Volume Title: New Developments in Productivity Analysis

Volume Author/Editor: Charles R. Hulten, Edwin R. Dean and Michael J. Harper, editors

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-36062-8

Volume URL: http://www.nber.org/books/hult01-1

Publication Date: January 2001

Chapter Title: Why Is Productivity Procyclical? Why Do We Care?

Chapter Author: Susanto Basu, John Fernald

Chapter URL: http://www.nber.org/chapters/c10128

Chapter pages in book: (p. 225 - 302)

7

Why Is Productivity Procyclical? Why Do We Care?

Susanto Basu and John Fernald

Productivity is procyclical. That is, whether measured as labor productivity or total factor productivity, productivity rises in booms and falls in recessions. Recent macroeconomic literature views this stylized fact of procyclical productivity as an essential feature of business cycles, largely because of the realization that each explanation for this stylized fact has important implications for the workings of macroeconomic models. In this paper, we seek to identify the empirical importance of the four main explanations for this stylized fact, and discuss the implications of our results for the appropriateness of different macroeconomic models.

Until recently, economists generally regarded the long-run *average* rate of productivity growth as important for growth and welfare; procyclical productivity, by contrast, seemed irrelevant for understanding business cycles. Economists presumed that high-frequency fluctuations in productivity reflected cyclical mismeasurement—for example, labor and capital worked harder and longer in booms—but these cyclical variations in utilization were not themselves important for understanding cycles.

In the past decade and a half, productivity fluctuations have taken center stage in modeling output fluctuations, and are now viewed as an essen-

Susanto Basu is associate professor of economics at the University of Michigan and a research associate of the National Bureau of Economic Research. John Fernald is senior economist at the Federal Reserve Bank of Chicago.

We are grateful for comments from Zvi Griliches, John Haltiwanger, Michael Horvath, Charles Hulten, Dale Jorgenson, Catherine Morrison, Plutarchos Sakellaris, and participants in the conference. Basu thanks the National Science Foundation and the Alfred P. Sloan Foundation for financial support. The views in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System.

tial part of the cycle. Figure 7.2 (discussed later) charts the Solow residual for the aggregate U.S. economy. The figure also shows growth in output and growth in "inputs," defined as a weighted average of labor and capital growth (we discuss data sources and definitions in section 7.4). Mean productivity growth is positive, so that over time society's ability to produce goods or services that satisfy final demand is rising faster than its inputs. In addition, productivity growth is quite volatile. The volatility is not random, but is significantly procyclical: The correlation with output growth is about 0.8.

Macroeconomists have become interested in the cyclical behavior of productivity because of the realization that procyclicality is closely related to the impulses or propagation mechanisms underlying business cycles. Even the cyclical mismeasurement that was formerly dismissed as unimportant turns out to be a potentially important propagation mechanism.

There are four main explanations for high-frequency fluctuations in productivity. First, procyclical productivity may reflect procyclical technology. After all, under standard conditions, total factor productivity measures technology. If there are high-frequency fluctuations in technology, it is not surprising that there are high-frequency fluctuations in output as well. Second, widespread imperfect competition and increasing returns may lead productivity to rise whenever inputs rise. (Increasing returns could be internal to a firm, or could reflect externalities from the activity of other firms.) Figure 7.2 (discussed later) shows the key stylized fact of business cycles, the comovement of inputs and output. With increasing returns, the fluctuations in inputs then cause endogenous, procyclical fluctuations in productivity. Third, as already mentioned, utilization of inputs may vary over the cycle. Fourth, reallocation of resources across uses with different marginal products may contribute to procyclicality. For example, if different industries have different degrees of market power, then inputs will generally have different marginal products in different uses. Then aggregate productivity growth is cyclical if sectors with higher markups have input growth that is more cyclical. Alternatively, if inputs are relatively immobile or quasi-fixed, then marginal products may temporarily differ across uses; as these resources eventually shift, productivity rises.[1]

Why do economists now care about the relative importance of these four explanations? In large part, changes in the methodology of macroeconomics have raised the fact of procyclical productivity to the forefront. Macroeconomists of all persuasions now use dynamic general equilibrium (DGE) models, and it turns out that each of the four explanations has important implications for the workings of these models. Hence, answering

1. For examples of these four explanations, see, respectively, Cooley and Prescott (1995); Hall (1988, 1990); Basu (1996), Bils and Cho (1994), and Gordon (1993); and Basu and Fernald (1997a, 1997b).

why productivity is procyclical sheds light on the relative merits of different models of the business cycle.

First, if high-frequency fluctuations in productivity reflect high-frequency fluctuations in technology, then comovement of output and input (i.e., business cycles) are a natural byproduct. The DGE approach to business cycle modeling began with so-called real business cycle models, which explore the extent to which the frictionless one-sector Ramsey-Cass-Koopmans growth model can explain business cycle correlations. Real business cycle models use Solow's productivity residual—interpreted as aggregate technology shocks—as the dominant impulse driving the cycle (e.g., Cooley and Prescott 1995). Other impulses may affect output in these models, but technology shocks must dominate if the model is to match the key stylized fact of business cycles: the positive comovement of output and labor input.[2]

Second, recent papers show that increasing returns and imperfect competition can modify and magnify the effects of various shocks in an otherwise standard DGE model. In response to government demand shocks, for example, models with countercyclical markups can explain a rise in real wages while models with increasing returns can explain a rise in measured productivity. Perhaps most strikingly, if increasing returns are large enough, they can lead to multiple equilibria, in which sunspots or purely nominal shocks drive business cycles.[3] Furthermore, if firms are not all perfectly competitive, then it is not appropriate to use the Solow residual as a measure of technology shocks, since the Solow residual becomes endogenous. Taking the Solow residual to be exogenous thereby mixes impulses and propagation mechanisms.

Third, variable utilization of resources turns out to improve the propagation of shocks in DGE models. If firms can vary the intensity of factor use, then the effective supply of capital and labor becomes more elastic. Small shocks (to technology or demand) can then lead to large fluctuations.[4] If the model has sticky nominal prices, these elastic factor supplies greatly increase the persistence of the real effects of nominal shocks.

Fourth, reallocations of inputs, without any change in technology, may cause aggregate productivity to be procyclical. For example, in the sectoral

2. Barro and King (1984) provide an early discussion of this issue. Dynamic general equilibrium models without technology shocks can match this stylized fact with countercyclical markups of price over marginal cost, arising from sticky prices (as in Kimball, 1995) or from game-theoretic interactions between firms (as in Rotemberg and Woodford 1992). Models with an extreme form of increasing returns—increasing marginal product of labor—can also produce a positive comovement between output and labor input; see, for example, Farmer and Guo (1994).

3. See, for example, Rotemberg and Woodford (1992), Farmer and Guo (1994), and Beaudry and Devereux (1994). Rotemberg and Woodford (1995) survey dynamic general equilibrium models with imperfect competition.

4. See, for example, Burnside and Eichenbaum (1996); Dotsey, King, and Wolman (1997); and Wen (1998).

shifts literature (e.g., Ramey and Shapiro 1998 and Phelan and Trejos 1996), demand shocks cause differences in the marginal product of immobile factors across firms. Output fluctuations then reflect shifts of resources among uses with different marginal products. Basu and Fernald (1997a) provide a simple stylized example in which aggregation over constant-returns firms with different levels of productivity lead to the existence of multiple equilibria. Weder (1997) calibrates a DGE model where durables manufacturing firms have increasing returns while all other producers have constant returns (calibrated from results in Basu and Fernald 1997a), and shows that reallocations make multiple equilibria possible in that model. Of course, reallocations can also help propagate sector-specific technology shocks, as in Lilien (1982).

In this paper, we seek to identify the importance of the four explanations. Our approach builds on the seminal contributions of Solow (1957) and Hall (1990). We allow for imperfect competition and nonconstant returns to scale, as well as variations in the workweek of capital or the effort of labor that are unobservable to the econometrician. In the Solow-Hall tradition, we take the production function residual as a measure of sectoral technology shocks. We then aggregate over sectors, since our ultimate focus is on explaining movements in aggregate productivity.

Our empirical work relies primarily on the tools developed by Basu and Fernald (1997b) and Basu and Kimball (1997). Both papers allow for increasing returns to scale and markups of price over marginal cost. Basu and Fernald stress the role of sectoral heterogeneity. They argue that for economically plausible reasons—for example, differences across industries in the degree of market power—the marginal product of an input may differ across uses. Then aggregate productivity growth depends in part on which sectors change inputs.

Basu and Kimball stress the role of variable capital and labor utilization. Solow's original (1957) article presumed that variations in capacity utilization were a major reason for the procyclicality of measured productivity, a presumption widely held thereafter. (See, for example, Gordon 1993; Abbott, Griliches, and Hausman 1998). In essence, the problem is cyclical measurement error of input quantities: True inputs are more cyclical than measured inputs, so that measured productivity is spuriously procyclical.

Basu and Kimball use the basic insight that a cost-minimizing firm operates on all margins simultaneously—whether observed or unobserved—ensuring that the marginal benefit of any input equals its marginal cost. As a result, increases in observed inputs can proxy for unobserved changes in utilization. For example, when labor is particularly valuable, firms will work existing employees both longer (increasing observed hours per worker) and harder (increasing unobserved effort).

Our work on these issues follows the Solow-Hall tradition, which makes minimal assumptions and focuses on the properties of the resulting resid-

ual. An alternative literature, surveyed by Nadiri and Prucha (ch. 4, this volume), addresses issues of technology change by attempting to estimate an extensively parameterized production (or cost) function. That approach can address a wider range of issues because it imposes much more structure on the problem. If the problem is correctly specified and all necessary data are available, the more parametric approach offers clear theoretical advantages, since one can then estimate second-order properties of the production function such as elasticities of substitution. However, that approach is likely to suffer considerable practical disadvantages, such as the increased likelihood of misspecification and the necessity of factor prices being allocative period by period. We are skeptical that observed factor prices are always allocative—with implicit contracts, for example, observed wages need only be right on average, rather than needing to be allocative period by period. Hence, we argue in favor of an explicitly first-order approach when possible, since results are likely to be more robust.

Note that this production function literature sometimes claims to solve the capacity utilization problem (e.g., Morrison 1992a,b). Suppose, for example, that the capital stock cannot be changed instantaneously but that it is nevertheless used all the time. It can still be used more intensively by combining the fixed capital services with more labor and materials. As a result, the shadow value of this quasi-fixed capital stock may vary from its long-run equilibrium factor cost. Capital's output elasticity may also vary over time, reflecting variations in this shadow value. More generally, with quasi-fixity, quantities of inputs (and output) may differ from long-run equilibrium values; similarly, on the dual (cost) side, short-run variable cost may differ from long run variable cost at a given output level. If we equate full capacity with the firm's optimal long-run equilibrium point, then capacity utilization (defined by how far actual output or variable cost is from their appropriately-defined long-run equilibrium values) may vary over time.

For estimating the technology residual, the relevant feature of this literature is that it attempts to estimate the shadow value of quasi-fixed inputs in order to calculate the time variation in the output elasticities. The production function literature thus tries to control for time-varying output elasticities, while the capacity-utilization literature tries to measure factor quantities (i.e., the service flow) correctly.[5]

The enormous confusion between these two logically distinct ideas stems purely from the semantic confusion caused by both literatures claiming to address the capacity utilization problem. In particular, adherents of the time-varying elasticity approach often claim that their opti-

5. For our purposes, this is the only difference that matters. The structural literature also estimates a variety of other quantities, such as elasticities of substitution, that may be of independent interest.

mization-based methods obviate the need to use proxies that control for unobserved variations in capital workweek or labor effort, with the implication that the use of proxies is somehow ad hoc. This implication is wrong. The production function literature assumes that the quantities of all inputs are correctly measured; the capacity utilization literature devotes itself to correcting for measurement error in the inputs. The two address separate issues.

The easiest way to see the difference between the two concepts is to suppose that output is produced via a Cobb-Douglas production function. Then the output elasticities are constant, so for our purposes—for example, estimating returns to scale—the time-varying elasticity literature has nothing to add. But there is still a capacity utilization problem—for example, if workers work harder in a boom than in recessions, but this fact is not captured in the statistics on labor input. Formally, therefore, the problem of time-varying elasticities is a second-order issue—it relates to deviations from the first-order Cobb-Douglas approximation—but capacity utilization is a first-order issue, since it concerns the accurate measurement of the inputs.

After presenting empirical results, we discuss implications for macroeconomics. We discussed some of these implications above, in motivating our interest in procyclical productivity, so we highlight other issues. Normative productivity analysis emphasizes the welfare interpretation of the productivity residual. But what if productivity and technology differ because of distortions such as imperfect competition? Recent macroeconomic literature often seems to assume that we measure productivity in order to measure technology; any differences reflect mismeasurement. Although variable input utilization is clearly a form of mismeasurement, we argue that other distortions (such as imperfect competition) are not: Productivity has clear welfare implications, even in a world with distortions. A modified Solow residual—which reduces to Solow's measure if there are no economic profits—approximates to first order the welfare change of a representative consumer. Intuitively, growth in aggregate output measures the growth in society's ability to consume. To measure welfare change, we must then subtract the opportunity cost of the inputs used to produce this output growth. Input prices measure that cost, even when they do not measure marginal products. Hence, if productivity and technology differ, then productivity most closely indexes welfare.

Section 7.1 provides a relatively informal overview of the issues. We highlight key issues facing the empirical literature and make recommendations. That section will allow the reader to proceed to the data and empirical sections 7.4 and 7.5. However, some of our choices in section 7.1 require more formal treatment to justify fully; section 7.2 provides this treatment. Section 7.3 discusses aggregation from the firm to the economy-wide level. Section 7.4 discusses data, and section 7.5 presents results. Sec-

tion 7.6 discusses several macroeconomic implications of our results. Section 7.7 concludes.

7.1 Methods of Estimating Technical Change: Overview

Why do macroeconomists care about fluctuations in productivity? First, and perhaps foremost, productivity yields information about the aggregate production of goods and services in the economy. Second, productivity analysis may provide information about firm behavior—for example, the markup and its cyclicality, the prevalence of increasing returns to scale, and the factors determining the level of utilization.

There are several possible approaches to empirical productivity analysis. Each has advantages and disadvantages. In this section, we assess these alternatives, and make recommendations about key decisions facing an empirical researcher. We summarize the microeconomic foundations of our preferred approach, which is in the spirit of Solow (1957) and Hall (1990). This discussion may satisfy the interests and needs of most readers, who can then proceed to the data and results. However, a full justification of our approach requires a somewhat more technical discussion. We present that technical discussion in sections 7.2 and 7.3.

Our ultimate goal is to understand the aggregate economy. At an aggregate level, the appropriate measure of output is national expenditure on goods and services, that is, GDP—the sum of consumption, investment, government purchases, and net exports. GDP measures the quantity of goods available to consume today or invest for tomorrow. GDP is intrinsically a value-added measure, since the national accounts identity assures us that when we aggregate firm-level value added (defined in nominal terms as total revenue minus the cost of intermediate inputs of materials and energy; in real terms, gross output with intermediate inputs netted out in some way, as discussed in section 7.3), aggregate value added equals national expenditure. The economy's resources for producing goods and services are capital and labor.

Nevertheless, despite our interest in macroeconomic aggregates, we should begin at the level where goods are actually produced: at the level of a firm or even a plant. The appropriate measure of output for a firm is gross output—shoes, books, computers, haircuts, and so forth. Firms combine inputs of capital, labor, and intermediate goods (materials and energy) to produce this gross output.

Our goal, then, is to explore how firm-level production of gross output translates into production of aggregate final expenditure. Macroeconomists often assume an aggregate production function that relates aggregate final expenditure (value added) to inputs of capital and labor. For many purposes, what matters is whether such a function provides at least a first-order approximation to the economy's production possibilities—even if an

explicit aggregate function does not exist (as it rarely does). For example, in calibrating a dynamic general equilibrium model, one may care about how much aggregate output increases if the equilibrium quantity of labor increases, and a first-order approximation should give the right magnitude.[6] We argue that in a world without frictions or imperfections, aggregation is generally very clean and straightforward. However, with frictions and imperfections such as imperfect competition or costs of reallocating inputs, the assumption that an aggregate production function exists (even as a first-order approximation) generally fails—but the failures are economically interesting. These failures also help explain procyclical productivity.

7.1.1 The Basic Setup

We write the firm's production function for gross output, Y_i, in the following general form:

$$Y_i = F^i(\tilde{K}_i, \tilde{L}_i, M_i, T_i) \tag{1}$$

Firms use capital services $\tilde{K}_i$, labor services $\tilde{L}_i$, and intermediate inputs of materials and energy M_i. We write capital and labor services with tildes to remind ourselves that these are the true inputs of services, which may not be observed by the econometrician. T_i indexes technology, which we define to include any inputs that affect firm-level production but are not compensated by the firm. For example, T_i comprises both standard exogenous technological progress and any Marshallian externalities that may exist; we take technology as unobservable. (For simplicity, we omit time subscripts.)

The services of labor and capital depend on both raw quantities (hours worked and the capital stock), and the intensity with which they are used. We define labor services as the product of the number of employees, N_i, hours worked per employee H_i, and the effort of each worker, E_i. We define capital services as the product of the capital stock, K_i, and the utilization of the capital stock, Z_i. (For example, K_i might represent a particular machine, whereas Z_i represents the machine's workweek—how many hours it is operated each period). Hence, input services are:

$$\begin{aligned} \tilde{L}_i &= E_i H_i N_i, \\ \tilde{K}_i &= Z_i K_i. \end{aligned} \tag{2}$$

We will generally assume that the capital stock and the number of employees are quasi-fixed, so firms cannot change their levels costlessly.

6. However, DGE models usually also need to relate changes in factor prices (e.g., wages) to changes in the quantities of inputs. For this purpose, the first-order approximation is not sufficient: One also needs to know the elasticities of substitution between inputs in production, which is a second-order property.

Let the firm's production function F^i be (locally) homogeneous of arbitrary degree γ_i in total inputs. Constant returns then corresponds to the case where γ_i equals 1. Formally, we can write returns to scale in two useful, and equivalent, forms. First, returns to scale equal the sum of output elasticities:

$$\gamma_i = \frac{F^i_1 \tilde{K}_i}{Y_i} + \frac{F^i_2 \tilde{L}_i}{Y_i} + \frac{F^i_3 M_i}{Y_i}, \tag{3}$$

where F^i_J denotes the derivative of the production function with respect to the Jth element (i.e., the marginal product of input J). Second, assuming firms minimize cost, we can denote the firm's cost function by $C_i(Y_i)$. (In general, the cost function also depends on the prices of the variable inputs and the quantities of any quasi-fixed inputs, although we omit those terms for simplicity here.) The *local* degree of returns to scale equals the inverse of the elasticity of cost with respect to output.[7]

$$\gamma_i(Y_i) = \frac{C_i(Y_i)}{Y_i C_i'(Y_i)} = \frac{C_i(Y_i)/Y_i}{C_i'(Y_i)} = \frac{AC_i}{MC_i}, \tag{4}$$

where AC_i equals average cost, and MC_i equals marginal cost. Increasing returns, in particular, may reflect overhead costs or decreasing marginal cost; both imply that average cost exceeds marginal cost. If increasing returns take the form of overhead costs, then $\gamma_i(Y_i)$ is not a constant structural parameter, but depends on the level of output the firm produces. As production increases, returns to scale fall as the firm moves down its average cost curve.

As equation (4) shows, there is no necessary relationship between the degree of returns to scale and the slope of the marginal cost curve. Indeed, increasing returns are compatible with increasing marginal costs, as in the standard Chamberlinian model of imperfect competition. One can calibrate the slope of the marginal cost curve from the degree of returns to scale only by assuming there are no fixed costs. An important point is that the slope of the marginal cost curve determines the slopes of the factor demand functions, which in turn are critical for determining the results of DGE models (for example, whether the model allows sunspot fluctuations). Several studies have used estimates of the degree of returns to scale to calibrate the slope of marginal cost; for a discussion of this practice, see Schmitt-Grohé (1997).

Firms may also charge a price P_i that is a markup, μ_i, over marginal cost. That is, $\mu_i \equiv P_i/\mathrm{MC}_i$. Returns to scale γ_i is a technical property of the production function, whereas the markup μ_i is essentially a behavioral parameter, depending on the firm's pricing decision. However, from equation (4), the two are inextricably linked:

7. See Varian (1984, 68) for a proof.

$$\gamma_i = \frac{C_i(Y_i)}{Y_i C_i'(Y_i)} = \frac{P_i}{C_i'(Y_i)} \frac{C_i(Y_i)}{P_i Y_i} = \mu_i(1 - s_{\pi i}), \tag{5}$$

where $s_{\pi i}$ is the share of pure economic profit in gross revenue. As long as pure economic profits are small (and in our data, we estimate the average profit rate to be at most 3 percent[8]), equation (5) shows that μ_i approximately equals γ_i. Large markups, for example, require large increasing returns. This is just what one would expect if free entry drives profits to zero in equilibrium—for example, in Chamberlinian monopolistic competition. Thus, although increasing returns and markups are not equivalent from a welfare perspective, they are forced to equal one another if competition eliminates profits. As a result, as we show in the next subsection, one can write the resulting wedge between output elasticities and factor shares in terms of either parameter. Given low estimated profits, equation (5) shows that strongly diminishing returns (γ_i much less than one) imply that firms consistently price output below marginal cost (μ_i less than one). Since price consistently below marginal cost makes no economic sense, we conclude that average firm-level returns to scale must either be constant or increasing. Increasing returns also require that firms charge a markup, as long as firms do not make losses.

7.1.2 The Solow-Hall Approach

Suppose we want to estimate how rapidly technology is changing. Solow's (1957) seminal contribution involves differentiating the production function and using the firm's first-order conditions for cost minimization. If there are constant returns to scale and perfect competition, then the first-order conditions (discussed below) imply that output elasticities are observed in the data by revenue shares. Hall (1988, 1990) extends Solow's contribution to the case of increasing returns and imperfect competition. Under these conditions, output elasticities are not observed, since neither returns to scale nor markups are observed. However, Hall derives a simple regression equation, which he then estimates. In this section, we extend Hall's approach by using gross-output data and taking account of variable factor utilization.

We begin by taking logs of both sides of equation (1) and then differentiating with respect to time:

$$dy_i = \frac{F^i_1 K_i}{Y_i} d\tilde{k}_i + \frac{F^i_2 L_i}{Y_i} d\tilde{l}_i + \frac{F^i_3 M_i}{Y_i} dm + dt_i. \tag{6}$$

Small letters denote growth rates (so dy, for example, equals $(1/Y)\dot{Y}$), and we have normalized the output elasticity with respect to technology equal to one for simplicity.

8. Rotemberg and Woodford (1995) also provide a variety of evidence suggesting that profit rates are close to zero.

As Solow (1957) and Hall (1990) show, cost minimization puts additional structure on equation (6), allowing us to relate the unknown output elasticities to observed factor prices. In particular, suppose firms charge a price P_i that is a markup, μ_i, over marginal cost. (The advantage of the cost minimization framework is that it is unnecessary to specify the potentially very complicated, dynamic profit maximization problem that gives rise to this price.) Perfect competition implies μ_i equals one. Suppose that firms take the price of all J inputs, P_{Ji}, as given by competitive markets. The first-order conditions for cost-minimization then imply that:

$$P_i F^i_J = \mu_i P_{Ji}. \tag{7}$$

In other words, firms set the value of a factor's marginal product equal to a markup over the factor's input price. Equivalently, rearranging the equation by dividing through by μ_i, this condition says that firms equate each factor's marginal revenue product $(P_i/\mu_i)F^i_J$ to the factor's price.

The price of capital, P_{Ki}, must be defined as the *rental price* (or shadow rental price) of capital. In particular, if the firm makes pure economic profits, these are generally paid to capital, since the owners of the firm typically also own its capital. These profits should not be incorporated into the rental price. Equation (7) still holds if some factors are quasi-fixed, as long as we define the input price of the quasi-fixed factors as the appropriate *shadow* price, or implicit rental rate (Berndt and Fuss, 1986). We return to this point in section 7.2 where we specify a dynamic cost-minimization problem.

Using equation (7), we can write each output elasticity as the product of the markup multiplied by total expenditure on each input divided by total revenue. Thus, for example,

$$\frac{F^i_1 Z_i K_i}{Y_i} = \mu_i \frac{P_{Ki} K_i}{P_i Y_i} \equiv \mu_i s_{Ki}. \tag{8}$$

(The marginal product of capital is $F^i_1 Z_i$, since the services from a machine depend on the rate at which it is being utilized; i.e., its workweek.) The shares s_{Ji} are total *cost* of each type of input divided by total *revenue.* Thus, the input shares sum to less than one if firms make pure profits.

We now substitute these output elasticities into equation (6) and use the definition of input services from equation (2):

$$\begin{aligned} dy_i &= \mu_i[s_{Ki} d\tilde{k}_i + s_{Li} d\tilde{l}_i + s_{Mi} dm_i] + dt_i \\ &= \mu_i[s_{Ki}(dk_i + dz_i) + s_{Li}(dn_i + dh_i + de_i) + s_{Mi} dm_i] + dt_i \\ &= \mu_i[s_{Ki} dk_i + s_{Li}(dn_i + dh_i) + s_{Mi} dm_i] + \mu_i(1 - s_{Mi}) \\ &\quad \left[\frac{s_{Ki} dz_i + s_{Li} de_i}{(1 - s_{Mi})}\right] + dt_i \end{aligned}$$

By defining dx as a share-weighted average of conventional (observed) input growth, and du as a weighted average of unobserved variation in capital utilization and effort, we obtain our basic estimating equation:

$$dy_i = \mu_i dx_i + \mu_i(1 - s_{Mi})du_i + dt_i. \tag{9}$$

Using equation (5), we could rewrite equation (9) in terms of returns to scale γ_i. In that case, the correct weights to calculate weighted average inputs are cost shares, which sum to one, rather than revenue shares, which might not. Hall (1990) used the cost-share approach, but there is no economic difference between Hall's approach and ours, and the data requirements are the same. In particular, once we allow for the possibility of economic profits, we must in any case estimate a required rental cost of capital. Writing the equation in terms of μ_i turns out to simplify some later derivations and also facilitates the welfare discussion in Section VI.

Suppose all firms have constant returns to scale, all markets are perfectly competitive, and all factor inputs are freely variable and perfectly observed. Then the markup μ_i equals one, du_i is identically zero (or is observed), and all factor shares are observed as data (since there are no economic profits to worry about). This case corresponds to Solow's (1957) assumptions, and we observe everything in equation (9) except technology change dt_i, which we can calculate as a residual. However, if Solow's conditions fail, we can follow Hall (1990) and treat equation (9) as a regression.

An alternative to the Solow-Hall approach involves estimating many more properties of the production function (equation [1])—Nadiri and Prucha (this volume) survey that approach. That approach requires much more structure on the problem. For example, one must usually postulate a functional form and specify the firm's complete maximization problem, including all constraints. If the problem is correctly specified and all necessary data are available, that approach offers clear theoretical advantages for estimating the second-order properties of the production function. However, the structural approach may suffer considerable practical disadvantages, such as the increased likelihood of misspecification and the need for all factor prices to be allocative period by period. In any case, for the first-order issues we focus on here, it should give similar results. (If it does not, our view is that the Solow-Hall approach is probably more robust).

Regarding equation (9) as an estimating equation, one immediately faces three issues. First, the econometrician usually does not observe utilization du directly. In particular, if capital and labor utilization vary, then growth rates of the observed capital stock and labor hours do not capture the full service flows from those inputs. In the short run, firms can vary their inputs of capital and labor only by varying utilization. In this case, the regression suffers from measurement error. Unlike classical measurement error, variations in utilization du are likely to be (positively) corre-

lated with changes in the measured inputs dx, leading to an upward bias in estimated elasticities. Below, we draw on recent work in the literature on capacity utilization, when we attempt to control for variable service flow from inputs.

Second, should one take the output elasticities as constant (appropriate for a Cobb-Douglas production function or for a first-order log linear approximation), or time varying? That is, should one allow the markup and the share weights in equation (9) to change over time? If the elasticities are not truly constant over time, then treating them as constant may introduce bias. However, as we discuss later, attempting to estimate the time-varying shares may lead to more problems than it solves.

Third, even if the output elasticities are constant and all inputs are observable, one faces the "transmission problem" noted by Marschak and Andrews (1944): The technical change term, dt, is likely to be correlated with a firm's input choices, leading to biased OLS estimates. In principle, one can solve this problem by instrumenting, but the need to use instruments affects our choice of the appropriate technique for estimating equation (6). (One might expect technology improvements to lead to an expansion in inputs, making the OLS bias positive; Gali 1999 and Basu, Fernald, and Kimball 1999, however, argue that there are theoretical and empirical reasons to expect a negative bias.)

7.1.3 Empirical Implementation

We now discuss the empirical issues noted above, beginning with capacity utilization. In the estimating equation (9), we need some way to observe utilization growth du or else to control for the measurement error resulting from unobserved changes in utilization.

Our approach builds on the intuition that firms view all inputs (whether observed by the econometrician or not) identically. Suppose a firm wants more labor input but cannot instantaneously hire more workers. Then the firm should equate the marginal cost of obtaining more services from the observed intensive margin (e.g., working current workers longer hours) and from the unobserved intensive margin (working them harder each hour). If the costs of increasing hours and effort are convex, firms will choose to use both margins. Thus changes in an observed input—for example, hours per worker—provide an index of unobserved changes in the intensity of work. This suggests a regression of the form

$$dy_i = \mu_i dx_i + a_i dh_i + dt_i, \tag{10}$$

where dh_i is the growth rate of hours per worker. Earlier work by Abbott, Griliches, and Hausman (1998) also runs this regression to control for utilization.

In section 7.2, we construct a dynamic model of variable utilization that

provides complete microfoundations for this intuition. That model shows that the regression in equation (10) appropriately controls for variable effort. In addition, if the cost of varying the workweek of capital takes the form of a shift premium—for example, one needs to pay workers more to work at night—then this regression corrects for variations in utilization of capital as well as effort. (If the cost of varying capital's workweek is "wear and tear"—i.e., capital depreciates in use—then the regression is somewhat more complicated.)

Variable utilization of inputs suggests two additional problems in estimating regressions like equations (9) or (10), both of which make it difficult to observe "true" factor prices at high frequencies. First, firms will vary utilization only if inputs of capital and labor are quasi-fixed—that is, costly to adjust. Varying utilization presumably costs the firm something—for example, a shift premium. Thus, if firms could vary the number of machines or workers without cost, they would always adjust along these extensive margins rather than varying utilization. However, if inputs are costly to adjust, then the shadow price of an input to the firm may not equal its current market price. For example, investment adjustment costs imply that the return to installed capital may differ from its frictionless rental rate.

Second, varying utilization—especially of labor—is most viable when workers and firms have a long-term relationship. A firm may increase work intensity when demand is high, promising to allow workers a break in the next downturn. Such a strategy cannot be implemented if most workers are not employed by the firm when the next downturn comes. The existence of such long-term relationships suggests that wages may be set to be right "on average," instead of being the correct spot market wages in every period. Thus, both quasi-fixity and implicit labor contracts imply that we may not be able to observe factor prices at high frequencies.

This inability to observe factor prices period by period implies that we probably want to assume constant, rather than time-varying, elasticities. This is unfortunate, since the first-order equations (8) suggest that if factor shares vary over time, then output elasticities may vary as well. But if, because of quasi-fixity and implicit labor contracts, we do not observe the relevant shadow values, then observed shares do not tell us how the elasticities vary. For estimating the average markup, a first-order approximation may suffice, and may be relatively unaffected by our inability to observe factor prices at high frequency. Of course, if our goal were to estimate elasticities of substitution between inputs, then we would need to find some way to deal with these problems—we could not use a first-order approximation that simply assumes the elasticities are one.[9]

9. For example, Berman, Bound, and Griliches (1994) and others investigate the hypothesis that technological progress is skill-biased, and hence contributed to the increase in income inequality in the United States in recent decades; Jorgenson (1987) argues that part of the productivity slowdown of the 1970s and 1980s is due to energy-biased technical change.

With these considerations in mind, we now assess four methods of estimating the parameters in the production function. First, one can simply take equation (6) as an estimating equation and estimate the output elasticities directly. This approach, which essentially assumes that the production function is Cobb-Douglas, can be justified as a first-order approximation to a more general production function. But this procedure requires us to estimate three parameters (or more, if we include proxies for variable utilization) for the output elasticities using data that are often multicollinear, and also subject to differing degrees of endogeneity and hence differing OLS biases. The use of instrumental variables is not a complete solution, since most plausible instruments are relatively weak. The literature on weak instruments suggests that instrumental-variables methods have difficulty with multiple parameters; see, for example, Shea (1997).

Second, one can impose cost minimization and estimate equations (9) or (10), while still taking a first-order approximation. Cost minimization is a relatively weak condition, and seems likely to hold at least approximately. Imposing it allows us to move from estimating three parameters to estimating only one (the markup), thus increasing efficiency. (Of course, assuming a particular model of cost minimization, if it is inappropriate, can lead to specification error. For example, if firms are *not* price takers in factor markets—for example, if firms have monopsony power or, in general, face price-quantity schedules for their inputs rather than single prices—then cost-minimizing conditions, and hence the Hall equations [9] or [10], are misspecified.)

In this case, one simply assumes that the shares used in constructing dx_i are constant, as is the markup, μ_i. As with the first approach, if our interest is in first-order properties such as average markups or returns to scale, it can in principle give an accurate answer, although the omitted second-order terms may well be important. (Risk aversion is a second-order phenomenon, but there is an active insurance industry.) A substantial advantage of the first-order approach is that it does not require that observed factor prices (or rental rates) be allocative in every period. The average shares are likely to be close to the steady-state shares, so the approximation typically will be correct even with quasi-fixed inputs or implicit contracts.

Third, one could continue to estimate equations (9) or (10), but allow the shares to change period by period. If markups are constant, if factor prices are allocative period by period, and if there are no quasi-fixed factors, then this approach can, in principle, give a second-order approximation to any production function (Diewert 1976). If these conditions fail, then this approach in essence incorporates some second-order effects, but not others; it is unclear whether this is preferable to including none of them.

Fourth, one could estimate a flexible, general functional form along with Euler equations for the quasi-fixed inputs; as Nadiri and Prucha (this volume) discuss, this approach provides a complete second-order approxima-

tion to any production technology. If we properly parameterize the firm's problem, then the markup need not be constant, and factors can be quasi-fixed—the model provides estimates of the true shadow values. In principle, this full structural approach provides a complete characterization of the technology; one can then calculate elasticities of substitution, biases of technological change, and so forth. For some macroeconomic questions, these parameters are crucial.[10]

However, this general approach has the disadvantage that one needs to estimate many parameters (a translog production function for equation [1], for example, has twenty-five parameters before imposing restrictions, to say nothing of the associated Euler equations). Efficient estimation requires various restrictions and identifying assumptions, such as estimating the first-order conditions (equation [7]) along with the production function itself. Hence, results may be sensitive to misspecification. For example, if wages are determined by an implicit contract, and hence are not allocative period by period, then one would not want an approach that relies heavily on high-frequency changes in observed factor prices (and hence factor shares) for identification. The structural approach can be estimated either through the production or cost function—that is, from the primal or the dual side. Our concerns apply to both. We discuss these issues at greater length in section 7.2.3.

Since our primary interest is in measures of technical change, average returns to scale, and the average markup, in this paper we follow the second approach outlined above, using an explicitly first-order approximation. (In practice, this gives qualitatively similar results to the third approach, which allows the factor shares to vary over time). This second method is essentially the procedure of Hall (1990), generalized to include materials input and controls for variable factor utilization. Although this approach allows us to estimate the parameters governing firm-level technology and behavior, our ultimate interest is describing the evolution of aggregates. Thus, we must take one more step and aggregate output growth across firms.

7.1.4 Aggregation

So far, we have discussed production and estimation in terms of firm-level gross output. Our ultimate interest is in aggregate value added. In general, no aggregate production function exists that links aggregate output to aggregate inputs—but the relationship between these aggregates remains of interest. It turns out that aggregation across firms can introduce a significant new source of procyclical productivity.

10. For example, a recent strand of business-cycle theory emphasizes the cyclical properties of markups as important propagation mechanisms for output fluctuations; see, for example, Rotemberg and Woodford (1992, 1995). Time-series variation in the markup is a second-order property that can only be estimated with a second-order (or higher) approximation to the production function.

In section 7.3, we derive the following equation for aggregate output (value added) growth dv:

$$dv = \bar{\mu}^V dx^V + du + R + dt^V. \tag{11}$$

$\bar{\mu}^V$ is the average "value added" markup across firms, du is an appropriately weighted average of firm-level utilization rates, R represents various reallocation (or aggregation) effects, and dt^V is an appropriately weighted average of firm-level technology. dx^V is an weighted average of the aggregate capital stock and labor hours. In all cases, the superscript V refers the fact that aggregate output is a value-added measure rather than a gross-output measure, which requires some minor changes in definitions. (As we discuss in section 7.3, for macroeconomic modeling it is the "value-added markup" that is likely to be of interest.)

The major implication of equation (11) is that output growth at the aggregate level is not completely analogous to output growth at the firm level. The firm-level equation (9) looks similar to equation (11). Firm-level output growth depends on input growth, the markup, variations in utilization, and technical change; aggregate output growth depends on aggregate input growth, the average markup, average variation in utilization, and average technical change. Equation (11), however, has a qualitatively new term, R.

The reallocation term reflects the effect on output growth of differences across uses in the (social) values of the marginal products of inputs. Output growth therefore depends on the distribution of input growth as well as on its mean: If inputs grow rapidly in firms where they have above-average marginal products, output grows rapidly as well. Thus, aggregate productivity growth is not just firm-level productivity growth writ large. There are qualitatively new effects at the aggregate level, which may be important both for estimating firm-level parameters and as powerful amplification and propagation mechanisms in their own right.

For example, suppose that some firms have large markups of price over marginal cost while others have low markups. Also suppose that all firms pay the same prices for their factors. Then resources such as labor have a higher marginal product in the firms with the larger markup, as shown by the first-order condition (equation [8]). Intuitively, because these firms have market power, they produce too little output and employ too few resources; hence, the social value of the marginal product of these inputs is higher. The reallocation term R captures the fact that aggregate output rises if resources shift from low- to high-markup firms.

To summarize this section, we argue in favor of the following approach to estimating technology change, controlling for utilization. First, interpret all results as first-order approximations, on the grounds that the data are probably insufficient to allow reliable estimation of second-order approximations. Second, estimate equation (10)—a Hall (1990)-style regression

with a theoretically justified utilization proxy—at a disaggregated level. Take the residuals as a measure of disaggregated technology change. Third, use the aggregation equation (11)—which incorporates the disaggregated estimates of technology change, markups, and utilization—to identify the importance of the various explanations for procyclical productivity.

The casual reader is now well prepared to proceed to the data and empirical sections of this paper in sections 7.4 and 7.5. However, our discussion passed quickly over several technical details. To fully justify our choice of method, it is useful to address those issues in detail. We do so in sections 7.2 and 7.3.

7.2 The Meaning and Measurement of Capacity Utilization

Section 7.1 raised several issues with empirical implementation of the Solow-Hall approach to estimating technology, markups, and variations in utilization at a disaggregated level. First, we must decide what prices to use in calculating weights. With quasi-fixed inputs, the appropriate shadow price is not, in general, the observed factor price. In addition, even if factors are freely variable, the observed factor prices may not be allocative: For example, firms and workers may have implicit contracts. Second, we must decide whether to use a first- or second-order approximation to the continuous time equation (10). Third, we must find suitable proxies for du.

This section provides explicit microfoundations for our preferred approach. We specify a dynamic cost-minimization problem to provide appropriate shadow values. We then discuss the pros and cons of first- versus second-order approximations. Finally, we use the first-order conditions from the cost minimization problem to find observable proxies for unobserved effort and capital utilization.

We argue in favor of using an explicitly first-order approximation for the estimating equation, on the grounds that the data necessary for an appropriate second-order approximation are unavailable. We also argue for using growth in hours per worker to adjust for variations in effort and capital utilization. We also spend some time clarifying conceptual confusions over the meaning of capacity utilization. In particular, we argue that correcting for capacity utilization requires correcting quantities, not merely output elasticities.

7.2.1 A Dynamic Cost-Minimization Problem

We now specify a particular dynamic cost-minimization problem. Although the problem is relatively complicated, specifying it provides insight into several practical issues in attempting to estimate equation (10) in the previous subsection, and, in section 7.2.4, provides proxies for unobserved utilization.

We model the firm as facing adjustment costs in both investment and hiring, so that both the amount of capital (number of machines and buildings), K, and employment (number of workers), N, are quasi-fixed. We model quasi-fixity for two reasons. First, we want to examine the effect of quasi-fixity per se on estimates of production function parameters and firm behavior. Second, quasi-fixity is necessary for a meaningful model of variable factor utilization. Higher utilization must be more costly to the firm, otherwise factors would always be fully utilized. If there were no cost to increasing the rate of investment or hiring, firms would always keep utilization at its minimum level and vary inputs using only the extensive margin, hiring and firing workers and capital costlessly. Only if it is costly to adjust along the extensive margin is it sensible to adjust along the intensive margin, and pay the costs of higher utilization.[11]

We assume that the number of hours per week for each worker, H, can vary freely, with no adjustment cost. In addition, both capital and labor have freely variable utilization rates. For both capital and labor, the benefit of higher utilization is its multiplication of effective inputs. We assume the major cost of increasing capital utilization, Z, is that firms may have to pay a shift premium (a higher base wage) to compensate employees for working at night, or at other undesirable times.[12] We take Z to be a continuous variable for simplicity, although variations in the workday of capital (i.e., the number of shifts) are perhaps the most plausible reason for variations in utilization. The variable-shifts model has had considerable empirical success in manufacturing data, where, for a short period of time, one can observe the number of shifts directly.[13] The cost of higher labor utilization, E, is a higher disutility on the part of workers that must be compensated with a higher wage. We allow for the possibility that this wage is unobserved from period to period, as might be the case if wage payments are governed by an implicit contract in a long-term relationship.

Consider the following cost minimization problem for the representative firm of an industry:

11. One does not require *internal* adjustment costs to model variable factor utilization in an aggregative model (see, e.g., Burnside and Eichenbaum, 1996), because changes in input demand on the part of the representative firm change the aggregate real wage and interest rate, so in effect the concavity of the representative consumer's utility function acts as an adjustment cost that is *external* to the firm. However, if one wants to model the behavior of firms that vary utilization in response to idiosyncratic changes in technology or demand—obviously the case in the real world—then one is forced to posit the existence of internal adjustment costs in order to have a coherent model of variable factor utilization. (Both of these observations are found in Haavelmo's 1960 treatment of investment.)

12. Our model can be extended easily to allow utilization to affect the rate at which capital depreciates, as in Basu and Kimball (1997). We consider the simpler case for ease of exposition. Nadiri and Prucha (this vol.) show that their approach of estimating a second-order approximation to the production function can also accommodate variable depreciation, but do not consider either a shift premium or variable labor effort.

13. See, for example, Beaulieu and Mattey (1998) and Shapiro (1996).

$$\min_{Z,E,H,M,I,A} C(Y) = \int_0^{\infty} e^{-\int_0^s r d\tau} [WNG(H,E)V(Z) + P_M M + WN\Psi(A/N) + P_I KJ(I/K)]ds \tag{12}$$

subject to

$$Y = F(ZK, EHN, M, T) \tag{13}$$

$$\dot{K} = I - \delta K \tag{14}$$

$$\dot{N} = A \tag{15}$$

The production function and inputs are as before. In addition, I is gross investment, and A is hiring net of separations. $WG(H, E)V(Z)$ is total compensation per worker (compensation may take the form of an implicit contract, and hence not be observed period by period). W is the base wage; the function G specifies how the hourly wage depends on effort, E, and the length of the workday, H; and $V(Z)$ is the shift premium. $WN\Psi(A/N)$ is the total cost of changing the number of employees; $P_I KJ(I/K)$ is the total cost of investment; P_M is the price of materials; and δ is the rate of depreciation. We continue to omit time subscripts for clarity.

Using a perfect-foresight model amounts to making a certainty equivalence approximation. However, even departures from certainty equivalence should not disturb the key results, which rely only on intratemporal optimization conditions rather than intertemporal ones.

We assume that Ψ, and J are convex and make the appropriate technical assumptions on G in the spirit of convexity and normality.[14] It is also helpful to make some normalizations in relation to the normal or steady-state levels of the variables. Let $J(\delta) = \delta$, $J'(\delta) = 1$, $\Psi(0) = 0$. We also assume that the marginal employment adjustment cost is zero at a constant level of employment: $\Psi'(0) = 0$.

We use the standard current-value Hamiltonian to solve the representative firm's problem. Let λ, q, and θ be the multipliers on the constraints in equations (13), (14), and (15), respectively. Numerical subscripts denote derivatives of the production function F with respect to its first, second, and third arguments, and literal subscripts denote derivatives of the labor cost function G. The six intratemporal first-order conditions for cost-minimization are

$$Z: \quad \lambda KF_1(ZK, EHN, M; T) = WNG(H,E)V'(Z). \tag{16}$$

$$H: \quad \lambda ENF_2(ZK, EHN, M; T) = WNG_H(H,E)V(Z) \tag{17}$$

14. The conditions on G are easiest to state in terms of the function Φ defined by $\ln G(H, E) = \Phi(\ln H, \ln E)$. Convex Φ guarantees a global optimum; assuming $\Phi_{11} > \Phi_{12}$ and $\Phi_{22} > \Phi_{12}$ ensures that optimal H and E move together.

(18) $$E: \quad \lambda HNF_2(ZK, EHN, M; T) = WNG_E(H, E)V(Z)$$

(19) $$M: \quad \lambda F_3(ZK, EHN, M; T) = P_M$$

(20) $$A: \quad \theta = W\Psi'(A/N)$$

(21) $$I: \quad q = P_I J'(I/K).$$

The Euler equations for the capital stock and employment are

(22) $$\dot{q} = (r + \delta)q - \lambda ZF_1 + P_I[J(I/K) - (I/K)J'(I/K)]$$

(23) $$\dot{\theta} = r\theta - \lambda EHF_2 + WG(H, E)V(Z) + W[\Psi(A/N) - (A/N)\Psi'(A/N)].$$

Since λ is the Lagrange multiplier associated with the level of output, one can interpret it as marginal cost. The firm internally values output at marginal cost, so λF_1 is the marginal value product of effective capital input, λF_2 is the marginal value product of effective labor input, and λF_3 is the marginal value product of materials input.[15] We defined the markup, μ, as the ratio of output price, and P, to marginal cost, so λ equals

(24) $$\lambda = C'(Y) = \frac{P}{\mu}.$$

Note that equation (24) is just a definition, not a theory determining the markup. The markup depends on the solution of the firm's more complex profit maximization problem, which we do not need to specify.

Equations (22) and (23) implicitly define the shadow (rental) prices of labor and capital:

(25) $$\lambda ZF_1 = (r + \delta)q - \dot{q} + P_I[J(I/K) - (I/K)J'(I/K)] \equiv P_K$$

(26) $$\lambda EHF_2 = r\theta - \dot{\theta} + WG(H, E)V(Z) + W[\Psi(A/N) - (A/N)\Psi'(A/N)] \equiv P_L.$$

As usual, the firm equates the marginal value product of each input to its shadow price. Note that with these definitions of shadow prices, the atemporal first-order condition in equation (7) holds for all inputs. For some intuition, note that equation (25) is the standard first-order equation from a q-model of investment. In the absence of adjustment costs, the value of installed capital q equals the price of investment goods P_I, and the "price" of capital input is then just the standard Hall-Jorgenson cost of capital $(r + \delta)P_I$. With investment adjustment costs, there is potentially

15. For the standard static profit-maximization problem, of course, marginal cost equals marginal revenue, so these are also the marginal revenue products.

an extra return to owning capital, through capital gains $\dot{q}$ (as well as extra terms reflecting the fact that investing today raises the capital stock, and thus lowers adjustment costs in the future).

The intuition for labor in equation (26) is similar. Consider the case where labor can be adjusted freely, so that it is not quasi-fixed. Then adjustment costs ψ are always zero. So is the multiplier θ, since the constraint in equation (15) does not bind. In this case, as we expect, equation (26) says that the shadow price of labor input to the firm—the right side of the equation—just equals the (effort adjusted) compensation $WG(H, E)V(Z)$ received by the worker. Otherwise, the quasi-fixity implies that the shadow price of labor to a firm may differ from the compensation received by the worker.

7.2.2 First-Order Approximations

We now turn to issues of estimation. This subsection discusses how to implement our preferred first-order approximation; the next subsection compares this approach with second-order approximations.

Equations (6) and (9) hold exactly in continuous time, where the values of the output elasticities adjust continuously. In discrete time, we can interpret these equations as first-order approximations (in logs) to any general production function if we assume the elasticities are constant. For a consistent first-order approximation to equation (9), one should interpret it as representing small deviations from a steady-state growth path, and evaluate derivatives of the production function at the steady-state values of the variables. Thus, to calculate the shares in equation (9), one should use steady-state prices and quantities, and hence treat the shares as constant over time. The markup is then also taken as constant.

For example, in the first order approach, we want the steady-state output elasticity for capital, up to the unknown scalar μ. Using asterisks to denote steady-state values, we use equations (21) and (25) and our normalizations to compute the steady-state output elasticity of capital:

$$\frac{F_1^* Z^* K^*}{Y^*} = \mu^* \frac{P_K^* K^*}{P^* Y^*} \equiv \mu^* \frac{(r^* + \delta) P_I^* K^*}{P^* Y^*}. \tag{27}$$

Note that the steady-state user cost of capital is the frictionless Hall-Jorgenson (1967) rental price.[16] Since quasi-fixity matters only for the adjustment to the steady state, in the steady state $q = P_I$ and $\dot{q} = 0$. Operationally, we calculate the Hall-Jorgenson user cost for each period and take the time average of the resulting shares as an approximation to the steady-state share. We proceed analogously for the other inputs. In the final esti-

16. In practice, one would also include various tax adjustments. We do so in the empirical work, but omit them in the model to keep the exposition simple.

mating equation for equation (9), we use logarithmic differences in place of output and input growth rates, and use steady-state shares for the weights.

Thus, we can construct the index of observable inputs, dx, and take the unknown μ^* multiplying it as a parameter to be estimated. We can use a variety of approaches to control for the unobserved du; we discuss some of them in section 7.2.4. Alternatively, under the heroic assumption that du is always zero or is uncorrelated with dx, we can estimate equation (9) while simply ignoring du. Hall (1988, 1990) and Basu and Fernald (1997a) follow this second procedure.

In any case, we have to use instruments that are orthogonal to the technology shock dt, since technology change is generally contemporaneously correlated with input use (observed or unobserved).[17]

7.2.3 Second-Order Approximations

The first-order approach of constant weights is, of course, equivalent to assuming that for small, stationary deviations from the steady-state balanced growth path, we can treat the production function as Cobb-Douglas. Parameters such as the average markup or degree of returns to scale are first-order properties, so the bias from taking a first-order approximation may not be large.

Nevertheless, the Cobb-Douglas assumption is almost surely not literally true. In principle, using a second-order approximation allows us to eliminate some of the bias in parameter estimates from using Cobb-Douglas and eliminate approximation errors that end up in the residual. It also allows us to estimate second-order properties of production functions, such as separability and elasticities of substitution. (Nadiri and Prucha, this volume) discuss the benefits of the second-order approach.) Unfortunately, the second order approach suffers severe practical disadvantages—particularly the increased likelihood of misspecification. Our view is that these sizeable disadvantages outweigh the potential benefits.

We begin this discussion with a simplest case where one obtains an appropriate second-order approximation to equation (9) simply by using weights that change period by period. Suppose the production function takes a more general form than Cobb-Douglas, for example, translog. With a more general production function, output elasticities will typically vary over time if relative factor prices are not constant. Also suppose the markup is constant, and that all factor shadow values (i.e., input prices) are observed, as will be the case if observed prices are allocative and there are no costs of adjusting inputs. With a translog function (which provides a flexible approximation to any functional form), the discrete-time Tornqvist

17. Olley and Pakes (1996) propose an insightful alternative to the usual instrumental-variables approach; see Griliches and Mairesse (1998) for an excellent discussion of the pros and cons.

approximation to the continuous time equation (9) turns out to be exactly correct.[18] This approximation requires replacing growth rates with log differences and replacing the continuous-time input shares with average shares in adjacent periods. Thus, for example, the output elasticity for capital is approximated by $(1/2)(s_{Kt} + s_{Kt-1})$, and is multiplied by the capital change term $(\ln K_t - \ln K_{t-1})$. In this case, using a Tornqvist index of input use allows a correct second-order approximation.

A major potential shortcoming with using changing shares is that observed factor prices may not be allocative period by period because of implicit contracts or quasi-fixity. Then observed factor shares might not be proportional to output elasticities period by period, and the Tornqvist index is misspecified.

A large literature has focused on the problem of quasi-fixity of capital and labor, while largely ignoring the concern about implicit contracts. We briefly review that approach to estimating second-order approximations. As discussed in the introduction, the time-varying elasticity literature—for example, Berndt and Fuss (1986), Hulten (1986), and the much more parametric literature surveyed in Nadiri and Prucha (this volume)—tries explicitly to deal with quasi-fixity. Quasi-fixity may lead to large variations in input shadow prices, which in turn may cause the output elasticities to vary over time.[19] The shadow prices may differ substantially from observed prices; in our cost-minimization setup of section 7.2.1, equations (25) and (26) show the relationship between shadow price and observed prices. For example, a firm may find that it has too many workers, so that the output elasticity with respect to labor is very low. If the firm cannot (or chooses not to) immediately shed labor, it will be forced to continue paying these workers. Then the observed labor share may be higher than the true output elasticity, since the shadow value of labor is less than the wage.

Hulten (1986) shows that even with quasi-fixity, if there are constant returns and perfect competition, then one can implement the Tornqvist approximation to equation (9) using observed input prices and output growth as long as there is only a single quasi-fixed input (e.g., capital). Under these conditions, the revenue shares sum to one. Since the observed input prices give the correct shares for all inputs other than capital, capital's share can be taken as a residual. (As we emphasize again below, this approach corrects only for variations in the shadow values of quasi-fixed inputs, not unobserved changes in the workweek of capital.)

18. This result follows from Diewert (1976), who shows that the Tornqvist index is "superlative"—that is, exact for functional forms such as the translog that are themselves flexible approximations to general functions—and shows that superlative indices have desirable index-number properties. For discussions of the concept of "flexibility" and its limitations, see Lau (1986) and Chambers (1988, ch. 5).

19. As our previous discussion indicates, the output elasticities will generally change whenever input prices (shadow values) change. Quasi-fixity is only one reason why shadow values change.

Multiple quasi-fixed inputs cannot be accommodated using the nonparametric approach of equation (9), but one can estimate the parameters of a structural cost function. Pindyck and Rotemberg (1983) and Shapiro (1986), for example, estimate particular parametric forms of the production function along with Euler equations for the quasi-fixed inputs, such as equations (22) and (23). As a by-product, this approach provides estimates of the shadow values of the state variables, and thus one can construct a measure of period-by-period shadow costs by valuing the input of each quasi-fixed factor at its shadow price.

The problem grows even more challenging when one allows for nonconstant returns to scale. Since one cannot estimate the scale elasticity from the first-order conditions for the cost shares (e.g., see Berndt 1991, chapter 9), one must estimate both the cost function and the share equations together. But output is a right-hand side variable in the cost function; since the error term is interpreted as technical change, output is clearly endogenous. Pindyck and Rotemberg (1983) deal with this problem by using lagged variables as instruments, but it is unclear to what extent this procedure alleviates the problem. For example, technical change—the error term in the cost function—can be (and usually is found to be) serially correlated. Finally, one can compute the markup from the estimates of returns to scale (in this context, a time-varying parameter) by using equation (5) and the observed prices and estimated shadow prices to construct a period-by-period estimate of economic profit.

Morrison (1992a,b) attempts to construct better instruments by estimating industry demand curves embodying shift variables and then imputes values of output using the simple, static, monopoly pricing formula for the markup. However, there are also major problems with this procedure. First and most importantly, it implies that μ is no longer identified from cost minimization conditions alone; one has to subscribe to a particular model of firm behavior and use profit maximization conditions as well. This change is a major loss in generality. Second, Morrison's specific model of firm behavior assumes that firms can collude perfectly in every period and that prices are completely flexible, and hence is misspecified if the degree of collusion varies (as stressed, e.g., by Rotemberg and Saloner 1986 and Green and Porter 1984), or if prices are sticky in the short run, leading the actual markup to deviate from the optimal markup (see, e.g., Ball and Romer 1990 and Kimball 1995).

The challenges raised by these issues suggest that misspecification is a serious concern. Even in the best of cases, where factor prices for variable factors are correctly observed and utilization does not vary, one needs to specify correctly the problem and constraints and then estimate a large number of parameters. This complexity makes misspecification harder to spot.

Of course, this best of cases is unlikely to hold, so misspecification is

probably much worse. First, we have strong reasons to think that input prices observed by the econometrician may *not* be allocative period by period because of implicit contracts. Since many economic relationships are long term, it may not be worth recontracting explicitly period by period. Then the correct shadow price of a variable input may differ from its observed price. The shadow prices may satisfy the first-order conditions period by period, while the observed factor payments may satisfy the conditions only on average. With implicit contracts, the data that are available hinder accurate estimation of a second-order approximation—even if the full problem and constraints were correctly specified.[20]

Second, variations in capacity utilization—that is, in labor effort E and capital's utilization/workweek Z—also lead to misspecification. Since these utilization margins are not observed by the econometrician, they remain as omitted variables in estimating a fully parametric production function. Using the dual approach does not help. In specifying the cost function, one needs to specify what margins the firm can use to adjust—and at what price. But it is virtually impossible for the econometrician to correctly observe the "prices" of varying effort or shifts to a firm. For example, Bils (1987) discusses the importance of the premium paid for overtime work. The wage function $G(H, E)$ in section 7.2.1 incorporates any overtime premium. However, the fully parameterized dual approach generally prices labor at the average wage rather than the correct marginal wage. Hence, the fully parameterized production or cost functions are almost surely misspecified.

The misspecification arising from omitted capacity utilization is sometimes obscured by the parameterized second-order literature, in part because that literature sometimes claims to solve the utilization problem. They are addressing a different issue. That is, there are two distinct concepts of utilization in the literature: what we call capacity utilization, and the Berndt-Fuss concern that inputs may not always be at their steady-state levels.

The Berndt-Fuss notion of utilization addresses the possible time-variation in output elasticities caused by time-varying shadow prices of quasi-fixed inputs. Capacity utilization, on the other hand, refers to the much earlier idea that there is a particular type of measurement error in the inputs: Certain factors—again, notably capital and labor—have variable service flow per unit of observed input (the dollar value of machinery or the number of hours worked). Many commentators have explained the

20. Carlton (1983) also stresses that observed materials prices may not be allocative if firms use delivery lags to clear markets. A similar problem arises from labor composition changes. Solon, Barsky, and Parker (1994) find that the marginal worker in a boom is of relatively low quality—and hence relatively low paid—so that if labor data do not adjust for this composition effect, the wage will appear spuriously countercyclical. Cost-function estimation that uses spuriously cyclical wage data is, of course, misspecified.

procyclicality of productivity and the short-run increasing returns to labor (SRIRL) puzzle by arguing that variations in the workweek of capital and changes in labor effort increase effective inputs much more than observed inputs. In terms of equation (9), they argue that *du* is nonzero, and is positively correlated with *dy* and *dx*.

It is easy to see the difference between the two concepts in the case where *F* is actually known to be a Cobb-Douglas production function. Since the elasticities are truly constant over time, the first-order discrete-time implementation of equation (9) is exact. The *observed* factor shares might vary over time, but since we know that *F* is Cobb-Douglas, these variations would reflect the effects of quasi-fixity rather than changes in output elasticities. Quasi-fixity of inputs thus creates no problems for estimation, since one can observe and use the steady-state prices and shares. Indeed, the Berndt-Fuss approach would "correct" the time-varying shares, using constant shares instead. Despite this correction, variations in factor utilization remains a problem. If one cannot somehow control for *du*, then estimates of the markup or returns to scale will generally be biased.

Formally, therefore, the problems posed by quasi-fixity are second order. If the first-order approximation of Cobb-Douglas is actually exact, then quasi-fixity—or, more generally, time-varying elasticities—cannot matter for the estimation of μ. On the other hand, capacity utilization remains a first-order problem, since it concerns the measurement of the right-hand-side variables. It continues to pose a problem even if the Cobb-Douglas approximation holds exactly.

The two concepts are often confused because, as we argued above, quasi-fixity is a necessary condition for capacity utilization to vary. As a result, the shadow value of an extra worker will be high at exactly the same time that unobserved effort is high. Because of this relationship, some authors in the production-function literature assert that the capacity utilization problem can be solved by allowing elasticities to vary over time, or by using a dual approach and allowing shadow values to vary from long-run equilibrium levels. Many of these authors do not appear to realize that there are two separate problems, and allowing for quasi-fixity does not solve the problem that the workweek of capital or the effort of workers may vary over time.[21] That is, there is no reason that the shadow value computed from a misspecified variable cost function will necessarily capture all of the effects of variable utilization.

We conclude this section by returning to the first-order approach. In

21. The survey by Nadiri and Prucha (this vol.) does try to address both issues. They focus primarily on the issue of time-varying elasticities. However, they also allow capital's workweek to vary, where the cost of increasing capital utilization is "wear and tear"—capital depreciates faster. They do not consider either a shift premium or variable labor effort (or the consequent problem that observed wages may not be allocative).

particular, though estimation using second-order approximations is likely to be misspecified, the problems that approach addresses—such as quasi-fixity—are clearly concerns. So given these concerns, how robust is the first-order approach that ignores them?

It is plausible that the first-order approach is relatively robust. That is, ignoring quasi-fixity probably does not significantly affect estimates of markups and returns to scale. First, quasi-fixity affects only the period-by-period computation of input shares, not the growth rate of capital (or any other quasi-fixed input). Since these shares are constant to a first-order Taylor approximation, any errors caused by failure to track the movements of the shares is likely to be small. Second, quasi-fixity is likely to be most important for capital. (Shapiro 1986 finds that quasi-fixity is not important for production-worker labor, although it is present for nonproduction workers). But mismeasurement of the rental rate of capital affects only capital's share, and since the growth rate of capital is almost uncorrelated with the business cycle, errors in measuring capital's share are unlikely to cause significant biases in a study of cyclical productivity. Caballero and Lyons (1989) present simulations indicating that maximum biases from ignoring quasi-fixity of capital are likely to be on the order of 3 percent of the estimated coefficients. (However, capital utilization *is* cyclical, and it is also multiplied by capital's share.)

In sum, the second-order approach offers great theoretical advantages that, in practice, it cannot deliver. The failures are understandable, since the problems are difficult. Given these practical difficulties, we prefer the explicitly first-order approach to estimation. This has costs as well, such as ignoring quasi-fixity for estimation purposes and the inability to estimate elasticities of substitution, time variation in the markup, and other second-order properties. Nevertheless, results on first-order properties are likely to be relatively robust.

7.2.4 Capacity Utilization

Before we can estimate μ from equation (9), we need to settle on a method for dealing with changes in utilization, du. A priori reasoning—and comparisons between results that control for du and those that do not—argue that du is most likely positively correlated with dx; thus ignoring it leads to an upward-biased estimate of μ^*. Three general methods have been proposed. First, one can try to observe du directly using, say, data on shiftwork. When possible this option is clearly the preferred one, but data availability often precludes its use.[22] Second, one can impose a priori restrictions on the production function. Third, one can derive links

22. In the United States, shift-work data are available only for a relatively short time period, and solely for manufacturing industries. The only data set on worker effort that we know of is the survey of British manufacturing firms used by Schor (1987).

between the unobserved *du* and observable variables using first-order conditions like equations (16)–(21). Both the second and third approaches imply links between the unobserved *du* and observable variables, which can be used to control for changes in utilization.

The approach based on a priori restrictions basically imposes separability assumptions on the production function. For example, Jorgenson and Griliches (1967) use an idea going back at least to Flux (1913), and assume that effective capital input is a function of capital services and energy (this idea has recently been revived by Burnside, Eichenbaum, and Rebelo 1995; we henceforth refer to it as the Flux assumption). To clarify this idea, it will be useful to separate intermediate goods M into two components: the flow of energy inputs, W (mnemonic: Wattage); and all other intermediate inputs O (mnemonic: Other). Then the Flux assumption implies:

$$F(\tilde{K}, EHN, W, O; T) = G[S(\tilde{K}, W), EHN, O; T]. \tag{28}$$

Jorgenson-Griliches and Burnside and colleagues generally also assume that

$$S(\tilde{K}, W) = \min[\tilde{K}, W], \tag{29}$$

so that energy input is a perfect index of capital input.[23]

Using this implication of equation (29) to substitute into equation (9), we have

$$\begin{aligned} dy &= \mu[s_K(dk + dz) + s_L(dn + dh + de) + s_W dw + s_o do] + dt \\ &= \mu[(s_K + s_w)dw + s_L(dn + dh) + s_o do] + \mu s_L de + dt. \end{aligned} \tag{30}$$

Note that under the maintained hypothesis of equation (28), electricity use proxies for changes in capital utilization, but does not capture variations in labor effort. (The same is true if one uses direct observations on utilization, like the number of shifts.) Electricity use is also a much more sensible proxy for the utilization of heavy machinery than for the services of structures or light machinery like computers, which are often left on day and night regardless of use. Thus, electricity use is probably a reasonable proxy only within manufacturing—and even there it does not capture variations in effort.

Basu (1996) attempts to control for both labor effort and capital utilization by using either of two different separability assumptions:

23. Burnside et al. (1995) also experiment with a CES functional form for H, and allow the ratio of capital to energy to change depending on their relative prices. The generalization to the CES does not affect their results significantly, but their measures of the rental price of capital services are questionable.

$$F(ZK, EHN, W, O; T) = TG[V(ZK, EHN), S(W, O)], \tag{31}$$

or

$$F(ZK, EHN, W, O; T) = TG[V(ZK, W, EHN), S(O)]. \tag{32}$$

In both cases, he assumes that G takes the Leontief form:[24]

$$G(V, S) = \min(V, S). \tag{33}$$

The basic intuition behind the separability assumption is the distinction between the materials inputs that are being used up in production and the inputs that are assembling the materials into final output. In the simplest case, equation (32), the estimating equation is just

$$dy = \mu do + dt. \tag{34}$$

Under the maintained hypothesis of equation (33), variations in materials use capture changes in both capital and labor utilization.

A problem with separability-based methods of controlling for utilization is that they rely crucially on the assumption that the production function is homothetic (i.e., relative input demands do not depend on the level of output).[25] In other words, it is important that the function S in the Flux case or the functions V and S in the Basu case all be homogenous of degree one. Basu (1996) discusses this issue in detail, and argues that if homotheticity does not hold (or if one does not take additional steps), the estimated values of μ from equations (30) and (34) are likely to be biased downward. With this caveat, however, the intuitive approaches presented here offer relatively easy ways to control for changes in utilization.

Bils and Cho (1994), Burnside and Eichenbaum (1996), and Basu and Kimball (1997) argue that one can also control for variable utilization using the relationships between observed and unobserved variables implied by first-order conditions like equations (16)–(21). Our discussion follows Basu and Kimball.

They begin by assuming a generalized Cobb-Douglas production function,

$$F(ZK, EHN, M; Z) = Z\Gamma[(ZK)^{\alpha_K}(EHN)^{\alpha_L}M^{\alpha_M}], \tag{35}$$

where Γ is a monotonically increasing function. In their case this assumption is not merely a first-order approximation, because they make use of the second-order properties of equation (35), particularly the fact that the ratios of output elasticities are constant. Although they argue that one can relax the Cobb-Douglas assumption, we shall maintain it throughout our discussion.

24. Basu also experiments with CES specifications for G, but allowing deviations from the Leontief assumption has little effect on his estimates.

25. See Chambers (1988) for a discussion of homotheticity.

Equations (17) and (18) can be combined into an equation implicitly relating E and H:

$$\frac{HG_H(H,E)}{G(H,E)} = \frac{EG_E(H,E)}{G(H,E)}. \tag{36}$$

The elasticity of labor costs with respect to H and E must be equal, because on the benefit side the elasticities of effective labor input with respect to H and E are equal. Given the assumptions on G, (36) implies a unique, upward-sloping E-H expansion path, so that we can write

$$E = E(H), \quad E'(H) > 0. \tag{37}$$

Equation (37) says that the unobservable intensity of labor utilization E can be expressed as a monotonically increasing function of the observed number of hours per worker H. This result also holds in growth rates; thus,

$$d\ln(EHN) = dn + dh + de = dn + (1 + \zeta)dh. \tag{38}$$

Finding the marginal product of labor from equation (35) and substituting into the first-order condition for hours per worker, equation (17), we find

$$WNH\,G_H(H,E)V(U) = \lambda\gamma\alpha_L Y. \tag{39}$$

Substituting the marginal product of capital and equation (39) into equation (16) yields

$$\lambda\gamma\alpha_K \frac{Y}{Z} = \lambda\gamma\alpha_L Y \frac{G(H,E)}{HG_H(H,E)} \frac{V'(Z)}{V(Z)}. \tag{40}$$

Define

$$g(H) = \frac{HG_H(H,E(H))}{G(H,E(H))} \tag{41}$$

and

$$v(Z) = \frac{ZV'(Z)}{V(Z)}. \tag{42}$$

$v(Z)$ is thus the ratio of the marginal shift premium to the average shift premium. Rearranging, we get

$$1 = \frac{\alpha_L}{\alpha_K} \frac{v(Z)}{g(H)}. \tag{43}$$

The labor cost elasticity with respect to hours given by the function $g(H)$ is positive and increasing by the assumptions we have made on $G(H,E)$. The labor cost elasticity with respect to capital utilization given by the

function $v(Z)$ is positive as long as there is a positive shift premium. We also assume that the shift premium increases rapidly enough with Z to make the elasticity increasing in Z.

First, define

$$\eta = \frac{H^* g'(H^*)}{g(H^*)}, \tag{44}$$

and

$$v = \frac{Z^* v'(Z^*)}{v(Z^*)}. \tag{45}$$

η indicates the rate at which the elasticity of labor costs with respect to hours increases. v indicates the rate at which the elasticity of labor costs with respect to capital utilization increases. Using this notation, the log-linearization of equation (45) is simply[26]

$$dz = \frac{\eta}{v} dh. \tag{46}$$

Thus, equations (46) and (38) say that the change in hours per worker should be a proxy for changes in both unobservable labor effort and the unmeasured workweek of capital. The reason that hours per worker proxies for capital utilization as well as labor effort is that shift premia create a link between capital hours and labor compensation. The shift premium is most worth paying when the marginal hourly cost of labor is high relative to its average cost, which is the time when hours per worker are also high.

Putting everything together, we have an estimating equation that controls for variable utilization

$$dy = \mu^* dx + \mu^* \left(\zeta s_L + \frac{\eta}{v} s_K \right) dh + dt. \tag{47}$$

This specification controls for both labor and capital utilization, without making special assumptions about separability or homotheticity. However, for our simple derivation, the Cobb-Douglas functional form is important. As we have noted before, our model can be generalized to allow depreciation to depend on capital utilization. This modification would introduce two new terms into the estimating equation (47), as in Basu and Kimball (1997).

26. This equation is where the Cobb-Douglas assumption matters. Basu and Kimball differentiate (35) assuming that α_L/α_K is a constant. Their theory allows for the fully general case where the ratio of the elasticities is a function of all four input quantities, but they argue that pursuing this approach would demand too much of the data and the instruments.

7.3 Aggregation over Firms[27]

Sections 7.1 and 7.2 emphasize production and estimation in terms of firm-level gross output. Section 7.1 also provided an overview of the relationship between firm-level gross output measures and aggregate value-added measures. In this section, we derive and interpret this relationship in greater detail. We aggregate in two steps. First, we relate firm-level gross output to value added. Second, we aggregate over firm-level value added. We then provide an economic interpretation of the various terms in the aggregation equation. We conclude by discussing how aggregation could spuriously generate apparent external effects across firms or industries.

The discussion highlights the potential pitfalls of using value added directly as a production measure. A long literature from the 1970s (see, for example, Bruno 1978) also argues against the use of value added, but on very different grounds. That literature emphasized the strong separability assumptions necessary for the existence of a stable value-added function. By contrast, our argument against value added does not rely on separability, which is a second-order property of the production function. Instead, we point out that value added is akin to a partial Solow residual, subtracting from gross output growth the growth in intermediates, weighted by their share in revenue. Firms equate revenue shares to output elasticities only with perfect competition. Hence, with imperfect competition, some of the productive contribution of intermediate inputs remains in measured value-added growth—a first-order issue that applies regardless of whether separability holds or fails.

7.3.1 The Conversion to Value Added

Measures of real value added attempt to subtract from gross output the productive contribution of intermediate goods. Hence, gross output is shoes, while value added is "shoes lacking leather, made without power" (Domar 1961, 716), or books without paper or ink.

Despite its unobservable and perhaps unintuitive nature at a firm level, we focus on value added for two reasons. First, discussing firm-level value added turns out to be a useful intermediate step in moving between firm-level gross output and aggregate value added. Second, many researchers use data on value added for empirical work, and the results in this section shed light on the (de)merits of that approach.

Nominal value added, $P^V V$, is defined unambiguously: $P^V V = PY - P_M M$. It is less clear how to decompose nominal value added into price and quantity: We must subtract real intermediate inputs from gross output in some way, but several methods are possible.

For the national accounting identity to hold, one must deflate nominal

27. The appendix contains detailed derivations of the equations in this section.

value added with the same method used to deflate nominal final expenditure (Sato 1976). Since the national accounts now use chain-linked indexes, we follow suit here. In particular, we use the continuous time analogue to a discrete-time chain-linked index, and define value added at a firm level, dv_i, using the standard Divisia definition[28]

$$dv_i \equiv \frac{dy_i - s_{Mi}dm_i}{1 - s_{Mi}}. \tag{48}$$

After substituting input for output growth from equation (9), we can write this as

$$dv_i = \mu_i(dx_i^V + du_i) + (\mu_i - 1)\frac{s_{Mi}}{1 - s_{Mi}}dm_i + \frac{dt_i}{1 - s_{Mi}}, \tag{49}$$

where primary input growth, dx_i^V, is defined analogously to aggregate primary input growth

$$dx_i^V = \frac{s_{Ki}}{1 - s_{Mi}}dk_i + \frac{s_{Li}}{1 - s_{Mi}}dl_i \equiv s_{Ki}^V dk_i + s_{Li}^V dl_i, \tag{50}$$

The main implication of equation (49) is that value-added growth is not, in general, simply a function of primary inputs dx_i^V. Value-added growth is calculated by subtracting from gross output the revenue-share-weighted contribution of intermediate goods. With markups, however, the output elasticity of intermediate inputs ($\mu_i s_{Mi}$) exceeds its revenue share. Hence, some of the contribution of materials and energy is attributed to value added; that is, value-added growth does not subtract off the full productive contribution of intermediate inputs. This extra productive contribution affects value-added growth.

Of course, value-added growth could still be a function of primary-input growth alone, to the extent that intermediate inputs move together with primary inputs. To explore this possibility, it is useful to rewrite equation (49) in the following way (the appendix shows the algebra)

$$dv_i = \mu_i^V dx_i^V + (\mu_i^V - 1)\left[\frac{s_{Mi}}{1 - s_{Mi}}\right](dm_i - dy_i) + \mu_i^V du_i + dt_i^V, \tag{51}$$

where

28. See Sato (1976). Until recently, the national accounts used a Laspeyres index for the expenditure side of GDP. That is, each component is valued using base-year prices, and then components are added. The national accounts identity then implied that one needed a Laspeyres or double-deflated index for real value added: $V^{DD} = Y - M$, where gross output Y and intermediate inputs M are valued in base-year prices. The chain-linked index gives cleaner results for productivity analysis, although the basic conclusions are unchanged. See the appendix to Basu and Fernald (1995).

$$\mu_i^V \equiv \mu_i \left[\frac{1 - s_{Mi}}{1 - \mu_i s_{Mi}} \right] \tag{52}$$

$$dt_i^V \equiv \frac{dt_i}{1 - \mu_i s_{Mi}}. \tag{53}$$

A useful benchmark case is where intermediate inputs are used in fixed proportions to output; this assumption is implicit in most dynamic general equilibrium models with imperfect competition (see, e.g., Rotemberg and Woodford 1995). With fixed proportions, equation (52) shows that value-added growth is, indeed, a function of primary inputs alone. The "value-added markup" μ_i^V, defined by equation (52), includes the productive contribution of primary inputs as well as the extra contribution of intermediates. Hence, if μ_i is greater than one, the value-added markup μ_i^V exceeds the gross-output markup μ_i. Nevertheless, equation (51) makes clear that with imperfect competition, if the assumption of fixed proportions fails, then value-added growth depends on the growth rate of the ratio of intermediate inputs to output.

The value-added markup is plausibly the appropriate concept for calibrating the markup charged by the representative firm in a one-sector macroeconomic model (see Rotemberg and Woodford 1995). The reason is that small markups at a plant level may translate into larger efficiency losses for the economy overall, because of "markups on markups." That is, firms buy intermediate goods from other firms at a markup, add value, and again price the resulting good with a markup, generally selling some of it to another firm to use as an intermediate good. We explore this (and other) intuition in greater detail in the appendix.

That the intensity of intermediate-input use affects value-added growth underlies our argument that the right approach to estimating even the value-added markup is to use gross-output data. Without making any auxiliary assumptions about whether materials are used in fixed coefficients, one can estimate (a utilization-corrected) μ from equation (10), and then transform it into its value-added analogue using equation (52). (Note that value added remains appropriate as a national accounting concept, because the wedge between the cost and marginal product of intermediate inputs represents actual goods and services available to society—it's just that we cannot in general allocate its production to primary inputs of capital and labor.)

The firm's revenue-weighted, value-added productivity residual, dp_i, equals $dv_i - dx_i^V$. Hence,

$$dp_i = (\mu_i^V - 1)dx_i^V + (\mu_i^V - 1) \left[\frac{s_{Mi}}{1 - s_{Mi}} \right] (dm_i - dy_i) + \mu_i^V du_i + dt_i^V. \tag{54}$$

A long literature in the 1970s explored whether a value-added function exists (see, e.g., Bruno 1978) and argued that the answer depends on separability properties of the production function. The equation above shows that with imperfect competition, taking value added to be a function only of primary inputs is generally misspecified regardless of whether the production function is separable between value added and intermediate inputs. Separability is a second-order property of a production function, so its presence or absence does not affect first-order approximations like equations (49) and (51). However, the fact that the output elasticity of materials is $\mu_i s_{Mi}$ instead of simply s_{Mi} is of first-order importance.

7.3.2 Aggregation

We define aggregate inputs as simple sums of firm-level quantities.

$$K \equiv \sum_{i=1}^{N} K_i$$

$$L \equiv \sum_{i=1}^{N} L_i$$

For simplicity, we assume that there is one type of capital and one type of labor. (This can be relaxed easily.)

In principle, different firms may face different shadow prices for a homogeneous input; this will generally be the case if some inputs are quasi-fixed. For any input J, let P_{Ji} be the shadow price it pays to rent or hire the input for one period; differences across firms in shadow prices could reflect factor-price differences or else adjustment costs, as in section 7.2. We define the aggregate (rental) prices of capital and labor as implicit deflators—that is, total factor payments divided by aggregate quantities.

$$P_K \equiv \frac{\sum_{i=1}^{N} P_{Ki} K_i}{K} \tag{55}$$

$$P_L \equiv \frac{\sum_{i=1}^{N} P_{Li} L_i}{L} \tag{56}$$

We use the standard Divisia definition of aggregate output. This measure weights goods by market prices and hence avoids substitution bias in the aggregate output and price indices. Divisia aggregates are defined most naturally in growth rates, and we denote the growth in aggregate output (equivalently, aggregate value added) by dv. From the national accounting identity, one can define aggregate output in terms of either production (aggregating value added over firms) or expenditure (aggregating sales for consumption, investment, government purchases, or export). Welfare (discussed in section 7.6) uses the expenditure side; production (emphasized

so far) relates to the value-added side. From the production side, aggregate output is a Divisia index of firm-level value added. In growth rates

$$dv = \sum_{i=1}^{N} w_i dv_i, \tag{57}$$

where w_i is the firm's share of nominal value added

$$w_i = \frac{P_i^V V_i}{P^V V}.$$

We can now substitute in from equation (51) for dv_i. Substantial algebraic manipulation (shown in the appendix) yields our basic aggregation equation

$$dv = \bar{\mu}^V dx^V + du + R + dt^V, \tag{58}$$

where

$$\begin{aligned} dx^V &= \left(\frac{P_L L}{P^V V}\right) dl + \left(\frac{P_K K}{P^V V}\right) dk \equiv s_L^V dl + s_K^V dk, \\ \bar{\mu}^V &= \sum_{i=1}^{N} w_i \mu_i^V, \\ du &= \sum_{i=1}^{N} w_i \mu_i^V du_i, \text{ and} \\ dt^V &= \sum_{i=1}^{N} w_i dt_i^V. \end{aligned} \tag{59}$$

dx^V is growth in aggregate primary inputs, $\bar{\mu}^V$ is the average firm value-added markup, du is average firm utilization growth (weighted by markups), and dt^V is average value-added technology change. R represents various reallocation effects

$$R = R_\mu + R_M + \bar{\mu}^V R_K + \bar{\mu}^V R_L, \tag{60}$$

where

$$\begin{aligned} R_\mu &= \sum_{i=1}^{N} w_i (\mu_i^V - \bar{\mu}^V) dx_i^V, \\ R_M &= \sum_{i=1}^{N} w_i (\mu_i^V - 1) \left[\frac{s_{Mi}}{1 - s_{Mi}}\right] (dm_i - dy_i), \\ R_K &= \sum_{i=1}^{N} w_i s_{Ki}^V \left[\frac{P_{Ki} - P_K}{P_{Ki}}\right] dk_i, \text{and} \\ R_L &= \sum_{i=1}^{N} w_i s_{Li}^V \left[\frac{P_{Li} - P_L}{P_{Li}}\right] dl. \end{aligned} \tag{61}$$

Note that aggregate utilization du from equation (59) is the weighted

average of firm-level value-added utilization. It is a value-added measure, since it is a form of primary input, albeit one that is unobserved. Since it is multiplied by value-added returns to scale, it captures the full effect on aggregate output of a change in utilization, incorporating both the contribution of a change in the unobserved *quantity* of input, and the contribution of the markup. (As with primary inputs, we could have separated this into a term reflecting the "average" markup and a "reallocation of the markup" term, but since our main interest in utilization is on the total effect, it is simpler to keep them together.)

Aggregate productivity growth is the difference between the growth rates of aggregate output, dv, and aggregate inputs, dx^V

$$dp \equiv dv - dx^V. \tag{62}$$

Note that this is a modified Solow residual, since the input weights in dx^V need not sum to one. The shares are factor payments in total nominal value added, and sum to one only if there are no economic profits. Thus

$$dp = (\bar{\mu}^V - 1)dx^V + du + R + dt^V. \tag{63}$$

Equation (63) shows the distinction between aggregate productivity and aggregate technology. If all firms are perfectly competitive, pay the same price for factors (perhaps reflecting perfect factor mobility), and do not vary utilization, then all terms other than dt^V disappear: Productivity equals technology. However, with imperfect competition or frictions in product or factor markets, productivity and technology are not equivalent.[29]

7.3.3 Productivity Interpretation of Reallocation Terms

Aggregate output combines goods using market prices, which in turn measure consumers' relative valuations. Suppose we want to know how much consumers (and, therefore, usually society) value having a marginal input allocated to a particular firm (which we assume produces a single good). That valuation equals the good's price times the factor's marginal product: $P_i F_J^i$. The first-order condition in equation (7) shows that this valuation equals the firm's markup times the input price paid by the firm: $\mu_i P_{Ji}$. If markups or input prices differ across firms, then the marginal "social value" of inputs also differs across firms. R_μ, R_K, and R_L reflect shifts of resources among uses with different marginal social values.

Consider the markup-reallocation term, R_μ. By definition, the markup represents the wedge between the value consumers place on the good—that is, its price—and its marginal cost. Reallocating resources from low-

29. Jorgenson, Gollop, and Fraumeni (1987) derive an equation for the case of case of constant returns to scale and perfect competition, so they omit the terms other than R_K and R_L. They also allow for heterogeneity in capital and labor, which we have ignored for simplicity. With heterogeneity, our results generalize easily: For example, if R_{Kk} is the factor-price reallocation term for capital of type k, then $R_K = \Sigma_k R_{Kk}$.

to high-markup firms thus shifts resources towards uses where consumers value them more highly. If the variability of firms' inputs are correlated with their market power, then imperfect competition affects aggregate productivity even if the average markup is small. For example, Basu and Fernald (1997a) estimate that durable goods industries have larger returns to scale and markups than nondurable goods industries. Durable industries are more cyclical, and employ a larger share of the marginal inputs in a boom. This marginal reallocation thus contributes to the procyclicality of aggregate productivity.

Now consider the input-reallocation terms R_K and R_L. Shifting labor from firms where it has a low shadow value to firms where it has a high shadow value increases aggregate output. Why might shadow values (or wages) differ across firms? First, labor may not be instantaneously mobile across sectors; sectoral shifts may lead workers in, say, defense industries to have lower marginal products than they would in health-care.[30] Second, efficiency wage considerations may be more important in some industries than others, as emphasized by Katz and Summers (1989). Third, unions with monopoly power might choose to charge different wages to different firms. Whatever the reason, shifting labor to more productive uses increases aggregate output, even if total input does not change.

Note that the first reason, costly factor mobility, is completely consistent with constant returns and perfect competition. Differences in marginal products that reflect factor immobility should be temporary—that is, P_K as defined in equation (25) should not differ persistently from the Hall-Jorgenson cost of capital, and P_L as defined in equation (26) should not differ persistently from the firm's compensation payments (see Berndt and Fuss 1986). With costly factor adjustment, these shadow values may differ substantially; hence reallocation effects on output and productivity may be significant, even in a world with perfect competition and constant returns.[31]

The materials-reallocation term, R_M, reflects the extent to which measured real value added depends on the intensity of intermediate-input use. Firm-level value added is useful for national accounting, regardless of technology or market structure. However, with imperfect competition, value-added growth does not subtract off the full marginal product of intermediate inputs. Growth in primary inputs captures some of this productive contribution (which is why μ_i^V differs from μ_i), but some wedge may

30. Whether differences in labor's marginal product lead to differences in wages depends on whether the adjustment costs are paid by workers or firms.

31. The "sectoral shifts" literature takes this approach; see, for example, Phelan and Trejos (1996). Horvath (1995) also incorporates adjustment costs into a dynamic general equilibrium model, generating effects on aggregate productivity from input reallocations. Microeconomic productivity literature (e.g., Baily, Hulten, and Campbell 1992) finds that there are systematic productivity level differences across firms within narrowly defined industries; to the extent these productivity differences are not measurement error, they show up either as higher profits from a higher markup, and hence are reflected in R_μ, or higher factor payments, and hence are reflected in R_K and R_L.

remain. R_M equals the sum of these wedges. It represents real goods and services, and hence affects aggregate output and productivity.

Note that R_M depends on the size of markups in firms *using* materials. Consider an economy where some firms produce intermediate goods using capital and labor, and other firms assemble intermediate goods into final goods (e.g., Beaudry and Devereux 1994). The importance of R_M depends on the size (and heterogeneity) of markups in the *final goods* industry. This is important because firms may be able to negotiate multi-part prices with long-term suppliers of their inputs, and thus partially offset the inefficiencies resulting from imperfect competition in intermediate goods industries.[32] The inefficiency in R_M, however, depends on markups in firms *using* intermediate goods, not those *selling* such goods; multipart pricing for intermediate goods does not eliminate this inefficiency. (However, the inefficiency is larger in symmetric models, such as Basu (1995), where all output is also used as materials input.)

7.3.4 The Definition of Technology Change

Conceptually, aggregate technology change measures the change in aggregate output in response to firm-level technology shocks, holding primary inputs fixed. Under what conditions does this correspond to our measure dt^V?

With constant returns and perfect competition, Domar (1961) and Hulten (1978) show that our definition properly measures the outward shift in society's production possibilities frontier (PPF) when firm-level technology changes. In figure 7.1 this case corresponds to point A, where society allocates resources optimally.

In this case, all of the μ_i equal 1, so this "Domar weighted measure" of aggregate technology equals

$$dt^V = \sum_i w_i \frac{dt_i}{1 - s_{Mi}}. \tag{64}$$

Conceptually, Domar weighting converts gross-output technology shocks to a value-added basis by dividing through by the value-added share, $-s_M$. These shocks are then weighted by the firm's value-added share.

Even with imperfect competition, the Domar weighted measure correctly shows how much final output (the sum of value added over firms) increases, if all of the increase in gross output goes to final sales, with primary and intermediate inputs remaining unchanged. With perfect competition and perfect factor mobility, Domar weighting is correct even if inputs adjust. That is, the consequences for aggregate output and technology are the same whether the firm sells all of its additional output for

32. We thank Robert Hall for this observation.

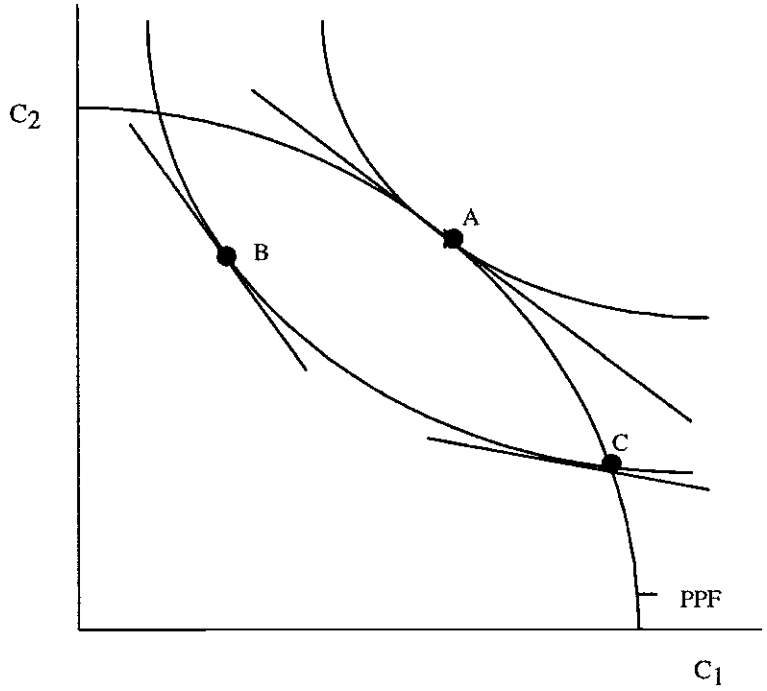

Fig. 7.1 Productivity and welfare

final consumption (with a valuation equal to its price), or instead sells the additional output to other producers, who in turn use the additional intermediate inputs (with a marginal product equated to the price) to produce goods for final consumption.

With imperfect competition, however, aggregate technology change is not unambiguously defined. The economy may produce inside its PPF, at a point like B in figure 7.1 or at an allocatively inefficient point on the PPF, like point C. At these points, the same firm-level technology shocks can affect the distribution of inputs across firms in different ways, and thus cause different changes in aggregate output. Imperfect competition or differences in factor payments across firms may lead the same factor to have a different social value for its marginal product in different uses. Hence, changes in the distribution of resources can affect aggregate output, even if there are no technology shocks and total inputs remain the same.

The definition in equation (59) is essentially a markup corrected measure. In terms of the gross-output shocks, it is equivalent to the following:

$$dt^V = \sum_i w_i \frac{dt_i}{1 - \mu_i s_{Mi}} \tag{65}$$

This definition correctly measures the increase in aggregate output under standard conditions where an aggregate production function exists. Suppose there are no factor market frictions, and that all firms have the same separable gross-output production function, always use materials in fixed proportions to gross output, and charge the same markup of price over marginal cost. Under these assumptions, which are implicit or explicit in most dynamic general-equilibrium models with imperfect competition, there is a representative producer and an aggregate value-added production function (see, e.g., Rotemberg and Woodford 1995). In this case, aggregate technology change corresponds to our definition dt^V.

Qualitatively, the Domar weighted and markup corrected measures turn out to give very similar correlations between technology change and various business-cycle indicators. (One empirical note of caution is that concern the markup corrected measure is sometimes very sensitive to estimates of the markup. When we estimate μ_i for each sector, we hope that our estimates are unbiased. However, the estimate of value-added technology change in equation (53), $dt_i/(1 - \mu_i s_{Mi})$, is convex in μ_i. Hence, it overweights sectors where $\hat{\mu}_i$ exceeds μ_i, and underweights sectors where $\hat{\mu}_i$ is less than μ_i.)

7.3.5 Externalities

So far, we have discussed internal increasing returns but not Marshallian externalities. With such external effects, the economy might display aggregate increasing returns even if all firms produce with constant returns. Hence, one can model increasing returns without having to model imperfect competition. As a result, models of growth and business cycles use such externalities extensively (e.g., Romer 1986; Baxter and King 1991; Benhabib and Farmer 1994). Externalities are almost surely important for modeling economic growth, as suggested by the extensive R&D-spillover literature surveyed by Griliches (1992). High-frequency *demand* spillovers (as discussed in Cooper and John 1988) also seem eminently sensible. However, apart from Diamond (1982), few have proposed models of short-run *technological* spillovers that operate at high frequencies.

Caballero and Lyons (1992) argued, however, that a basic prediction of externality models is a robust feature of the data: Estimated returns to scale should rise at higher levels of aggregation, as the increasing returns become internalized. In other words, aggregate productivity should be more procyclical than would be implied by estimates of industry-level returns to scale. Thus, despite lacking a formal model of short-run external effects, they concluded that there is strong prima facie evidence for such externalities.

Caballero and Lyons (1990, 1992) proposed an empirical model to estimate any short-run externalities. They augmented a firm-level estimating equation like equation (10) with aggregate inputs as well as firm-level in-

puts. In practice, they used industry-level data on output and inputs, and added aggregate output to the estimating equation to capture any externalities that would otherwise be relegated to the error term, *dt*. They used NIPA data on real value added by industry. Thus, their basic estimating equation was essentially

$$dv_i = \mu_i^V dx_i^V + \beta dx^V + dt_i, \tag{66}$$

where *i* indexed a two-digit manufacturing industry, and dx^V was growth of aggregate capital and labor in manufacturing. (Their 1992 paper also included various utilization controls.) They found large, positive, and statistically-significant values of β in data from the United States and a number of European countries.

Numerous authors have questioned the interpretation and robustness of the Caballero-Lyons's results. For example, their results may reflect inappropriate data (Basu and Fernald 1995; Griliches and Klette 1996), incorrect econometric method (Burnside 1996), or inadequate utilization proxies (Sbordone 1997).

Nevertheless, even if one can dismiss their interpretation, the stylized fact remains that productivity is more procyclical at higher levels of aggregation. However, equation (63) suggests an alternative explanation of the Caballero-Lyons stylized fact. Aggregate productivity is more procyclical because the reallocation terms, *R*, are procyclical. Thus, we can explain the Caballero-Lyons stylized fact based only on firm-level heterogeneity, without invoking external effects that have questionable theoretical basis in the business cycle context. Indeed, we note that the one well-known economic model of short-run externalities, Diamond's (1982) search model, relies on increasing returns to scale in the "matching function" that produces new hires as a function of economy-wide vacancies and unemployment. Efforts to estimate this matching function directly (e.g., Blanchard and Diamond 1990), however, show that it exhibits approximately constant returns to scale, not large increasing returns. Thus, in the absence of any direct evidence for *short-run* externalities, we do not model them explicitly in this paper.[33]

7.4 Data and Method

7.4.1 Data

We now construct a measure of "true" aggregate technology change and explore its properties. As discussed in section 7.1, we estimate technology

33. The search for short-run spillovers remains an active area of ongoing research. See, for example, Bartelsman, Caballero, and Lyons (1994), Cooper and Johri (1997), and Paul and Siegel (1999).

change at a disaggregated level, and then aggregate using the theory in section 7.3. Our aggregate is the private U.S. economy, and our "firms" are thirty-three industries; for manufacturing, these industries correspond roughly to the two-digit Standard Industrial Classification (SIC) level.

Given that each industry includes thousands of firms, it may seem odd to take industries as firms. Unfortunately, no firm-level data sets span the economy. In principle, we could focus on a subset of the economy, using, say, the Longitudinal Research Database. However, narrowing the focus requires sacrificing a macroeconomic perspective, as well as panel length and data quality. By focusing on aggregates, our paper complements existing work that uses small subsets of the economy.[34]

We use data compiled by Dale Jorgenson and Barbara Fraumeni on industry-level inputs and outputs. These data comprise gross output and inputs of capital, labor, energy, and materials for a panel of thirty-three private industries (including twenty-one manufacturing industries) that cover the entire U.S. nonfarm private economy. These sectoral accounts seek to provide accounts that are, to the extent possible, consistent with the economic theory of production. (For a complete description of the dataset, see Jorgenson, Gollop, and Fraumeni 1987.) These data are available from 1947 to 1989; in our empirical work, however, we restrict our sample to 1950 to 1989, since our money shock instrument (described below) is not available for previous years.

We weight growth rates (measured as log changes) of capital, labor, and intermediate inputs using the *average* shares in revenue over the entire period. To compute capital's share, s_K, for each industry, we construct a series for required payments to capital. Following Hall and Jorgenson (1967) and Hall (1990), we estimate the user cost of capital C. For any type of capital, the required payment is then $CP_K K$, where $P_K K$ is the current-dollar value of the stock of this type of capital. In each sector, we use data on the current value of the fifty-one types of capital, plus land and inventories, distinguished by the BEA in constructing the national product accounts. Hence, for each of these fifty-three assets, indexed by s, the user cost of capital is

$$C_s = (r + \delta_s)\frac{(1 - ITC_s - \tau d_s)}{(1 - \tau)}, \quad s = 1 \text{ to } 53. \tag{67}$$

r is the required rate of return on capital, and δ_s is the depreciation rate for assets of type s. ITC_s is the asset-specific investment tax credit, τ is the corporate tax rate, and d_s is the asset-specific present value of depreciation allowances. We follow Hall (1990) in assuming that the real required re-

34. See, for example, Baily et al. (1992), Haltiwanger (1997), Bartelsman and Dhrymes (1998), and Foster, Haltiwanger, and Krizan (this vol.). The aggregation theory in section 7.3 implies that our industry data include various intra-industry reallocation terms, including the analogous terms to R_μ, R_K, and R_L.

turn r equals the dividend yield on the S&P 500. Jorgenson and Yun (1991) provide data on ITC_s and d_s for each type of capital good. Given required payments to capital, computing s_K is straightforward.

As discussed in section 7.1, we require instruments uncorrelated with technology change. We use two of the Hall-Ramey instruments: the growth rate of the price of oil deflated by the GDP deflator and the growth rate of real government defense spending.[35] (We use the contemporaneous value and one lag of each instrument.) We also use a version of the instrument used by Burnside (1996): quarterly Federal Reserve "policy shocks" from an identified Vector Autoregression. We sum the four quarterly policy shocks in year $t - 1$ as instruments for year t.[36]

7.4.2 Estimating Technology Change

To estimate firm-level technology change, we estimate a version of equation (10) for each industry. Although we could estimate these equations separately for each industry (and indeed do so as a check on results), some parameters—particularly on the utilization proxies—are then estimated rather imprecisely. To mitigate this problem, we combine industries into four groups, estimating equations that restrict the utilization parameters to be constant within industry groups.

As discussed in section 7.2, this estimating equation corresponds to the special case where the cost of higher capital utilization is a shift premium. In that case, variations in hours per worker fully captures variations in capital utilization and effort. Thus, for each group we have

$$dy_i = c_i + \mu_i dx_i + adh_i + dt_i. \tag{68}$$

The markup μ_i differs by industries within a group (Burnside 1996 argues for allowing this variation). The groups are durables manufacturing (eleven industries); nondurables manufacturing (ten); mining and petroleum extraction (four); and all others, mainly services and utilities (eight). To avoid the transmission problem of correlation between technology

35. We drop the third instrument, the political party of the President, because it appears to have little relevance in any industry. Burnside (1996) shows that the oil price instrument is generally quite relevant, and defense spending explains a sizeable fraction of input changes in the durable-goods industries.

36. The qualitative features of the results in section 7.3 appear robust to using different combinations and lags of the instruments. On a priori grounds, the set we choose seems preferable to alternatives—all of the variables have strong grounds for being included. In addition, the set we choose has the best overall fit (measured by mean and median F statistic) of the a priori plausible combinations we considered. Of course, Hall, Rudebusch, and Wilcox (1996) argue that with weak instruments, one does not necessarily want to choose the instruments that happen to fit best in sample; for example, if the "true" relevance of all the instruments is equal, the ones that by chance fit best in sample are in fact those with the largest small sample bias. That case is probably not a major concern here, since the instrument set we choose fits well for all industry groupings; for example, it is the one we would choose based on a rule of, say, using the instruments that fit best in durables industries as instruments for nondurables industries, and vice versa.

shocks and input use, we estimate each system using Three-Stage Least Squares, using the instruments noted above.

After estimating equation (68), we take the sum of the industry-specific constant c_i and residual $d\hat{t}_t$ as our measure of technology change in the gross-output production function. We then insert these industry estimates in the aggregation equation (63), derived in section 7.3. Note that this aggregation equation is an accounting identity. It allows us to decompose aggregate productivity into a technological component plus various non-technological components, including the effects of markups and various reallocations.

One problem in implementing the decomposition is that we may not, in fact, observe period-by-period the appropriate "shadow" factor prices P_{Li} and P_{Ki} defined by equations (25) and (26). We deal with this problem by taking an explicitly first-order approach to the estimating equation in equation (68), using fixed weights on capital and labor in constructing dx_i. (It is straightforward, though it requires some care, to ensure that the aggregation equation (63) remains an accounting identity with fixed weights.) This approach is unlikely to lead to major problems in estimating the markup and materials reallocation terms (R_μ and R_M) in the accounting identity, since those terms are driven primarily by changes in *quantities,* rather than changes in the weights (which, in turn, incorporate the shadow prices).

However, the inability to measure prices is a major problem for measuring the input reallocation terms (R_K and R_L), as those terms depend explicitly on differences in prices across sectors. Jorgenson, Gollop, and Fraumeni (1987) estimate these terms under the assumption that factor prices are allocative, and that there are zero profits in all sectors (so that they can "back out" capital's input P_K as a residual). Although we do not require these assumptions elsewhere, we will show summary statistics from Jorgenson and colleagues' estimates—we emphasize that these are meant to be suggestive only. Since the aggregation equation (63) is an accounting identity, changing our estimate of one component requires a change in other components as well. We then remove the effect of the input-reallocation terms from the average-markup term (a natural place to take it from, given appendix equation [A.17]). It is worth emphasizing that even if we mismeasure the input reallocation terms, this primarily affects our measurement of *other* nontechnological components of aggregate productivity. In particular, it does not directly affect our estimate of aggregate technology, which we built up from disaggregated residuals.

7.5 Results

We now investigate empirically why productivity is procyclical. We seek to identify the importance of (1) imperfect competition, (2) reallocations,

Table 7.1 Descriptive Statistics for Technology Residuals

	Mean	Standard Deviation	Minimum	Maximum
		A. Private Economy		
Solow residual	0.011	0.022	−0.044	0.066
Technology residual (no utilization correction)	0.012	0.016	−0.034	0.050
Technology residual (hours corrected)	0.013	0.013	−0.013	0.042
		B. Manufacturing		
Solow residual	0.023	0.035	−0.081	0.080
Technology residual (no utilization correction)	0.014	0.030	−0.085	0.072
Technology residual (hours corrected)	0.018	0.028	−0.030	0.082

Note: Sample period is 1950–89. The Solow residual is calculated using aggregate data alone. The two technology residuals are calculated by aggregating residuals from sectoral regressions of gross-output growth on input growth, as described in the text. The "hours corrected" residual corrects for variable utilization by including growth in hours per worker as a separate explanatory variable, in line with the theory developed in section 7.2.

and (3) variable utilization. Controlling for these influences allows us to move from aggregate productivity to aggregate technology. We then explore the cyclical properties of the "corrected" technology series.

We define aggregate productivity growth as the modified Solow residual defined in equation (62). This measure differs from the standard Solow residual since the revenue weights need not sum to one; the difference reflects economic profits or losses. However, we estimate that profits are small (about 3 percent on average, using our estimates of required payments to capital), so the results we report are essentially unchanged using the standard Solow residual instead.

Table 7.1 reports summary statistics for three series: the Solow residual; a series that makes no utilization corrections, but corrects only for aggregation biases; and a "technology" measure based on equation (68), which uses growth in hours per worker to correct for utilization. The first measure uses aggregate data alone. The other two are based on sectoral Solow-Hall–style regression residuals, as described in the previous section; these residuals are then aggregated using equation (63). Hence, aggregate technology change is the weighted sum of sectoral technology changes, as described in sections 7.1 and 7.3 (see equations [53] and [59]).

Panel A shows results for the entire nonfarm business economy. Our corrected series have about the same mean as the Solow residual. However, the variance is substantially smaller: The variance of the fully corrected series is less than one-third that of the Solow residual, so the standard deviation (shown in the second column) is only about 55 percent as large.

The reported minimums show that we do estimate negative technical change in some periods, but the lower variance of the technology series implies that the probability of negative estimates is much lower. For example, the Solow residual is negative in twelve out of forty years; the fully-corrected residual is negative in only five out of forty years.

Panel B gives results within manufacturing alone. Data within manufacturing (especially for output) are often considered more reliable than data outside manufacturing. In addition, some other papers (such as Burnside 1996) focus only on manufacturing, so these results provide a basis for comparison. The results are qualitatively similar to those for the aggregate economy.

Some simple plots and correlations summarize the comovement in our data: Output and inputs are strongly positively correlated, and all are positively correlated with the Solow residual. Figure 7.2 plots business cycle data for the nonfarm private economy: output (value-added) growth dv, primary input growth, dx^V, and the Solow residual dp (all series are demeaned). These series comove positively, quite strongly so in the case of dp and dv. Table 7.2 shows correlations for these three variables, as well as growth of total hours worked ($dh + dn$). Hours correlate more strongly with productivity than do total inputs, reflecting the low correlation of changes in the capital stock with the business cycle. The 95 percent confidence intervals show that all are significant.

Figure 7.3 plots our fully corrected technology series against these three variables. The top panel shows that technology fluctuates much less than

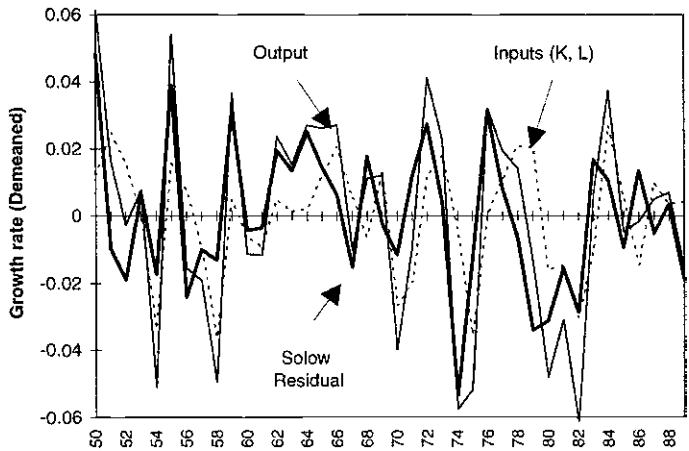

Fig. 7.2 Aggregate Solow residual, input growth, and output growth
Note: All series are demeaned. Sample period is 1950–89. Data is from Jorgenson, Gollop, and Fraumeni (1987). Inputs are a share weighted average of capital and labor growth.

Table 7.2 **Basic Data Correlations**

	Output Growth (dv)	Input Growth (dx^V)	Hours Growth ($dh+dn$)	Solow Residual
	A. Private Economy			
Output growth (dv)	1			
Input growth (dx^V)	0.78	1		
	(0.62, 0.88)			
Hours growth ($dh+dn$)	0.80	0.91	1	
	(0.64, 0.89)	(0.83, 0.92)		
Solow residual	0.84	0.33	0.44	1
	(0.72, 0.91)	(0.02, 0.59)	(0.15, 0.66)	
	B. Manufacturing			
Output growth (dv)	1			
Input growth (dx^V)	0.81	1		
	(0.66, 0.90)			
Hours growth ($dh+dn$)	0.86	0.98	1	
	(0.75, 0.92)	(0.96, 0.99)		
Solow residual	0.84	0.36	0.46	1
	(0.71, 0.91)	(0.05, 0.61)	(0.17, 0.68)	

Notes: 95 percent confidence intervals in parentheses, calculated using Fisher transformation. Sample period is 1950–89.

the Solow residual, consistent with intuition that nontechnological factors, such as variable input utilization, increase the volatility of the Solow residual. In addition, some periods show a phase shift: The Solow residual lags technology change by one to two years. This phase shift reflects the utilization correction. In our estimates, technology improvements are associated with low levels of utilization, thereby reducing the Solow residual relative to the technology series. The phase shift, in particular, appears to reflect primarily movements in hours per worker, which generally increase a year after a technology improvement. In the model from section 7.2, increases in hours per worker imply increases in unobserved effort, which in turn increase the Solow residual.

The middle panel plots aggregate value-added output growth (dv) against technology. There is no clear contemporaneous comovement between the two series although, again, the series appear to have a phase shift: Output comoves with technology, lagged one to two years.

Finally, the bottom panel plots the growth rate of primary inputs of capital and labor (dx^V) and the same technology series. These two series clearly comove negatively over the entire sample period.

Table 7.3 shows the correlations between our technology measures and business cycle variables. Panel A shows results for the aggregate private economy. With full corrections, the correlation of technology with output is about zero, and the correlations with inputs are strongly negative: -0.42

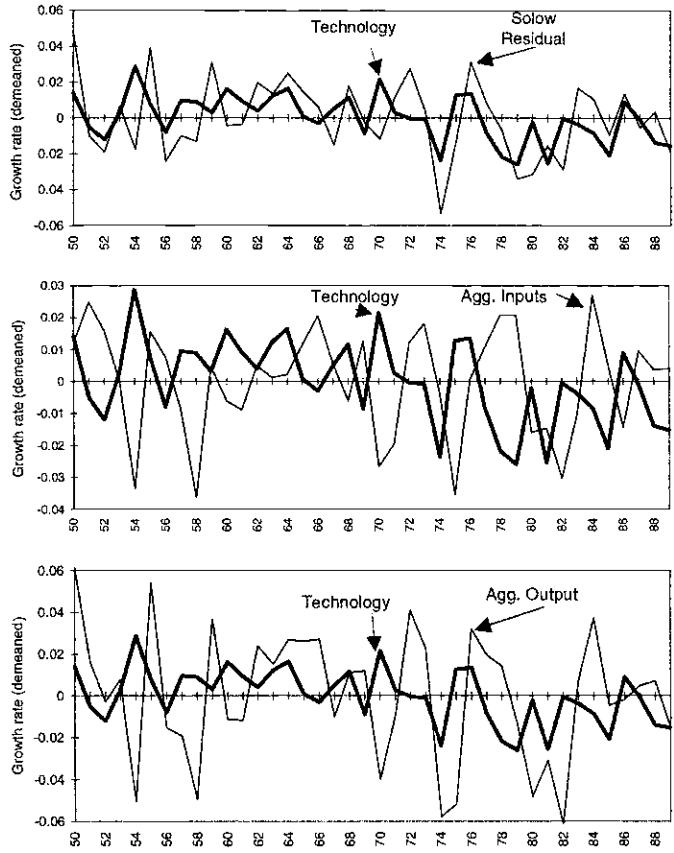

Fig. 7.3 Technology residual, Solow residual, output and input growth
Note: The technology series is the hours adjusted aggregate residual, which measures technology change (adjusted for variations in utilization) for the nonfarm business economy. Aggregate inputs are a share weighted average of capital and labor growth. All series are demeaned. Entries are log changes. Sample period is 1950–89.

for total primary inputs, and −0.44 for hours alone. Both correlations are statistically significantly negative at the 95 percent level.

The correlations with aggregate technology change differ sharply from those predicted by the usual RBC model (e.g., Cooley and Prescott 1995). In particular, in calibrated dynamic general equilibrium models with flexible prices, technology shocks generally cause a contemporaneous increase in both inputs and output. These kinds of standard, real business cycle models explore whether technology shocks lead to comovement that matches the stylized facts of business cycles. Given that the central stylized fact of business cycles is the comovement between inputs and output, if technology shocks drive the cycle then almost any sensible calibration implies that technology improvements increase inputs and output.

Table 7.3 **Correlations of Technology Residuals with Basic Data**

	Output Growth (dv)	Input Growth (dx^V)	Hours Growth ($dh+dn$)	Solow Residual
		A. Private Economy		
Technology residual (no utilization correction)	0.46 (0.17, 0.68)	−0.12 (−0.41, 0.21)	−0.06 (−0.37, 0.26)	0.77 (0.63, 0.88)
Technology residual (hours corrected)	0.04 (−0.28, 0.35)	−0.42 (−0.65, −0.12)	−0.44 (−0.66, −0.14)	0.40 (0.10, 0.64)
		B. Manufacturing		
Technology residual (no utilization correction)	0.42 (0.12, 0.65)	−0.14 (−0.44, 0.18)	−0.04 (−0.35, 0.28)	0.79 (0.63, 0.89)
Technology residual (hours corrected)	−0.40 (−0.64, 0.10)	−0.64 (−0.80, −0.41)	−0.62 (−0.78, −0.38)	−0.05 (−0.36, 0.27)

Notes: 95 percent confidence intervals in parentheses, calculated using Fisher transformation. Sample period is 1950–89. Technology residuals are calculated by aggregating residuals from sectoral regressions of gross-output growth on input growth, as described in the text. The "hours corrected" residual corrects for variable utilization by including growth in hours per worker as a separate explanatory variable.

Basu, Fernald, and Kimball (1999) explore the negative contemporary comovement between technology and input growth at length. They argue that this negative comovement is consistent with sticky price models. For example, suppose a firm's technology improves but the firm cannot change its price. If its demand curve does not change, then it cannot change its sales—but it can produce that same quantity with fewer inputs. Over time, of course, the firm (and economy) adjust to the technology change.

In terms of our accounting identities, what explains the movement away from a strong positive correlation? Figure 7.4 shows the our estimated utilization series—aggregated from the implicit utilization series for each industry using the equation for *du* from equation (59)—and our estimated reallocation series, both plotted against the Solow residual. Both utilization and reallocations are procyclical, as shown here by the positive comovement with the Solow residual. Each contributes about equally to generating the negative correlation. To see this, we first subtracted the

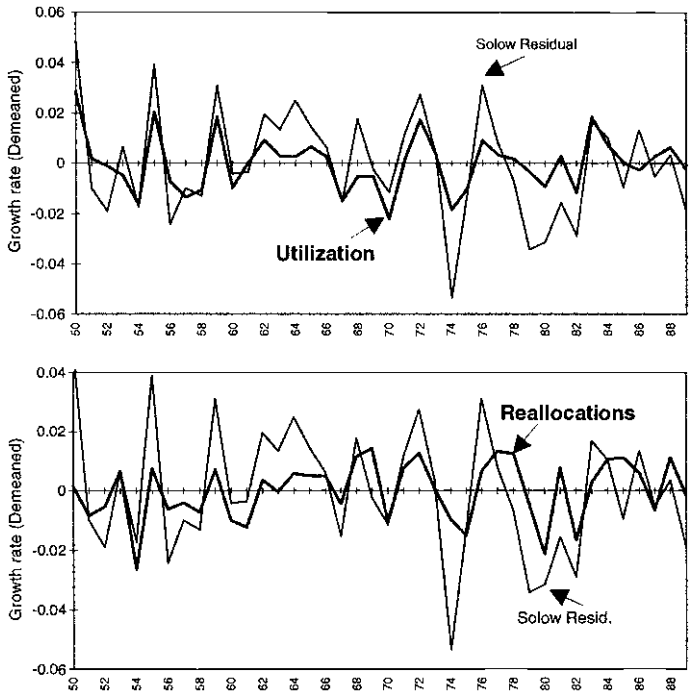

Fig. 7.4 Nontechnological adjustments to the Solow residual

Note: Aggregate utilization growth is a weighted average of estimated industry-level utilization growth (which includes the estimated markup for the industry). Estimated reallocations are the sum of R_{μ} (reallocations of inputs among industries with different markup estimates), R_M (reallocations of intermediate inputs), and R_K and R_L (reallocations of capital and labor among uses with different factor prices).

Table 7.4 **Reallocation by Component, 1959–89 (instrumental variables estimates)**

	R	R_μ	R_M	$\bar{\mu}^V R_K + \bar{\mu}^V R_L$
Mean	−0.11	−0.17	0.02	−0.03
Standard deviation	1.04	0.71	0.47	0.24
Minimum	−2.86	−2.05	−1.09	−0.51
Maximum	1.36	1.08	0.74	0.64

Notes: Entries are percentage points per year. R is the sum of the components shown in the columns to the right. R_μ is reallocations of inputs among industries with different markup estimates. R_M is reallocations of intermediate inputs. R_K and R_L are reallocations of capital and labor among uses with different factor prices, as calculated by Jorgenson and Fraumeni. See the text for further details.

estimated utilization change from the Solow residual; the correlation with inputs fell to about zero. Similarly, we then subtracted the estimated reallocation terms from the Solow residual; again, the correlation with inputs fell to about zero.

It is not surprising that utilization is procyclical. After all, utilization is a form of primary input, and inputs are procyclical. It is less obvious why reallocations are procyclical, as research at a highly disaggregated level—for example, using firms within a narrowly defined industry—finds that reallocations tend to be countercyclical.[37] Low productivity firms enter and expand disproportionately in booms, and contract or disappear disproportionately in recessions. Hence, within narrowly defined industries, reallocations appear to make aggregate productivity *less* cyclical, not more. (Because we use much more aggregated data on industries, these intra-industry reallocations will tend to appear as decreasing returns to scale or—given the close link between returns to scale and markups discussed in equation (5) of section 7.1—markups less than one.)

A different process clearly must be at work across aggregated industries. Table 7.4 presents summary statistics for the components of the reallocation terms from equation (61). (As discussed in the previous section, we obtained estimates of the capital and labor reallocation terms from Dale Jorgenson and Barbara Fraumeni.) None of these components has a sizeable mean, but they do have substantial standard deviations. R_μ and R_M are the most important components of the reallocation term. For the instrumented series, these terms have standard deviations of 0.72 and 0.51 respectively, compared with a total standard deviation of 1.11 for R. The sum of R_K and R_L is much less volatile. R_μ, in particular, comoves strongly with inputs, with a correlation of 0.7 with aggregate dx.

Why is R_μ so important over the cycle? Its variation reflects the fact that high-markup sectors tend to be more cyclical than average. Hence, in a boom, high-markup (and hence, high marginal product) firms produce a

37. See, for example, Foster, Haltiwanger, and Krizan, chapter 8 in this volume.

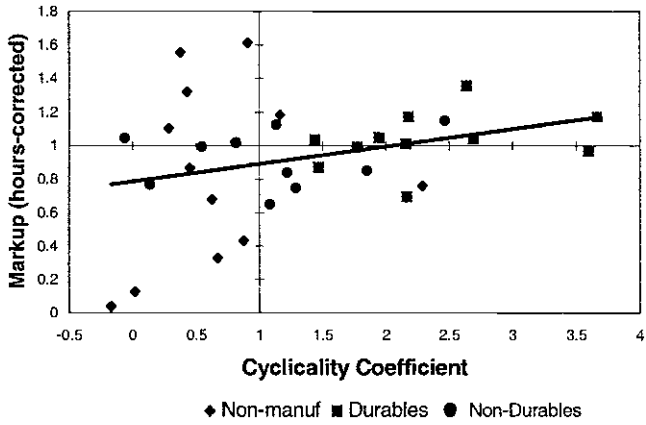

Fig. 7.5 Markups and cyclicality

Note: The horizontal axis shows the utilization corrected markup estimate by industry. The vertical axis shows the estimated cyclicality coefficient for the industry, estimated by regressing industry growth in primary inputs on aggregate growth in primary inputs.

disproportionate share of the marginal output. We can see this in figure 7.5, which shows estimates of the gross-output markup by industry, plotted against the relative cyclicality of the industry.[38] (The cyclicality was estimated by regressing $dx_i^V - dx^V$ on dx^V.) High markup industries tend to be durables manufacturing industries, which also tend to be the most cyclical. In a simulated DGE model, Basu, Fernald, and Horvath (1996) find that reallocations between durable and non-durable producers provide a significant propagation mechanism for shocks to the economy.

One might ask whether the reallocation effects we have identified represent the important gap between (utilization corrected) productivity and technology, or whether the difference between the two is still driven mostly by the "average markup" effect identified by Hall (1988, 1990). We compared our results to what we would have found had we used only the "average" correction, that is, that coming from the presence of the $\overline{\mu}^V - 1)dx^V$ term. Using the correlations with dx^V as a benchmark, we found that the reallocation effects are the more important. For the OLS results, productivity corrected for the average effect would have yielded a correlation with input growth of 0.18, as opposed to 0.23 for the uncorrected series and 0.05 for our estimated technology series. For the section 7.4 results, the correlation would have been 0.12, as opposed to -0.13 for the estimated series. Thus, our reallocation effects are responsible for at least two-thirds

38. The mean markup in figure 7.5 is less than one. As Basu and Fernald (1997a) discuss, the average of sectoral estimates of gross-output markups is less than one, even though the average of value-added markups tends to be slightly greater than one. This primarily reflects the difference in weights used to aggregate the two.

of the correction. This result should not be surprising since, as we noted, the recent literature finds small average markups. Our results echo this finding: Without any utilization correction, we found $\overline{\mu}^V = 1.12$, for the hours-corrected case we found $\overline{\mu}^V = 1.05$. The surprising result is that even such small average markups are consistent with important differences between aggregate productivity and aggregate technology coming from reallocations across sectors.

7.6 Implications for Macroeconomics

7.6.1 Productivity as Welfare

Why is aggregate productivity growth interesting? The usual justification is Solow's (1957) proof that with constant returns to scale, perfect competition, and no frictions, it measures aggregate technology change. But in a world with distortions, is the Solow residual merely mismeasured technology?

In this section we summarize the argument of Basu and Fernald (1997b), who suggest that the answer is no. They show that productivity growth correctly computed from aggregate data (i.e., after eliminating the mismeasurement caused by changes in utilization) has a natural welfare interpretation, whether it also measures technology change. In particular, the modified aggregate Solow residual ($dp - dx^V$) defined in section 7.3—which reduces to Solow's residual if there are no economic profits—measures welfare change for a representative consumer. This result holds even with imperfect competition in output markets and nonconstant returns to scale in production. Intuitively, growth in aggregate output measures the growth in society's ability to consume. To measure welfare change, we must then subtract the opportunity cost of the inputs used to produce this output growth. Input prices measure that cost, regardless of whether they also reflect marginal products. For example, increasing the stock of capital requires foregoing consumption, and the rental price of capital measures the consumer's opportunity cost of providing new capital, just as the wage measures the opportunity cost of providing extra labor. Proving our welfare result requires simply that consumers take prices parametrically. Hence, if productivity and technology differ, then it is productivity that most closely indexes welfare.

This conclusion is appealing for two reasons. First, it shows that productivity rather than, say, GDP, is the right measure of economic welfare under fairly general conditions. Second, it shows that even with distortions, policymakers can compute interesting quantities from aggregate data—we do not always need to calculate firm-level or aggregate technology change. In the short to medium run, productivity can change for reasons unrelated to technology change. Thus, even with distortions such as imperfect com-

petition, when the aggregate Solow residual does not in general index technology change, it remains an excellent index of welfare change. Hence, it remains an appropriate target for policy, as well as a convenient indicator.

In section 7.3, we showed that aggregate technology change needs to be measured using disaggregated—ideally, firm-level—data. So why do aggregate data yield a meaningful measure of welfare change? The welfare properties of the Solow residual follow from the equality of relative market prices to the consumer's marginal rates of substitution (MRS) between goods; this includes the equality of the real wage to the MRS between goods and leisure. These equalities hold *even when market prices do not reflect the economy's marginal rate of transformation (MRT) between those goods.* Equivalently, we need only to investigate the expenditure side of the National Income and Product Accounts (NIPA) identity; we do not need to know the production technology of firms or the competitive structure of industries.

There are two qualifications to our argument. First, the ratio of factor prices may not equal the consumer's marginal rate of substitution: Taxes, for example, create a wedge between the two, since the wage paid by firms then differs from the wage received by households. The welfare interpretation of the residual requires factor prices received by households, but this modification is straightforward: All prices should be those perceived by the household. Second and more seriously, the representative-consumer assumption may fail. Consumers may have different marginal utilities of wealth or, as in standard efficiency-wage models or bargaining models, they may face different prices. In this case, one cannot compute aggregate welfare change from aggregate statistics alone. However, we do not claim that our proposed productivity measure is a completely general measure of welfare change, merely that it is one under much more general conditions than the usual Solow residual. It seems a particularly apt measure in the context of recent macroeconomic models with a representative consumer but with imperfect competition in product markets (e.g., Rotemberg and Woodford 1992, 1995), or with multiple sectors and costly factor reallocation (e.g., Ramey and Shapiro 1998).

Figure 7.1 shows the economic intuition underlying our argument. Suppose the economy produces two goods, both of which are consumed (and possibly also used as intermediate inputs). To keep the graph simple, assume that the supplies of capital and labor are fixed. The PPF depicts all feasible (C_1, C_2) pairs. An economy without distortions attains the social optimum at point A, supported by relative prices P_1/P_2. Now suppose there are distortions. Then the economy might be at an allocatively inefficient point on the PPF, like point B, or even within the PPF, like point C. As shown in figure 7.1, these outcomes can be supported by price ratios different from the MRT between C_1 and C_2.

Note that in all cases the consumer's budget line shows the economy's

iso-output line, which aggregates heterogeneous output using market prices (regardless of whether these prices reflect technological tradeoffs.).[39] Thus, in this example welfare increases only if output increases. (This is a special case of our general result that welfare increases only if productivity increases, since in this example $dk = dl = 0$.) Hulten (1978) shows that under Solow's conditions—perfect competition and constant returns—aggregate productivity growth represents both technology improvement and welfare increase. In terms of figure 7.1, Hulten's result applies to an economy at point A: Output (productivity) can increase only if the PPF shifts out at point A, that is, if there is (local) technological improvement.

However, the same is not true at points B and C: Output (productivity) can increase without any change in technology, as long as distortions lessen. But these productivity improvements raise welfare, since output and inputs are weighted using prices that reflect the consumer's MRS between goods. Thus, Basu and Fernald's (1997b) finding generalizes Hulten's (1978) result to the case of imperfect competition and nonconstant returns, and clarifies the essence of his argument linking productivity and welfare.

7.6.2 Reallocations as Propagation Mechanisms

It is reasonable for practical macroeconomists to ask when and how they can avoid the perils posed by the nonexistence of an aggregate production function (even to a first-order approximation). It is also reasonable to ask whether the reallocation effects that lead to aggregation failures can also serve a positive function, by providing new amplification and propagation mechanisms in macro models. In this section we take a first pass at these large and difficult questions. Much of our intuition comes from the extended example of Basu and Fernald (1997a, section 5).

When can we ignore heterogeneity in production and act as if a representative firm produces all output? Doing so means modeling the production side of the economy using an aggregate production function for GDP. We ask three questions. First, is this procedure ever sensible if the world actually has significant heterogeneity? Second, what parameters should one use to calibrate the assumed aggregate production function? In particular, can one use estimates from aggregate data and ignore heterogeneity? Third, will the model with heterogeneity reproduce some of the interesting macroeconomic properties of the single-firm model it replaces?

We emphasize that in this subsection, as in the previous one, we abstract from variations in utilization. Capacity utilization is an important empirical issue, because it implies that certain inputs are unobserved by the econometrician. From the standpoint of theory, however, the possibility of changing utilization just changes the calibration of a model by changing

39. For a proof, differentiate the consumer's budget constraint holding income and prices fixed.

elasticities of supply and demand, but is not a qualitatively new effect. (For example, variable effort implies that the correct labor supply elasticity is larger than the usual data suggest; variable capital utilization implies that the labor demand curve is flatter than the standard production function would indicate.) In this section, by contrast, we ask whether reallocations of inputs constitute a qualitatively new propagation mechanism.

We now discuss whether a representative-firm model can capture the important features of a world where firms have approximately constant returns to scale, but where there are large reallocation effects. When it comes to fluctuations in output conditional on changes in aggregate inputs, the answer is often yes. The reason is that in many ways reallocations act like increasing returns to scale at a representative firm—in both cases, a one percent change in aggregate inputs is associated with a greater than one percent change in output. The major difference is that in the representative-firm economy, the degree of returns to scale is a structural parameter. In the economy with reallocations, however, the "effective returns to scale" depends on the nature of the reallocations induced by the driving shocks. Different shocks are likely to induce different degrees and types of reallocation, leading to variations in the effective returns to scale parameter under different circumstances—a classic example of the Lucas (1976) critique. However, this variation may actually prove an advantage under some circumstances. To the extent that the economy responds differently to shocks of different types, variable reallocation effects may help explain why. Basu, Fernald, and Horvath (1996) address this question, among others.

This answer to the first question that we posed at the beginning of this subsection also partially answers the second, on the issue of calibration. We believe that if any single summary statistic is useful, it is likely to be the degree of returns to scale (or markups) estimated from aggregate data without composition corrections. If the Lucas critique problem is not too severe, this parameter will correctly capture the procyclical behavior of aggregate output and productivity, but a one-sector model calibrated with the average of the firm-level parameters would be unable to replicate this behavior. But, as the discussion above indicated, a one-sector model can never be a perfect substitute for a multi-sector model if one wishes to understand the response to multiple shocks.

The final question—whether the model with heterogeneity reproduces some of the interesting macroeconomic properties of the single-firm model—is the hardest to answer. Indeed, the example of Basu and Fernald (1997a) shows that there are no general answers. Here we discuss one of many interesting issues, the possibility of positive-feedback loops that might magnify the effects of shocks, and in the limit give rise to multiple equilibria. The work on real rigidities (Ball and Romer 1990) and "indeterminacy" (e.g., Farmer and Guo 1994) has drawn attention to the importance of positive feedback as a propagation mechanism. Farmer and Guo show that a very strong form of increasing returns—increasing marginal

product of inputs—provides such positive feedback. The intuition is simple. Suppose the labor demand curve is upward-sloping (due to increasing marginal product of labor) and steeper than the labor supply curve. Also suppose a shock increases lifetime wealth, causing workers to supply less labor at each real wage (without affecting labor demand). This leftward shift in the labor supply curve causes equilibrium labor supply to increase rather than decrease, as standard neoclassical theory predicts. If this effect is sufficiently strong, the increase in labor and output can be self-justifying. Workers expect higher real wages now and in the future, which reduces their marginal utility of wealth, which increases the equilibrium real wage—validating the initial expectation of higher lifetime wealth.[40]

An interesting question, then, is whether reallocation effects can create such positive-feedback loops. Note that positive feedback depends on changes in marginal factor prices, since these are the prices that are relevant for economic decisions (saving and labor supply). In fact, increasing returns in the normal sense is not sufficient to create positive feedback. As we discussed in Section I, returns to scale can come from fixed costs, and can be quite consistent with diminishing marginal products. Thus, we conjecture that reallocations coming from differences in markups, R_{μ}, and the failure of a value-added production function, R_M, will have a positive-feedback effect only if the increasing returns accompanying the markups take the form of diminishing marginal cost.[41]

Similar caveats apply to R_K and R_L. Recall that these terms capture differences in shadow prices of the same input across firms. If the variance in shadow prices comes only from adjustment costs, and adjustment costs are paid by the firm, then the differences in marginal product will not translate directly to differences in factor prices. (Of course, adjustment costs usually have general-equilibrium effects on both prices and quantities.) Again, the reallocations induced by quasi-fixity will typically affect the dynamics of aggregate output following a shock (as discussed by Ramey and Shapiro 1998).

However, if the R_K and R_L terms come from steady-state differences in factor payments across sectors, matters might be different. (In the case of labor, for example, these differences might come from efficiency wages or union wage premia.) If the wage differences are allocative for the firm—that is, if the firm equates the marginal product of labor to the above-market-clearing wage, as in Katz and Summers (1989), instead of the wage premium reflecting efficient bargaining—then these reallocations can have feedback effects. Even in this case, however, positive feedback is not guar-

40. In a dynamic model one also needs an increase in capital accumulation to increase the rate of return to capital.

41. As discussed in section 7.1, we believe on theoretical and empirical grounds that free entry eliminates long-run economic profits, forcing markups to approximately equal the degree of returns to scale.

anteed, because it is unclear what form of rationing rule supports long-run differences in wages for identical labor—that is, how high-paying jobs are restricted to a subset of workers. Basu and Fernald (1997a, Section V) show that the form of the rationing rule determines whether long-run wage premia that give rise to a non-zero R_L also generate positive feedback. Thus, the implications of reallocation for macroeconomic models are likely to be quite sensitive to institutional assumptions, including institutions that do not directly concern production and firm behavior.

The general point of this discussion is that reallocation effects are not just a "nuisance term"—they are potentially important propagation mechanisms for shocks, and in the limit can give rise to multiple equilibria. This is an important lesson, as recent empirical estimates (including those we provide above) suggest that average, firm-level returns to scale are approximately constant. We argue that this finding does not necessarily imply that one should reject macroeconomic parables in which increasing returns at a representative plant play a central role in explaining economic fluctuations. Ascertaining which paradigm provides better macroeconomic insights is an important, unresolved question, and the focus of on-going research.

7.7 Conclusions

In this paper, we explore the meaning and measurement of productivity in a world with frictions and distortions. In such a world, productivity growth might not estimate aggregate technology change. We provide a general accounting framework that relates growth in aggregate productivity and aggregate technology. We identify various nontechnological terms that reflect not only variations in utilization, but also changes in the allocation of factors across uses with different marginal products. Marginal products, in turn, can differ because of frictions or distortions in the economy. These reallocations affect aggregate output and productivity, without necessarily reflecting technology. Hence, computing aggregate technology change requires micro data.

The nontechnological components should not necessarily be considered "mismeasurement." Variable input utilization clearly constitutes mismeasurement, but reallocations do not. In fact, we argue in section 7.6 that even with distortions such as imperfect competition, a modified Solow residual appropriately measures welfare change. Thus, though much of the recent productivity literature emphasizes the use of micro data, in some circumstances welfare measurement requires only readily-available national income accounts data.[42]

Several existing studies provide models of fluctuations in economies that

42. In practice, of course, welfare depends on a broader measure of output than just GDP—for example, household production, investment in human capital, and changes in environmental policy. Unfortunately, these items are hard to measure.

deviate in a variety of ways from the standard one-sector model of production. In those models, the nontechnological sources of productivity fluctuations are not always clear, nor is the relationship to other models. Our general framework can aid in understanding and interpreting the fluctuations arising, for example, from sector-specific technology shocks, vintage capital effects, or imperfect competition with heterogeneity (see, respectively, Phelan and Trejos 1996, Gilchrist and Williams 1996, and Basu, Fernald, and Horvath 1996).

Applying our decomposition to the data raises several practical and methodological issues. We discuss the pros and cons of estimating first- and second-order approximations to the production function, and advocate an explicitly first-order approach. We use a model of a dynamically cost-minimizing firm to derive a proxy for unobserved variations in labor effort and the workweek of capital. We estimate aggregate technology change by aggregating industry-level shocks that are "purged" of the effects of variable utilization and imperfect competition.

Variable utilization and cyclical reallocations appear to explain much of the cyclicality of aggregate productivity. We find that in the short run, technology improvements significantly reduce input use while appearing to reduce output slightly as well. These results are inconsistent with standard parameterizations of RBC models, which imply that technology improvements raise input use at all horizons. However, Basu, Fernald, and Kimball (1999) argue that they *are* consistent with standard sticky-price models.

Finally, we discuss implications for macroeconomics. We conclude that reallocations are welfare-relevant, and hence not biases. We also conclude that while accounting for reallocation reduces the average markup and the average degree of returns to scale, an economy with strong reallocation effects can sometimes display the same behavior as an economy with large increasing returns. Hence, reallocations constitute a potentially-important propagation mechanism, which can be utilized in multi-sector dynamic models of business cycles.

Appendix

Derivations from Section 7.3

This appendix presents detailed derivations of the equations in section 7.3. We first derive the relationship between firm-level gross output and firm-level value added, and discuss the interpretation of the value-added equation. We then aggregate firm-level value-added growth to obtain aggregate output growth as a function of aggregate inputs, technology, utilization, and reallocations of resources.

As discussed in Section III, the Divisia definition of value-added growth is:[43]

$$(A.1) \qquad dv_i = \frac{dy_i - s_{Mi}dm_i}{1 - s_{Mi}} = dy_i - \left[\frac{s_{Mi}}{1 - s_{Mi}}\right](dm_i - dy_i).$$

We now want to substitute in for output growth dy_i. In section 7.1, we obtained the following equation for output growth

$$dy_i = \mu_i dx_i + \mu_i(1 - s_{Mi})du_i + dt_i.$$

Inserting this equation using the definition of input growth dx_i gives

$$(A.2) \quad dy_i = \mu_i[s_{Ki}dk_i + s_{Li}(dn_i + dh_i) + s_{Mi}dm_i] + \mu_i(1 - s_{Mi})du_i + dt_i.$$

We can rewrite this equation as

$$(A.3) \qquad \begin{aligned} dy_i &= \mu_i(1 - s_{Mi})\left[\frac{s_{Li}}{(1 - s_{Mi})}dl_i + \frac{s_{Ki}}{(1 - S_{Mi})}dk_i\right] \\ &\quad + \mu_i(1 - s_{Mi})du_i + \mu_i s_{Mi}dm_i + dt_i \\ &= \mu_i(1 - s_{Mi})(dx_i^V + du_i) + \mu_i s_{Mi}dm_i + dt_i, \end{aligned}$$

where primary-input growth, dx_i^V, is defined analogously to aggregate primary input growth

$$dx_i^V = \frac{s_{Ki}}{1 - s_{Mi}}dk_i + \frac{s_{Li}}{1 - s_{Mi}}dl_i \equiv s_{Ki}^V dk_i + s_{Li}^V dl_i.$$

Note that s_{Ki}^V and s_{Li}^V are shares of capital and labor costs in nominal value added.

Now subtract $\mu_i s_{Mi}dy_i$ from both sides of equation (A.3) and divide through by $(1 - \mu_i s_{Mi})$. This gives

$$\begin{aligned} dy_i &= \left[\frac{\mu_i(1 - s_{Mi})}{1 - \mu_i s_{Mi}}\right](dx_i^V + du_i) + \\ &\quad \left[\frac{\mu_i s_{Mi}}{1 - \mu_i s_{Mi}}\right](dm_i - dy_i) + \frac{dt_i}{1 - \mu_i s_{Mi}}. \end{aligned}$$

43. Double-deflation is an alternative method, where $V \equiv Y - M$, where Y and M are valued at base-year prices. By differentiating the double-deflated index, we can express the growth in double-deflated value added in a form completely parallel to equation (7A.1)—the only difference is that the intermediate share is calculated using *base-year* prices, rather than current prices. An implication of this difference is that just as substitution bias affects Laspeyres indices of aggregate expenditure, substitution bias affects double-deflated value added. Hence, as Basu and Fernald (1995, appendix) show, double-deflated value added suffers all of the biases we identify in this section for Divisia value-added, *plus* an additional substitution bias.

We can write this equation as

$$(A.4) \quad dy_i = \mu_i^V(dx_i^V + du_i) + \mu_i^V\left(\frac{s_{Mi}}{1 - s_{Mi}}\right)(dm_i - dy_i) + dt_i^V,$$

where

$$(A.5) \quad \mu_i^V \equiv \mu_i \frac{1 - s_{Mi}}{1 - \mu_i s_{Mi}}$$

$$(A.6) \quad dt_i^V \equiv \frac{dt_i}{1 - \mu_i s_{Mi}}.$$

Equation (A.4) relates gross output growth to growth rates of primary inputs dx_i^V, utilization du_i, the intermediates-output ratio $dm_i - dy_i$, and technology. Growth in primary inputs and utilization are multiplied by a "value-added markup" μ_i^V, defined by equation (A.5). We provide an economic interpretation of μ_i^V below.

We can now substitute for dy_i from equation (A.4) into the definition of value-added growth (equation [A.1])

$$(A.7) \quad dv_i = \mu_i^V dx_i^V + (\mu_i^V - 1)\left[\frac{s_{Mi}}{1 - s_{Mi}}\right](dm_i - dy_i) + \mu_i^V du_i + dt_i^V.$$

The firm's revenue-weighted value-added productivity residual, dp_i, equals $dv_i - dx_i^V$. Hence,

$$(A.8) \quad dp_i = (\mu_i^V - 1)dx_i^V + (\mu_i^V - 1)\left[\frac{s_{Mi}}{1 - s_{Mi}}\right] (dm_i - dy_i) + \mu_i^V du_i + dt_i.$$

It is obvious from equations (A.7) and (A.8) that value-added growth is not, in general, simply a function of primary inputs dx_i^V. A long literature in the 1970s (e.g., Bruno 1978) explored whether a "value-added function" exists, and argued that the answer depends on separability properties of the production function. The equation above shows that with imperfect competition, taking value added to be a function only of primary inputs is generally misspecified—regardless of whether the production function is separable between value added and intermediate inputs. Separability is a second-order property of a production function, so its presence or absence does not affect first-order approximations like equation (A.7). However, the fact that the output elasticity of materials is $\mu_i s_{Mi}$ instead of simply s_{Mi} is of first-order importance.

Nevertheless, it will be useful to make that further assumption of separability in order to provide a simple interpretation of value-added growth. In particular, suppose the production function is separable, as follows

$$\text{(A.9)} \qquad Y_i = F^i(\tilde{K}_i, L_i, M_i, T^V_i) = G^i[V^{Pi}(\tilde{K}_i, L_i, T^V_i), H^i(M_i)].$$

The firm combines primary inputs to produce "productive value added," V^{Pi}, which it then combines with intermediate inputs to produce gross output. (Note that V^{Pi} does not necessarily correspond to the national-accounting measure of value added—that is, the sum of firm-level productive value added need not equal national expenditure.) We can break the cost-minimization problem into two stages: First, minimize the cost of using primary inputs to produce any level of V^{Pi}; second, minimize the cost of using productive value added and intermediate inputs to produce any level of gross output.

In the first stage, the logic from equation (15) implies that the "productive" value-added growth, dv^P, depends on the revenue-weighted growth in primary inputs dx^V, plus technology shocks (without loss of generality we normalize to one the elasticity of productive value added V^{Pi} with respect to technology)

$$\text{(A.10)} \qquad dv^P_i = \mu^V_i dx^V_i + dt^V_i.$$

In the second stage, the firm seeks to minimize the cost of using value added and intermediate inputs to produce gross output. The cost in the minimation problem is $\text{MC}^V_i V_i + P_{Mi} M_i$, where MC^V_i is the marginal cost of value added to the firm. The first-order condition is then $\text{MC}^V_i = P_i G^i_V / \mu_i$. Analogously to the problem in section 7.1, we can interpret the value-added markup μ^V_i as the ratio of the price of productive value added to the marginal cost of producing it: $\mu^V_i = P^V_i / \text{MC}^V_i$. Hence,

$$\text{(A.11)} \qquad \frac{G_V V^{Pi}}{G} = \frac{\mu_i}{\mu^V_i} s_{Vi}.$$

Note that s_{Vi} equals $P^V_i V_i / P_i Y_i = (P_i Y - P_{Mi} M_i)/P_i Y$, which equals $(1 - s_{Mi})$. However, without knowing more about the shape of the production function (and hence, the slopes of marginal cost of producing V^P and Y), we cannot make any general statements about the magnitude of the value-added markup μ^V_i.

To do so, we make the further substantive assumption that all returns to scale are in V^P, arising perhaps from overhead capital or labor. This requires that G be homogeneous of degree one in V^P and H, and that H be homogeneous of degree one in M. Under these assumptions, the left-hand-side of equation (A.11) equals $(1 - \mu_i s_{Mi})$. Equation (A.11) then reads

$$\text{(A.12)} \qquad (1 - \mu_i s_{Mi}) = \frac{\mu_i}{\mu^V_i}(1 - s_{Mi}).$$

Rearranging this equation verifies that the value-added markup μ^V_i is the same as we defined in equation (A.5). Hence, just as μ_i creates a wedge between an input price and the input's marginal product in terms of gross

output, μ_i^V appropriately measures the wedge between the input price and the marginal product in terms of value added.

Note that for macroeconomic modeling, the value-added markup μ^V is likely to be the parameter of interest. Rotemberg and Woodford (1995), for example, make this point in a dynamic general-equilibrium model with imperfect competition. In their model, there are no factor-market frictions, all firms have the same separable gross-output production function, always use materials in fixed proportions to gross output, and charge the same markup of price over marginal cost. Under these assumptions (which are implicit or explicit in most other dynamic general-equilibrium models with imperfect competition), there is a representative producer and an aggregate value-added production function. Rotemberg and Woodford show that the correct aggregate markup corresponds to μ^V.

One source of economic intuition for μ^V is that under some circumstances, it correctly captures "economy-wide" distortions, as small markups at the plant level translate into larger efficiency losses for the economy overall. Suppose, for example, that final output is produced at the end of an infinite number of stages. At each stage a representative firm with markup μ uses all the output of the previous stage as intermediate input, and also uses primary inputs of capital and labor. Then the percent change in national income—the output of the last (nth) stage—is

$$\begin{aligned} dy_n &= \mu(1 - s_M)dx_n^V + \mu s_M dy_{n-1} \\ &= \mu(1 - s_M)dx_n^V + \mu s_M \mu(1 - s_M)dx_{n-1}^V + (\mu s_M)^2 dy_{n-2}. \end{aligned}$$

We can substitute into this equation for dy_{n-j}, and let j go to infinity. Since each firm is identical, dx_i^V is the same for all i as for the aggregate. This gives an infinite sum

$$\begin{aligned} \text{(A.13)} \qquad dy_n \equiv dv &= \mu(1 - s_M)dx^V \sum_{j=1}^{\infty} (\mu s_M)^j \\ &= \mu(1 - s_M)dx^V \frac{1}{1 - \mu s_M} = \mu^V dx^V. \end{aligned}$$

Thus, μ^V is plausibly the appropriate concept for calibrating the markup charged by the representative firm in a one-sector macroeconomic model.

Thus, μ^V correctly captures the idea that small deviations from perfect competition "cascade" in going from gross output to value added, because of the "markup on markup" phenomenon: Firms buy intermediate goods from other firms at a markup, add value, and again price the resulting good with a markup, generally selling some of it to another firm to use as an intermediate good. Nevertheless, this derivation shows that there is a limit to how much the effects can cascade or build up.

Even if we want this value-added markup, however, we still in general require data on intermediate inputs. The reason is that we do not observe

V^{Pi} directly, but must infer it from observable gross output and intermediate inputs.

Returning to equation (A.7), real value-added growth depends on primary input growth, changes in the materials-to-output ratio, variations in utilization of capital and labor, and technology. The first term shows that primary inputs are multiplied by the value-added markup. The second term reflects the extent to which the standard measure of value added differs from "productive" value added V^{Pi}, and hence does not properly measure the productive contribution of intermediate inputs. Intuitively, the standard measure of value added subtracts off intermediate input growth using revenue shares, whereas with imperfect competition the productive contribution of these inputs exceeds the revenue share by the markup. The third term shows that variations in utilization are also multiplied by the value-added markup. The fourth term is the value-added-augmenting technology shock.

We now aggregate over firms to get aggregate output growth as a function of technology, aggregate primary inputs growth, and the distribution of inputs. As in section 7.3, we define aggregate output growth dv as a share-weighted average of firm-level value-added growth

$$dv = \sum_{i=1}^{N} w_i dv,$$

where $w_i = P_i^V V / P^V V$. Substituting in from equation (A.7) for dv_i gives

$$\text{(A.14)} \quad dv = \sum_{i=1}^{N} w_i \mu_i^V dx_i^V + \sum_{i=1}^{N} w_i (\mu_i^V - 1)\left[\frac{s_{Mi}}{1 - s_{Mi}}\right](dm_i - dy_i) + \sum_{i=1}^{N} w_i \mu_i^V du_i + \sum_{i=1}^{N} w_i dt_i^V.$$

As in the text, we will define

$$du = \sum_{i=1}^{N} w_i \mu_i^V du$$

$$R_M = \sum_{i=1}^{N} w_i (\mu_i^V - 1)\left[\frac{s_{Mi}}{1 - s_{Mi}}\right](dm_i - dy_i), \text{ and}$$

$$dt^V = \sum_{i=1}^{N} w_i dt_i^V.$$

Hence, equation (A.14) becomes

$$\text{(A.15)} \quad dv = \sum_{i=1}^{N} w_i \mu_i^V dx_i^V + R_M + du + dt^V$$

We now decompose the first term into the effects of the "average" value-added markup $\bar{\mu}^V$, and the distribution of the markup. Rearranging equation (A.15) gives

$$\text{(A.16)} \quad dv = \bar{\mu}^V \sum_{i=1}^{N} w_i dx_i^V + \sum_{i=1}^{N} w_i(\mu_i^V - \bar{\mu}^V) dx_i^V + R_M + du + dt^V$$

$$= \bar{\mu}^V \sum_{i=1}^{N} w_i dx_i^V + R_\mu + R_M + du + dt^V,$$

where

$$R_\mu = \sum_{i=1}^{N} w_i(\mu_i^V - \bar{\mu}^V) dx_i^V.$$

At this point, we are almost finished. However, we still need to relate the first term on the right-hand side of equation (A.16) to aggregate input growth. As in section 7.3, consider the case where there was only one type of capital and one type of labor.[44] Aggregate labor and capital are arithmetic sums across firms, so that $K = \Sigma_{i=1}^{N} K_i$ and $L = \Sigma_{i=1}^{N} L_i$. Aggregate primary input growth is the share-weighted growth in aggregate capital and labor growth

$$dx^V = s_K^V dk + s_L^V dl.$$

Using the definitions of s_{Ki}^V and s_{Li}^V, and differentiating the definitions of aggregate K and L, we can write this as

$$dx^V = \frac{P_K K}{P^V V} \sum_{i=1}^{N} \frac{K_i}{K} dk_i + \frac{P_L K}{P^V V} \sum_{i=1}^{N} \frac{L_i}{L} dl_i$$

$$= \sum_{i=1}^{N} \frac{P_i^V V_i}{P^V V} \frac{P_{Ki} K_i}{P_i^V V_i} \frac{P_K}{P_{Ki}} dk_i + \sum_{i=1}^{N} \frac{P_i^V V_i}{P^V V} \frac{P_{Li} L_i}{P_i^V V_i} \frac{P_L}{P_{Li}} dl_i.$$

Noting the definitions of w_i, s_{Ki}^V, and s_{Li}^V, we can write this as

$$\text{(A.17)} \quad dx^V = \sum_{i=1}^{N} w_i s_{Ki}^V \frac{P_K}{P_{Ki}} dk_i + \sum_{i=1}^{N} w_i s_{Li}^V \frac{P_L}{P_{Li}} dl_i$$

$$= \sum_{i=1}^{N} w_i (s_{Ki}^V dk_i + s_{Li}^V dl_i) - \sum_{i=1}^{N} w_i s_{Ki}^V \left(\frac{P_{Ki} - P_K}{P_{Ki}} \right) dk_i$$

$$- \sum_{i=1}^{N} w_i s_{Li}^V \left(\frac{P_{Li} - P_L}{P_{Li}} \right) dl_i = \sum_{i=1}^{N} w_i dx_i^{VV} - R_K - R_L,$$

where

44. In general, suppose there are N firms, N_L types of labor, and N_K types of capital. For each type of capital K^k and labor L^l, the aggregate is an arithmetic sum across firms, so that $K^k = \Sigma_{i=1}^{N} K_i^K$ and $L^l = \Sigma_{i=1}^{N} L_i^l$. Aggregate capital and labor are then defined as a Divisia index across these types of capital, so that, for example, aggregate labor growth is $dl \equiv \Sigma_{L^l=1}^{N} P_L^l L^l / P_L L dl^l$, where $P_L^l L^l$ is total labor compensation to labor of type l. The derivations that follow extend easily to this general case, except that there is a separate input-reallocation term for each type of labor and capital. Jorgenson, Gollop, and Fraumeni (1987) derive this result explicitly.

$$R_K = \sum_{i=1}^{N} w_i s_{Ki}^V \left[\frac{P_{Ki} - P_K}{P_{Ki}} \right] dk_i ,$$

$$R_L = \sum_{i=1}^{N} w_i s_{Li}^V \left[\frac{P_{Li} - P_L}{P_{Li}} \right] dl_i ,$$

By substituting equation (A.17) into equation (A.16), we find

$$\text{(A.18)} \quad dv = \bar{\mu}^V dx^V + du + R_\mu + R_M + \bar{\mu}^V R_K + \bar{\mu}^V R_L + dt^V$$
$$= \bar{\mu}^V dx^V + du + R + dt^V .$$

Since aggregate productivity equals $dv - dx^V$, it immediately follows that

$$\text{(A.19)} \quad dp = (\bar{\mu}^V - 1) dx^V + du + R + dt .$$

References

Abbott, Thomas A., Zvi Griliches, and Jerry Hausman. 1998. Short run movements in productivity: Market power versus capacity utilization." In *Practicing econometrics: Essays in method and application,* ed. Zvi Griliches, 333–42. Cheltenham, UK: Elgar.

Baily, Martin N., Charles Hulten, and David Campbell. 1992. Productivity dynamics in manufacturing plants. *Brookings Papers on Economic Activity (Microeconomics),* issue no. 1:187–267.

Ball, Laurence, and David Romer. 1990. Real rigidities and the non-neutrality of money. *Review of Economic Studies* 57:183–203.

Barro, Robert J., and Robert G. King. 1984. Time-separable preferences and intertemporal substitution models of business cycles. *Quarterly Journal of Economics* 99 (November): 817–39.

Bartelsman, Eric J., Ricardo J. Caballero, and Richard K. Lyons. 1994. Customer- and supplier-driven externalities. *American Economic Review* 84 (4): 1075–84.

Bartelsman, Eric J., and Phoebus Dhrymes. 1998. Productivity dynamics: U.S. manufacturing plants, 1972–1986. *Journal of Productivity Analysis* 9 (1): 5–34.

Basu, Susanto. 1995. Intermediate goods and business cycles: Implications for productivity and welfare. *American Economic Review* 85 (June): 512–31.

———. 1996. Cyclical productivity: Increasing returns or cyclical utilization? *Quarterly Journal of Economics* 111 (August): 719–51.

Basu, Susanto, and John G. Fernald. 1995. Are apparent productive spillovers a figment of specification error? *Journal of Monetary Economics* 36 (December): 165–88.

———. 1997a. Returns to scale in U.S. manufacturing: Estimates and implications. *Journal of Political Economy* 105 (April): 249–83.

———. 1997b. Aggregate productivity and aggregate technology. International Finance Discussion Paper no. 593. Federal Reserve System, Board of Governors.

Basu, Susanto, John G. Fernald, and Michael T. K. Horvath. 1996. Aggregate production function failures. Manuscript.

Basu, Susanto, John G. Fernald, and Miles S. Kimball. 1999. Are technology improvements contractionary? Manuscript.

Basu, Susanto, and Miles S. Kimball. 1997. Cyclical productivity with unobserved input variation. NBER Working Paper no. 5915. Cambridge, Mass.: National Bureau of Economic Research, February.

Baxter, Marianne, and Robert King. 1991. Productive externalities and business cycles. Discussion Paper no. 53. Federal Reserve Bank of Minneapolis, Institute for Empirical Macroeconomics.

Beaudry, Paul, and Michael Devereux. 1994. Monopolistic competition, price setting, and the effects of real and nominal shocks. Boston University, Department of Economics, Manuscript.

Beaulieu, John J., and Joseph Mattey. 1998. The workweek of capital and capital utilization in manufacturing. *Journal of Productivity Analysis* 10 (October): 199–223.

Benhabib, Jess, and Roger E. A. Farmer. 1994. Indeterminacy and increasing returns. *Journal of Economic Theory* 63 (1): 19–41.

Berman, Eli, John Bound, and Zvi Griliches. 1994. Changes in the demand for skilled labor within U.S. manufacturing: Evidence from the annual survey of manufactures. *Quarterly Journal of Economics* 109 (2): 367–97.

Berndt, Ernst R. 1991. *The practice of econometrics: Classic and contemporary.* Reading, Mass.: Addison Wesley.

Berndt, Ernst R., and Melvin A. Fuss. 1986. Productivity measurement with adjustments for variations in capacity utilization and other forms of temporary equilibrium. *Journal of Econometrics* 33 (October/November): 7–29.

Bils, Mark. 1987. The cyclical behavior of marginal cost and price. *American Economic Review* 77:838–55.

Bils, Mark, and Jang-Ok Cho. 1994. Cyclical factor utilization. *Journal of Monetary Economics* 33:319–54.

Blanchard, Olivier J., and Peter Diamond. 1990. The aggregate matching function. In *Growth/productivity/unemployment: Essays to celebrate Bob Solow's birthday,* ed. Peter Diamond, 159–201. Cambridge, Mass.: MIT Press.

Bruno, Michael. 1978. Duality, intermediate inputs, and value added. In *Production economics: A dual approach to theory and applications,* vol. 2, 3–16. ed. Melvyn Fuss and Daniel McFadden. Amsterdam: North-Holland.

Burnside, Craig. 1996. What do production function regressions tell us about increasing returns to scale and externalities? *Journal of Monetary Economics* 37 (April): 177–201.

Burnside, Craig, and Martin Eichenbaum. 1996. Factor-hoarding and the propagation of business-cycle shocks. *American Economic Review* 86:1154–74.

Burnside, Craig, Martin Eichenbaum, and Sergio Rebelo. 1995. Capital utilization and returns to scale. In *NBER Macroeconomics Annual 1995,* ed. Ben S. Bernanke and Julio J. Rotemberg, 67–110. Cambridge, Mass.: MIT Press.

Caballero, Ricardo J., and Richard K. Lyons. 1989. The role of external economies in U.S. manufacturing. NBER Working Paper no. 3033. Cambridge, Mass.: National Bureau of Economic Research.

———. 1990. Internal and external economies in European industries. *European Economic Review* 34:805–30.

———. 1992. External effects in U.S. procyclical productivity. *Journal of Monetary Economics* 29:209–26.

Carlton, Dennis W. 1983. Equilibrium fluctuations when price and delivery lag clear the market. *Bell Journal of Economics* 14 (2): 562–72.

Chambers, Robert G. 1988. *Applied production analysis: A dual approach.* Cambridge: Cambridge University Press.

Cooley, Thomas F., and Edward C. Prescott. 1995. Economic growth and business cycles. In *Frontiers of business cycle research,* ed. Thomas F. Cooley, 1–38. Princeton: Princeton University Press.

Cooper, Russell, and Andrew John. 1988. Coordinating coordination failures in Keynesian models. *Quarterly Journal of Economics* 103 (3): 441–63.
Cooper, Russell, and Alok Johri. 1997. Dynamic complementarities: A quantitative analysis. *Journal of Monetary Economics* 40 (1): 97–119.
Diamond, Peter A. 1982. Aggregate demand management in search equilibrium. *Journal of Political Economy* 90:881–94.
Diewert, Erwin. 1976. Exact and superlative index numbers. *Journal of Econometrics* 4:115–46.
Domar, Evsey D. 1961. On the measurement of technical change. *Economic Journal* 71 (December): 710–29.
Dotsey, Michael, Robert King, and Alexander Wolman. 1997. Menu costs, staggered price setting, and elastic factor supply. University of Virginia, Department of Economics, Manuscript.
Farmer, Robert, and Jang-Ting Guo. 1994. Real business cycles and the animal spirits hypothesis. *Journal of Economic Theory* 63:42–72.
Flux, A. W. 1913. Gleanings from the Census of Productions report. *Journal of the Royal Statistical Society* 76 (6): 557–85.
Gali, Jordi. 1999. Technology, employment, and the business cycle: Do technology shocks explain aggregate fluctuations? *American Economic Review* 89 (March): 249–74.
Gilchrist, Simon, and John Williams. 1996. Investment, capacity, and output: A putty-clay approach. Finance and Economics Discussion Series no. 1998–44. Federal Reserve System, Board of Governors.
Gordon, Robert J. 1993. Are procyclical productivity fluctuations a figment of measurement error? Northwestern University, Department of Economics, Mimeograph.
Green, Edward J., and Robert H. Porter. 1984. Noncooperative collusion under imperfect price information. *Econometrica* 52 (January): 87–100.
Griliches, Zvi. 1992. The search for R&D spillovers. *Scandinavian Journal of Economics* 94 (Supplement): 29–47.
Griliches, Zvi, and Tor Jacob Klette. 1996. The inconsistency of common scale estimators when output prices are unobserved and endogenous. *Journal of Applied Econometrics* 11 (July/August): 343–61.
Griliches, Zvi, and Jacques Mairesse. 1998. Production functions: The search for identification. In *Practicing econometrics: Essays in method and application,* ed. Zvi Griliches, 343–411. Cheltenham, U.K.: Elgar.
Haavelmo, Trygve. 1960. *A study in the theory of investment.* Chicago: University of Chicago Press.
Hall, Alastair R., Glenn D. Rudebusch, and David W. Wilcox. 1996. Judging instrument relevance in instrumental variables estimation. *International Economic Review* 37 (2): 283–98.
Hall, Robert E. 1988. The relation between price and marginal cost in U.S. industry. *Journal of Political Economy* 96 (October): 921–47.
———. 1990. Invariance properties of Solow's productivity residual. In *Growth/productivity/unemployment: Essays to celebrate Bob Solow's birthday,* ed. Peter Diamond. Cambridge, Mass.: MIT Press.
Hall, Robert E., and Dale W. Jorgenson. 1967. Tax policy and investment behavior. *American Economic Review* 57 (June): 391–414.
Haltiwanger, John C. 1997. Measuring and analyzing aggregate fluctuations: The importance of building from microeconomic evidence. *Review of the Federal Reserve Bank of St. Louis* (79) (3): 55–77.
Horvath, Michael T. K. 1995. Cyclicality and sectoral linkages: Aggregate fluctu-

ations from independent sectoral shocks. *Review of Economic Dynamics* 1 (4): 781–808.

Hulten, Charles. 1978. Growth accounting with intermediate inputs. *Review of Economic Studies* 45:511–18.

———. 1986. Productivity change, capacity utilization, and the sources of efficiency growth. *Journal of Econometrics* 33 (October/November): 31–50.

Jorgenson, Dale W. 1987. Productivity and postwar U.S. economic growth. *Journal of Economic Perspectives* 2 (4): 23–42.

Jorgenson, Dale W., Frank Gollop, and Barbara Fraumeni. 1987. *Productivity and U.S. economic growth.* Cambridge, Mass.: Harvard University Press.

Jorgenson, Dale W., and Zvi Griliches. 1967. The explanation of productivity change. *Review of Economic Studies* 34:249–83.

Jorgenson, Dale W., and Kun-Young Yun. 1991. *Tax reform and the cost of capital.* Oxford: Oxford University Press.

Katz, Lawrence F., and Lawrence H. Summers. 1989. Industry rents: Evidence and implications. *Brookings Papers on Economic Activity (Microeconomics)* (1): 209–90.

Kimball, Miles S. 1995. The quantitative analytics of the basic neomonetarist model. *Journal of Money, Credit, and Banking* 27 (November): 1241–77.

Lau, Laurence. 1986. Functional forms in econometric model building. In *Handbook of Econometrics,* vol. 3, ed. Zvi Griliches and Michael Intriligator, 1515–66. Amsterdam: North-Holland.

Lilien, David M. 1982. Sectoral shifts and cyclical unemployment. *Journal of Political Economy* 90 (August): 777–93.

Lucas, Robert E., Jr. 1976. Econometric policy evaluation: A critique. *Journal of Monetary Economics* (Suppl. Series): 19–46.

Marschak, Jacob, and William H. Andrews Jr. 1994. Random simultaneous equations and the theory of production. *Econometrica* 12 (3/4): 143–205.

Morrison, Catherine. 1992a. Markups in U.S. and Japanese manufacturing: A short-run econometric analysis. *Journal of Business and Economic Statistics* 10 (1): 51–63.

———. 1992b. Unraveling the productivity growth slowdown in the United States, Canada and Japan: The effects of subequilibrium, scale economies and markups. *Review of Economics and Statistics* 74 (3): 381–93.

Olley, G. Steven, and Ariel Pakes. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64:1263–97.

Paul, Catherine J. Morrison, and Donald S. Siegel. 1999. Scale economies and industry agglomeration externalities: A dynamic cost function approach. *American Economic Review* 89 (1): 273–90.

Phelan, Christopher, and Alberto Trejos. 1996. On the aggregate effects of sectoral reallocation. *Journal of Monetary Economics* 45 (2): 249–68.

Pindyck, Robert S., and Julio J. Rotemberg. 1983. Dynamic factor demands and the effects of energy price shocks. *American Economic Review* 73 (December): 1066–79.

Ramey, Valerie A., and Matthew D. Shapiro. 1998. Costly capital reallocation and the effects of government spending. *Carnegie-Rochester Conference Series on Public Policy* 48 (June): 145–94.

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

Rotemberg, Julio J., and Garth Saloner. 1986. A supergame-theoretic model of price wars during booms. *American Economic Review* 76 (June): 390–407.

Rotemberg, Julio J., and Michael Woodford. 1992. Oligopolistic pricing and the

effects of aggregate demand on economic activity. *Journal of Political Economy* 100 (December): 1153–07.
———. 1995. Dynamic general equilibrium models with imperfectly competitive product markets. In *Frontiers of business cycle research,* ed. Thomas F. Cooley, 243–93. Princeton: Princeton University Press.
Sato, K. 1976. The meaning and measurement of the real value added index. *Review of Economics and Statistics* 58:434–42.
Sbordone, Argia M. 1997. Interpreting the procyclical productivity of manufacturing sectors: External effects or labor hoarding? *Journal of Money, Credit, and Banking* 29 (1): 26–45.
Schmitt-Grohé, Stephanie. 1997. Comparing four models of aggregate fluctuations due to self-fulfilling expectations. *Journal of Economic Theory* 72 (1): 96–147.
Schor, Juliet B. 1987. Does work intensity respond to macroeconomic variables? Evidence from British manufacturing, 1970–1986. Harvard University Department of Economics, Manuscript.
Shapiro, Matthew D. 1986. Capital utilization and capital accumulation: Theory and evidence. *Journal of Applied Econometrics* 1:211–34.
———. 1996. Macroeconomic implications of variation in the workweek of capital. *Brookings Papers on Economic Activity* issue no. 2:79–119.
Shea, John. 1997. Instrument relevance in multivariate linear models: A simple measure. *Review of Economics and Statistics* 79 (2): 48–52.
Solon, Gary, Robert Barsky, and Jonathan A. Parker. 1994. Measuring the cyclicality of real wages: How important is composition bias? *Quarterly Journal of Economics* 109 (February): 1–25.
Solow, Robert M. 1957. Technological change and the aggregate production function. *Review of Economics and Statistics* 39:312–20.
Varian, Hal. 1984. *Microeconomic analysis.* New York: W. W. Norton.
Weder, Mark. 1997. Animal spirits, technology shocks, and the business cycle. *Journal of Economic Dynamics and Control* 24 (2): 273–95.
Wen, Yi. 1998. Capacity utilization under increasing returns to scale. *Journal of Economic Theory* 81 (July): 7–36.

Comment Catherine J. Morrison Paul

My comments might be entitled, "Where Are the Microfoundations, and Do We Care?" In my view, yes, we do—or at least should—care. We care because the questions we are interested in asking, the explanatory power and interpretability we seek, and the implications for welfare and policy we pursue, all of which are crucial aspects of productivity analysis, are not effectively addressed or exhibited in the macro-oriented Basu-Fernald approach. In addition, it will not surprise anyone who knows of my own work over the past twenty years or so that I think this paper in a sense "reinvents the wheel." The paper raises numerous issues of technological,

Catherine J. Morrison Paul is professor of economics at the University of California, Davis.

market, and cross-market structure that are not novel, but have been addressed in various perspectives in a number of literatures and for a long time.

However, I must say that the macro literature on which the paper builds has been critical for both theoretical and conceptual development of productivity analysis. Also, the issues addressed are indeed important, and the many literatures in which these issues have been raised are not yet integrated. Perhaps this meeting may be used as a forum in which these different types of perspectives may begin to be aired and linked more effectively.

The different perspectives building the existing foundations of productivity analysis have been raised in the microproduction theory (parametric and nonparametric), efficiency, macro, "new growth," and "new IO" literatures, most of which have at least some representation in this gathering. The varying perspectives facilitate the creation of additional insights over any one viewpoint, and they are at least starting to converge.

The Basu and Fernald paper is an important case in point because it (implicitly) recognizes the importance of the microproduction theory literature to the macro issues addressed. It at least gives lip service to the basis of the theory of the firm—although the model developed is not really used for analysis—and recognizes the usefulness of a "bottom-up" approach. However, rather than building on the existing micro foundation, or synthesizing the different perspectives, this treatment primarily sweeps the existing literature under the rug, so rather than working together as the complements they have the potential to be, the different literatures become like two ships passing in the night. Perhaps a better analogy, given the contentious competition that sometimes rears its head between opposing camps, might be that of the *Andrea Doria,* where two massive ships crashed in thick fog without seeing each other.

So what is "new" here? Not much, I think. Important seminal work by Solow and later elaborations by Hall are extended to recognize issues raised already in the macro, as well as the micro, foundations literature. Bringing them together is a useful exercise, but I'm not sure exactly what we learn from it. I *am* left, however, contemplating a number of distinctions they make that are useful to think about in the context of the existing micro foundations literature, and the gap between this and the macro treatments. These distinctions will take the form of seven general points I wish to raise.

First-versus Second-Order Analysis

The authors emphasize that their analysis focuses on first-order effects, because that is mainly what is of interest for macroeconomic applications. However, most of the intriguing issues focused on in productivity analysis—including those raised in the paper about utilization, "biases," scale

effects, externalities, and spillovers—are based fundamentally on second-order relationships.

For example, utilization has to do with over- or underuse of existing stocks of capital and labor (or increases or decreases in the service flow of these stocks) by differential use of substitutable and complementary inputs. This is a second-order phenomenon. Biases in input use (or output production if reallocation among outputs occurs) also have to do with second-order effects (although in the macro literature, biases are often raised in the context of statistical biases, rather than real biases with respect to technology or market valuation stemming from technological and market forces). Recognizing these relationships, as in the microproduction theory literature overviewed by Nadiri and Prucha in this volume, allows structural modeling and separate identification of their impacts rather than relying on instruments, proxies, and control variables to "back-out" these relationships. Without this, little interpretation of the results is forthcoming.

Top-Down versus Bottom-Up

The authors indicate that macro questions require a top-down perspective but also motivate their analysis via a bottom-up approach, which in this case means aggregating over two-digit industries.

I have sympathy for the industry-oriented approach, although I would have preferred it to stem from something like the four-digit level, which represents much more homogeneous industry divisions. A true micro (say, plant-level) approach sometimes tends to get lost in the immense heterogeneity within even the most homogeneous divisions (as is evident from the Ellerman, Stoker, and Berndt paper in this volume). Also, typical questions of interest about patterns observed within and across industries require an industry focus for analysis. Thus, beginning with an industry-level micro perspective has its merits. Again, though, more structured analysis of the technological and market structure is important. The simple first-order analysis here glosses over the determinants of utilization, scale effects, and other technological and market conditions of interest in the simple average relationships. Once these patterns are determined at the industry level, bottom-up analysis means to me that they are summarized across industries to obtain insights about overall patterns, rather than just lumped into one, or a very limited number of, parameters.

Mismeasurement versus Mismodeling

I am not speaking here of the data mismeasurement issues raised by Triplett and others, which focus on quality. However, one aspect of the quality issue that may be addressed in a more complete microproduction theory structural model that allows consideration of differential input and output patterns is (second-order) changes in input and output com-

position. In this context, for example, changes in the proportion of high-tech capital, or educational attainment levels for labor, as well as trends in output mix, may be incorporated.

The main question I am raising about mismeasurement instead has to do with the literature in which the Basu-Fernald piece falls, which refers to mismeasurement in terms of distortions. That is, productivity is considered a combination of technology and distortions. However, much of the technological and market structure underlying these distortions, such as utilization and scale patterns, can be identified separately in a microproduction theory structural model, which is in fact represented by the theory-of-the-firm model in their paper. This could potentially facilitate the interpretation and use of measures of these production characteristics, instead of collapsing them together as mismeasurement, as is done in the implementation for this macro treatment.

For example, basic micro theory provides us information about the distinction between capital stock and service flow resulting from fixity of the capital factors and thus fluctuations in the intensity of their use. Therefore, the resulting input use patterns may be evaluated in terms of this conceptual framework, rather than mismeasurement or a "measure of our ignorance." This in turn allows these impacts to be separated from what is left of the measure of our ignorance, which productivity analysis is designed to illuminate. This relates also to the next distinction.

Primal versus Dual

Although there are some conceptual differences between the (first-order) variable utilization and (second-order) capacity utilization concepts, they are inextricably linked. The authors suggest that the first concept focuses on the distinction between service flow and stock, and the latter on valuing the stock. However, in both cases the underlying question is the service flow. In the first case, it relies on a more primal notion—revising the measure of the capital or labor level to correspond to its service flow. In the second case it is somewhat more of a nuance. Additional services or effort from a given stock of capital (or labor) results from more intensive application of other inputs, which in turn affects its marginal valuation in terms of other inputs.

That is, if, given input prices, greater output demand causes more capital effort to be expended, this raises the amount of other inputs applied to production from a given amount of capital stock, increasing its marginal valuation. This revaluation embeds the utilization issue in the dual or price term of the price times quantity "total value" of the stock, rather than directly in the quantity measure. But the effect is the same: greater utilization implies a higher capital share. This primal/dual distinction allows different perspectives on the issue, but they are essentially mirror images. The dual perspective also provides a structure in which shadow values (of

both inputs and outputs), which are alluded to many times in the Basu-Fernald paper, may be measured and analyzed, whereas this is not possible in the (first-order) primal model.

Technical versus Market Structure

In the Basu-Fernald treatment, the notions of imperfect competition and scale economies are often used nearly interchangeably. However, our micro structure, again learned from principles classes, emphasizes the different motivations and thus interpretations resulting from these production structure characteristics. Scale economies, which arise from technology, may be good in the sense of cost efficiency. Effective representation of the cost structure is therefore crucial for appropriate analysis of these effects. Market imperfections, which arise from the output demand structure, may be bad in the context of resource allocation. Recognition of the market structure, and what might be driving evidence of market power, is therefore important for justifiable interpretation and use of "markup" measures. They need to be separately distinguished, identified, and analyzed, which is not possible here.

External versus Internal Effects

Many production and market characteristics mentioned in the productivity literature, raised in this conference, and alluded to in the Basu-Fernald piece may have external or spillover, as well as internal, effects. These include, but are not restricted to, R&D, trade (import penetration or competition), high-tech capital investment (capital composition), and education (labor composition). They also may more generally be characterized as agglomeration effects. These externalities may generate scale effects, as emphasized by the endogenous growth literature. However, these external effects may not be disentangled within the simple first-order model used in the Basu-Fernald paper, and thus again collapse to just another component of the distortions, mismeasurement, or measure of our ignorance captured in the residual. Appropriate characterization of these effects must recognize their impact on the cost-output relationship, as external shift factors. Note also that the notion of the internalization of these effects at more macro levels of aggregation has a clear representation in this context; spillovers that are external at low aggregation levels will be internal at high levels.

Welfare versus Productivity versus Technical Change

This final distinction is a crucial one in terms of interpretation of measures. Although addressed by Basu and Fernald, I believe it again requires a microfoundations perspective for appropriate representation, because the different types of distortions or components of the productivity growth measure independent of technical change may be separately distinguished

in such a framework. For example, returning to the scale economy/market power distinction, these different production and market structure characteristics have widely varying connotations in terms of welfare analysis and thus policy implications, which thus require separate identification for evaluation of welfare implications.

Therefore, many technological, market, and cross-market issues raised by Basu and Fernald, and in the many linked micro- and macroproductivity studies in this area, seem to require a more detailed structural microfoundations emphasis for effective implementation, interpretation, and use. The different components of the Solow residual ("measure of our ignorance") need to be independently captured, identified, and unraveled. There are disadvantages of a more complex approach to modeling these relationships for more microunits, and ultimately summarizing them across industries to obtain macro implications, because implementation is more complicated. However, such an exercise can provide useful guidance, or at least underpinnings, for the first-order, top-down macro perspective of these issues overviewed in the Basu and Fernald piece. Hopefully, these views or perspectives are converging. The complementary insights provided by the different approaches should be investigated and synthesized.