australia.

Levinsohn, J. and Petrin, A. (2003). Estimating Production Functions Using Inputs to Control for Unobservables. *Review of Economic Studies*, 70(2), 317-41.

Lewis, W. A. (1954). Economic Development with Unlimited Supplies of Labour. *The Manchester School*, 22(2), 139-191.

Lucas, Robert E., Jr (1990). Why Doesn't Capital Flow from Rich to Poor Countries? *American Economic Review*, 80(2): 92-96.

Mankiw, N. G., Romer, D. and Weil, D. (1992). A contribution to the empirics of economic growth. *Quarterly Journal of Economics*, 107(2), 407-437.

Marschak, J. and Andrews, W. (1944). Random simultaneous equations and the theory of production. *Econometrica*, 12(3-4), 143-205.

Melitz, M. J. (2003). The Impact of Trade on Intra-industry Reallocations and Aggregate Industry Productivity. *Econometrica*, 71(6): 1695-1725.

Obstfeld, M. and Rogoff, K. (1996). *Foundations of international macroeconomics*. Cambridge, Mass. and London: MIT Press.

Olley, G. S. and Pakes, A. (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica*, 64(6): 1263-97.

Robertson, P. E. (1999). Economic growth and the return to capital in developing economies. *Oxford Economic Papers*, 57(4), 577-594.

Ros, J. (2000). *Development Theory and the Economics of Growth.* Ann Arbor, University of Michigan Press.

Samuelson, P. (1961-62). Parable and realism in capital theory: the surrogate production function. *Review of Economic Studies*, 29, 193-206.

Shaikh, A. (1974). Laws of Production and Laws of Algebra: The Humbug Production Function. *Review of Economics and Statistics*, 56(1): 115-20.

Solow, R. M. (1966). Review of Capital and growth. *American Economic Review*, 56(5), 1257-1260.

Temple, J. R. W. (1998). Equipment Investment and the Solow Model. *Oxford Economic Papers*, 50(1): 39-62.

Temple, J. R. W. (1999). The new growth evidence. *Journal of Economic Literature*, 37(1), 112-156.

Temple, J. R. W. (2001a). Generalizations that aren't? Evidence on education and growth. *European Economic Review*, 45(4-6), 905-918.

Temple, J. R. W. (2001b). Structural change and Europe's Golden Age. University of Bristol discussion paper no. 01/519.

Temple, J. R. W. (2003). The long-run implications of growth theories. *Journal of Economic Surveys*, 17(3), 497-510.

Temple, J. R. W. (2005). Dual economy models: a primer for growth economists. *The Manchester School*, 73(4), 435-478.

Temple, J. R. W. and Woessmann, L. (2004). Dualism and cross-country growth regressions. University of Bristol discussion paper no. 04/560.

Tesfatsion, L. and Judd, K. L. (2006). *Handbook of Computational Economics, Volume 2*. Amsterdam: North-Holland, forthcoming.

Trefler, D. (1995). The Case of the Missing Trade and Other Mysteries. *American Economic Review*, 85(5): 1029-46.

Ventura, J. (1997). Growth and Interdependence. *Quarterly Journal of Economics*, 112(1): 57-84.