1

Total Factor Productivity
A Short Biography

Charles R. Hulten

1.1 Introduction

Colonial Americans were very poor by today's standard of poverty. On the eve of the American Revolution, GDP per capita in the United States stood at approximately $765 (in 1992 dollars).[1] Incomes rose dramatically over the next two centuries, propelled upward by the Industrial Revolution, and by 1997, GDP per capita had grown to $26,847. This growth was not always smooth (see fig. 1.1 and table 1.1), but it has been persistent at an average annual growth rate of 1.7 percent. Moreover, the transformation wrought by the Industrial Revolution moved Americans off the farm to jobs in the manufacturing and (increasingly) the service sectors of the economy.

Understanding this great transformation is one of the basic goals of economic research. Theorists have responded with a variety of models. Marxian and neoclassical theories of growth assign the greatest weight to productivity improvements driven by advances in the technology and the organization of production. On the other hand, the New Growth Theory and another branch of neoclassical economics—the theory of capital and investment—attach primary significance to the increase in investments in human capital, knowledge, and fixed capital.

The dichotomy between technology and capital formation carries over to empirical growth analysis. Generally speaking, the empirical growth

Charles R. Hulten is professor of economics at the University of Maryland, a research associate of the National Bureau of Economic Research, and chairman of the Conference on Research in Income and Wealth.

1. Estimates of real GDP per capita and TFP referred to in this section are pieced together from Gallman (1987), *Historical Statistics of the United States, Colonial Times to 1970,* and the 1998 *Economic Report of the President.*

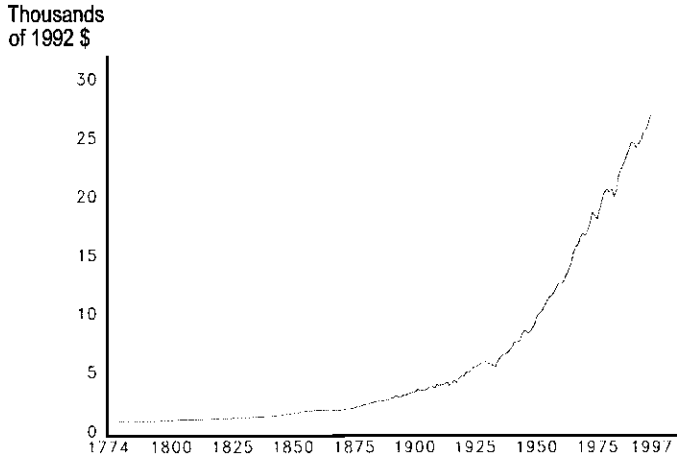

Fig. 1.1 Real GNP/GDP per capita in the United States.

economist has had two main tasks: first, to undertake the enormous job of constructing historical data on inputs and outputs; and second, to measure the degree to which output growth is, in fact, due to technological factors ("productivity") versus capital formation. This last undertaking is sometimes called "sources of growth analysis" and is the intellectual framework of the TFP residual, which is the organizing concept of this survey.

A vast empirical literature has attempted to sort out the capital technology dichotomy, an example of which is shown in table 1.2, but no clear consensus has emerged. Many early studies favored productivity as the main explanation of output growth (see Griliches 1996), and this view continues in the "official" productivity statistics produced by the U.S. Bureau of Labor Statistics (BLS). However, Jorgenson and Griliches (1967) famously disagreed, and their alternative view finds support in subsequent work (e.g., Young 1995) and in the New Growth literature.

In recent years, attention has turned to another issue: the slowdown in productivity that started in the late 1960s or early 1970s. This issue has never been resolved satisfactorily, despite significant research efforts. This, in turn, has been supplanted by yet another mystery: Why has the widely touted information revolution not reversed the productivity slowdown? In a review in the *New York Times* (12 July 1987, p. 36), Robert Solow puts the proposition succinctly: "We can see the computer age everywhere but in the productivity statistics." Recent research seems to have located some of the missing effect (Oliner and Sichel 2000; Jorgenson and Stiroh 2000) as the productivity pickup of the late 1990s has correlated well with the IT revolution. However, Nordhaus (1997) reminds us that the "Solow Par-

Table 1.1 **Historical Growth Rates of Output per Person and Total Factor Productivity in the United States (by decade)**

	Real GNP/GDP per Capita	TFP	Contribution of TFP (percent)
1779–1789	−0.002	n.a.	
1789–1799	−0.008	n.a.	
1799–1809	0.007	0.006	73.5
1809–1819	−0.009	0.006	64.4
1819–1829	0.008	0.006	69.7
1829–1839	0.012	0.006	44.0
1839–1849	0.018	0.007	38.4
1849–1859	0.016	0.007	45.1
1859–1869	0.004	0.007	161.7
1869–1879	0.023	0.007	30.7
1879–1889	0.017	0.007	42.7
1889–1899	0.023	0.003	12.6
1899–1909	0.018	0.002	13.5
1909–1919	0.019	0.003	16.3
1919–1929	0.024	0.002	7.7
1929–1939	0.016	0.003	16.6
1939–1949	0.026	0.003	9.6
1949–1959	0.034	0.002	6.2
1959–1969	0.027	0.003	12.0
1969–1979	0.023	n.a.	
1979–1989	0.017	n.a.	
1989–1997	0.009	n.a.	
1799–1979	0.018	0.005	26.0
	Private Business Economy Only		
1948–1973	0.033	0.021	64
1973–1979	0.013	0.006	46
1979–1990	0.012	0.002	17
1990–1996	0.011	0.003	27
1948–1996	0.023	0.012	52

Sources: Gallman (1987), U.S. Department of Commerce, Bureau of the Census (1975), and the 1998 *Economic Report of the President.* Data for "Private Business Economy Only" are from the Bureau of Labor Statistics, miscellaneous press releases subsequent to Bulletin 2178 (1983).

Note: n.a. = not available.

adox" is not limited to computers. Based on his study of the history of lighting, he argues that official price and output data "miss the most important technological revolutions in economic history" (Nordhaus 1997, 54). Moreover, the Advisory Commission to Study the Consumer Price Index (1996) assigns an upward bias of 0.6 percentage points per year in the Consumer Price Index (CPI) as a result of missed quality improvement, with a corresponding understatement of quantity.

In this New Economy critique of productivity statistics, the growth path

Table 1.2 **Sources of Growth in the U.S. Private Business Sector (selected intervals)**

	Real Output	Labor Input	Capital Services	TFP
1948–1996	3.4	1.4	3.7	1.2
1948–1973	4.0	1.0	3.8	2.1
1973–1996	2.7	1.9	3.5	0.3
1973–1979	3.1	1.8	4.1	0.6
1979–1990	2.7	2.0	3.8	0.2
1990–1996	2.4	1.9	2.5	0.3

Source: Bureau of Labor Statistics miscellaneous press releases subsequent to Bulletin 2178 (1983).

evident in figure 1.1, impressive as it may seem, seriously understates the true gains in output per person occurring over the last two centuries. However, there is another New Economy paradox that has been largely overlooked: If the missed quality change is of the magnitude suggested by the figure, then the quality of the goods in past centuries—and the implied standard of living—must have been much lower than implied by official (and allegedly quality-biased) statistics (Hulten 1997). Indeed, taken to its logical extreme, the correction of figure 1.1 for quality bias would result in a quality-adjusted average income in 1774 that is dubiously small.[2]

A second line of attack on the New Economy view comes from environmentalists, who argue that GDP growth overstates the true improvement in economic welfare because it fails to measure the depletion of natural resources and the negative spillover externalities associated with rapid GDP growth. This attack has been broadened to include what are asserted to be the unintended consequences of the Industrial Revolution: poverty, urban decay, crime, and loss of core values, among others. This view is represented by a statement that appeared on the cover of *Atlantic Monthly:* "The gross domestic product (GDP) is such a crazy mismeasure of the economy that it portrays disaster as gain" (Cobb, Halstead, and Rowe 1995).

In other words, conventional estimates of productivity growth are either much too large or much too small, depending on one's view of the matter. The truth undoubtedly lies somewhere between the two extremes, but where? This essay attempts to illuminate, if not answer, this question. Its

2. If all prices (not just the CPI prices) grew at a rate that was actually 0.6 percent lower than official price statistics, the corresponding quantity statistics would have an offsetting downward bias. If this bias occurred all the way back to 1774, real GDP per capita would have been $202 in that year, not $765.

first objective is to explain the origins of the growth accounting and productivity methods now under attack. This explanation, a biography of an idea, is intended to show which results can be expected from the productivity framework and which cannot. The ultimate objective is to demonstrate the considerable utility of the idea, as a counterweight to the often erroneous and sometimes harsh criticism to which it has been subjected. The first part of the essay is a critical bibliography of the research works that have defined the field. The second part consists of a somewhat personal tour of recent developments in the field and includes tentative answers to some of the unresolved issues.

1.2 The "Residual": A Critical Bibliography to the Mid-1980s

1.2.1 National Accounting Origins

Output per unit input, or TFP, is not a deeply theoretical concept. It is, in fact, an implicit part of the circular income flow model familiar to students of introductory economic theory. In that model, the product market determines the price, p_t, and quantity, Q_t, of goods and services sold to consumers. The total value of these goods is p_tQ_t dollars, which is equally the expenditure of consumers and the revenue of producers. The factor markets determine the volume of the inputs (labor, L_t, and capital, K_t), as well as the corresponding prices, w_t and r_t. The payment to these inputs, $w_tL_t + r_tK_t$, is a cost to the producer and the gross income of consumers. The two markets are connected by the equality of revenue and cost, on the producer side, and gross income and expenditure, on the consumer side, leading to the fundamental GDP accounting identity

$$p_tQ_t = w_tL_t + r_tK_t. \tag{1}$$

This is, in effect, the budget constraint imposed on an economy with limited resources of capital, labor, and technology.

However, GDP in current prices is clearly an unsatisfactory metric of economic progress. Economic well-being is based on the quantity of goods and services consumed, not on the amount spent on these goods. Because the volume of market activity as measured by equation (1) can change merely because prices have risen or fallen, it can be a misleading indicator of economic progress. What is needed is a parallel accounting identity that records the volume of economic activity that holds the price level constant—that is, a revision of equation (1) using the prices of some baseline year for valuing current output and input.

The construction of a parallel constant-price account is a deceptively simple undertaking. If constant dollar value of output is equal to the con-

stant dollar value of input in any one year, the equality cannot hold in the following year if an improvement in productivity allows more output to be obtained from a given quantity of inputs.[3] To bring the two sides of the constant dollar account into balance, a scaling factor, S_t, is needed. The correct form of the constant-price identity is thus

$$p_0 Q_t = S_t[w_0 L_t + r_0 K_t]. \tag{2}$$

The scaling factor has a value of 1 in the base year 0 but varies over time as the productivity of capital and labor changes. Indeed, if both sides of equation (2) are divided by $w_0 L_t + r_0 K_t$, it is apparent that the scaling factor S_t is the ratio of output to total factor input.

Growth accounting is largely a matter of measuring the variable S_t and using the result to separate the growth of real output into both an input component and a productivity component. Griliches (1996) credits the first mention of the output per unit input index to Copeland (1937), followed by Copeland and Martin (1938). The first empirical implementation of the output per unit input index is attributed to Stigler (1947).

Griliches also observes that Friedman uncovered one of the chronic measurement problems of productivity analysis—the index number problem—in his comment on the research by Copeland and Martin. The problem arises because, with some rearrangement, equation (2) can be shown to be a version of the fixed weight Laspeyres index:

$$\frac{S_t}{S_0} = \frac{\dfrac{Q_t}{Q_0}}{\dfrac{w_0 L_t + r_0 K_t}{w_0 L_0 + r_0 K_0}}. \tag{3}$$

This is a widely used index formula (e.g., the CPI) and was employed in early productivity literature (e.g., Abramovitz 1956). However, the substitution bias of the Laspeyres index is also well known (and was recently pointed out by the Advisory Commission [1996] in its analysis of the CPI). Substitution bias arises when relative prices change and agents (producers or consumers, depending on the context) substitute the relatively cheaper item for the more expensive. The problem can sometimes be reduced by the use of chained (i.e., frequently reweighted) Laspeyres indexes, and both Kendrick (1961) and Denison (1962) endorse the use of chain-indexing procedures, although they primarily use fixed weight procedures.

3. The basic problem is illustrated by the following situation. Suppose that output doubles from one year to the next while labor and capital remain unchanged. If the accounting is done in the constant prices of the first year, the left-hand side of the constant price identity doubles while the right-hand side remains unchanged, violating the adding-up condition.

A more subtle problem arises in the interpretation of the ratio S_t. The basic accounting identities shown in equations (1) and (2) can be read from the standpoint of either the consumer or the producer. Virtually all productivity studies have, however, opted for the producer-side interpretation, as witnessed by terms like "output per unit input" and "total factor productivity." Moreover, discussions about the meaning of S_t have typically invoked the rationale of the production function (see, e.g., the long discussion in Kendrick 1961). However, the consumer-welfare side has lurked in the background. The early literature tended to regard S_t as an indicator of the welfare benefits of innovation, with the consequence that "real" national income, or real net national product, was preferred to output measured gross of real depreciation when calculating the numerator of the TFP ratio.[4] This preference was based on the argument that an increase in gross output might be achieved by accelerating the utilization (and thus deterioration and retirement) of capital, thereby increasing TFP without conveying a long-run benefit to society. This argument had the effect of commingling consumer welfare considerations with supply-side productivity considerations. This introduced a fundamental ambiguity about the nature of the TFP index that has persisted to this very day in a variety of transmuted forms.

1.2.2 The Production Function Approach and the Solow Solution

Solow (1957) was not the first to tie the aggregate production function to productivity. This link goes back at least as far as Tinbergen (1942). However, Solow's seminal contribution lay in the simple, yet elegant, theoretical link that he developed between the production function and the index number approach. Where earlier index number studies had interpreted their results *in light* of a production function, Solow *started* with the production function and deduced the consequences for (and restrictions on) the productivity index. Specifically, he began with an aggregate production function with a Hicksian neutral shift parameter and constant returns to scale:

$$Q_t = A_t F(K_t, L_t). \tag{4}$$

4. The concept of depreciation has been a source of confusion in the productivity and national income accounting literatures, and elsewhere (Hulten 1990; Triplett 1996). Depreciation is a price concept that refers to the loss of capital *value* because of wear, tear, obsolescence, and approaching retirement. The loss of productive capacity as a piece of capital ages is not, strictly speaking, depreciation. The capital stock loses capacity through in-place deterioration and retirement.

The following terminology will be adopted in this paper: The net value of output is the difference between the gross value and depreciation; real net output is the difference between constant-price (real) gross output and a constant-price measure of depreciation; net capital stock is the difference between the gross stock and deterioration.

In this formulation, the Hicksian A_t measures the shift in the production function at given levels of labor and capital. It is almost always identified with "technical change," although this generally is not an appropriate interpretation.[5]

Once the production function is written this way, it is clear that the Hicksian A_t and the ratio of output per unit input S_t of the preceding section are related. The terms of the production function can be rearranged to express relative Hicksian efficiency, A_t/A_0, as a ratio with Q_t/Q_0 in the numerator and with the factor accumulation portion of the production function, $F(K_t, L_t)/F(K_0, L_0)$, in the denominator. The indexes A_t and S_t are identical in special cases, but A_t is the more general indicator of output per unit input (TFP). In the vocabulary of index number theory, the Laspeyres S_t is generally subject to substitution bias.

Solow then addressed the key question of measuring A_t using a nonparametric index number approach (i.e., an approach that does not impose a specific form on the production function). The solution was based on the total (logarithmic) differential of the production function:

$$\frac{\dot{Q}_t}{Q_t} = \frac{\partial Q}{\partial K}\frac{K_t}{Q_t}\frac{\dot{K}_t}{K_t} + \frac{\partial Q}{\partial L}\frac{L_t}{Q_t}\frac{\dot{L}_t}{L_t} + \frac{\dot{A}_t}{A_t}. \tag{5}$$

This expression indicates that the growth of real output on the left-hand side can be factored into the growth rates of capital and labor, both weighted by their output elasticities, and the growth rate of the Hicksian efficiency index. The former growth rates represent movements along the production function, whereas the latter growth rate is the shift in the function.

The output elasticities in equation (5) are not directly observable; but if each input is paid the value of its marginal product, that is, if

$$\frac{\partial Q}{\partial K} = \frac{r_t}{p_t} \text{ and } \frac{\partial Q}{\partial L} = \frac{w_t}{p_t}, \tag{6}$$

then relative prices can be substituted for the corresponding marginal products. This, in turn, converts the unobserved output elasticities into observable income shares, s^K and s^L. The total differential in equation (5) then becomes

5. The difference between the Hicksian shift parameter, A_t, and the rate of technical change arises for many reasons. The most important is that the shift parameter captures only costless improvements in the way an economy's resources of labor and capital are transformed into real GDP (the proverbial manna from heaven). Technical change that results from R&D spending will not be captured by A_t *unless* R&D is excluded from L_t and K_t (which it generally is not). A second general reason is that changes in the institutional organization of production will also shift the function, as will systematic changes in worker effort. I will emphasize these and other factors at various points throughout this paper.

$$\Re_t = \frac{\dot{Q}_t}{Q_t} - s_t^K \frac{\dot{K}_t}{K_t} - s_t^L \frac{\dot{L}_t}{L_t} = \frac{\dot{A}_t}{A_t}. \tag{7}$$

$\Re_t$ is the Solow residual—the residual growth rate of output not explained by the growth in inputs. It is a true index number in the sense that it can be computed directly from prices and quantities. The key result of Solow's analysis is that $\Re_t$ is, in theory, equal to the growth rate of the Hicksian efficiency parameter.

This is the theory. In practice, $\Re_t$ is a "measure of our ignorance," as Abramovitz (1956) put it, precisely because $\Re_t$ is a residual. This ignorance covers many components, some wanted (such as the effects of technical and organizational innovation), others unwanted (such as measurement error, omitted variables, aggregation bias, and model misspecification).

1.2.3 A Brief Digression on Sources of Bias

The unwanted parts of the residual might cancel if they were randomly distributed errors, leaving the systematic part of the residual unbiased. However, New Economy and environmentalist complaints arise precisely because the errors are thought to be systematic; these issues are addressed in the second half of this paper. Three other general criticisms will, however, be addressed here, in part because they involve challenges to the basic assumptions of the Solow model, and in part because they inform the evolution of the residual described in the next few sections.

First, there is the view that the Solow model is inextricably linked to the assumption of constant returns to scale. This view presumably originated from the close link between the GDP accounting identity shown in equation (1) and the production function. If the production function happens to exhibit constant returns to scale *and* if the inputs are paid the value of their marginal products as in equation (6), then the value of output equals the sum of the input values. This "product exhaustion" follows from Euler's Theorem and implies that the value shares, s^K and s^L, sum to 1. However, there is nothing in the sequence of steps in equations (4) to (7), leading from the production function to the residual, that requires constant returns (see Hulten 1973). Constant returns are actually needed for another purpose: to estimate the return to capital as a residual, as per Jorgenson and Griliches (1967). If an independent measure of the return to capital is used in constructing the share weights, the residual can be derived without the assumption of constant returns.

A second general complaint against the residual is that it is married to the assumption of marginal cost pricing (i.e., to the marginal productivity conditions shown in equation [6]). When imperfect competition leads to a price greater than marginal cost, Hall (1988) shows that the residual yields a biased estimate of the Hicksian shift parameter, A_t. There is, unfortunately, no way around this problem within the index number approach

proposed by Solow, which is by nature nonparametric, meaning that it produces estimates of A_t directly from prices and quantities. The essence of the Solow method is to use prices to estimate the slopes of the production function at the observed input-output configurations, without having to estimate the shape of the function at all other points (i.e., without the need to estimate all the parameters of the technology). The residual is thus a parsimonious method for getting at the shift in the production function, but the price of parsimony is the need to use prices as surrogates for marginal products.

A third issue concerns the implied nature of technical change. In general, the Hicksian formulation of the production function shown in equation (7) is valid if innovation improves the marginal productivity of all inputs equally. In this case, the production function shifts by the same proportion at all combinations of labor and capital. This is clearly a strong assumption that may well lead to biases if violated. A more general formulation allows (costless) improvements in technology to augment the marginal productivity of each input separately:

$$(4') \qquad Q_t = F(a_t K_t, b_t L_t).$$

This is the "factor augmentation" formulation of technology. It replaces the Hicksian A_t with two augmentation parameters, a_t and b_t. If all the other assumptions of the Solow derivation are retained, a little algebra shows that the residual can be expressed as

$$(7') \qquad \Re_t = s_t^K \frac{\dot{a}_t}{a_t} + s_t^L \frac{\dot{b}_t}{b_t}.$$

The residual is now the share weighted average of the rates of factor augmentation, but it still measures changes in TFP. Indeed, when the rates of factor augmentation are equal, and the sum of the shares is constant, we effectively return to the previous Hicksian case.

Problems may arise if the rates of factor augmentation are not equal. In this situation, termed "Hicks-biased technical change," it is evident that productivity growth depends on the input shares as well as on the parameters of innovation. A change in the income shares can cause output per unit input (TFP) to increase, even if the underlying rate of technical change remains unchanged. This reinforces the basic point that productivity growth is not the same thing as technical change.

Some observers have concluded that the bias in technical change translates into a measurement bias in the residual. This is true only if one insists on identifying TFP with technical change. However, the productivity residual does not get off free and clear: Factor-biased technical change may not lead to measurement, but it does generally lead to the problem of path dependence, discussed in the following section.

1.2.4 The Potential Function Theorem

Solow's derivation of the residual deduces the appropriate index number formulation from the production function and, as a by-product, shows that it is not the Laspeyres form. But what type of index number is it? It was soon noted that equation (7) is the growth rate of a Divisia index (e.g., Richter 1966), a continuous-time index related to the discrete-time chain index mentioned above. This linkage is important because it allows Solow's continuous formulation to be implemented using discrete-time data, while preserving the theoretical interpretation of the residual as the continuous shift in an aggregate production function.

However, this practical linkage has one potential flaw. Solow showed that the production function shown in equation (4) and the marginal productivity conditions shown in equation (6) lead to the growth rate form in equation (7). He did *not* show that a researcher who starts with equation (7) will necessarily get back to the shift term A_t in the production function. Without such a proof, it is possible that the calculation in equation (7) could lead somewhere besides A_t, thus robbing the index of its conventional interpretation.

This issue was addressed in my 1973 paper, which shows that the Solow conditions are both necessary and sufficient. The expression in equation (7) yields a unique index only if there is a production function (more generally, a potential function) whose partial derivatives are equal to the prices used to compute the index. The production function (cum potential function) is the integrating factor needed to guarantee a solution to equation (7), which is in fact a differential equation. If there is no production function, or if it is nonhomothetic, the differential equation (7) cannot be (line) integrated to a unique solution. This problem is called "path dependence."[6]

The Potential Function Theorem imposes a good deal of economic structure on the problem in order to avoid path dependence. Unfortunately, these conditions are easily met. First, aggregation theory demonstrates that the necessary production function exists only under very restrictive assumptions (Fisher 1965), essentially requiring all the micro-production units in the economy (plants, firms, industries) to have production functions that are identical up to some constant multiplier (see also Diewert 1980). If the aggregation conditions fail, a discrete-time ver-

6. The problem of path dependence is illustrated by the following example. Suppose that there is a solution to the Divisia line integral, but only for a particular path of output and input Γ_1 between points A and B. If a different path between these two points, Γ_2, gives a different value, then path dependence arises. If the Divisia index starts with a value of 100 at point A and the economy subsequently moves from A to B along Γ_1, and then back to A along Γ_1, the Divisia index will not return to 100 at A. Because the path can cycle between A and B along these paths, the index can, in principle, have a purely arbitrary value.

sion of the Divisia index might still be cobbled together, but the resulting numbers would have no unique link to the efficiency index A_t. Indeed, the theoretical meaning of A_t itself is ambiguous if the aggregation conditions fail.[7]

When the Divisia index is path independent, Solow's procedures yield an estimate of the productivity residual that is uniquely associated with the shift in the production function. This result carries the important implication that the residual must be given a capacity interpretation, in this case, rather than a welfare interpretation. Or, more accurately, any welfare interpretation must be consistent with this main interpretation.

The Potential Function Theorem also sheds light on the debate over net versus gross measures of output and capital. The theorem requires the units of output or input selected for use in equation (7) to be consistent with the form of the production function used as the integrating factor. To choose net output for computing the Solow residual, for example, is to assert that the production process generates net output from capital and labor, and that factor prices are equal to the *net* value of marginal product rather than to the gross value of standard theory. This is an unusual view of real-world production processes, because workers and machines actually make gross units of output and the units of output emerging from the factory door are not adjusted for depreciation. Nor do we observe a price quoted for net output. Similar reasoning leads to the use of a net-of-deterioration concept of capital.

1.2.5 Jorgenson and Griliches versus Denison

The 1967 paper by Jorgenson and Griliches is a major milestone in the evolution of productivity theory. It advanced the hypothesis that careful measurement of the relevant variables should cause the Solow measure of total factor productivity to disappear. This is an intellectually appealing idea, given that the TFP index is a residual "measure of our ignorance." Careful measurement and correct model specification should rid the residual of unwanted components and explain the wanted ones.

Jorgenson and Griliches then proceeded to introduce a number of measurement innovations into the Solow framework, based on a strict application of the neoclassical theory of production. When the renovations were

7. Path dependence also rises if the aggregate production function exists but fails to satisfy any of the basic assumptions: namely, marginal productivity pricing, constant returns to scale, and Hicksian technical change. This statement must, however, be qualified by the remarks of the preceding section. If an independent estimate of the return of capital is used when constructing the share weight of capital, s^K, then the Divisia productivity index is path independent even under nonconstant returns to scale (Hulten 1973). Moreover, if costless technical change is Harrod neutral, line integration of the residual $\Re$ is subject to path dependence, but integration of the ratio $\Re/s^L$ is not, and leads to a path-*in*dependent index of the labor augmentation parameter, b_t, shown in equation (7′). The Divisia residual is more versatile than is commonly believed.

complete, they found that the residual had all but disappeared. This result stood in stark contrast to contemporary results, which found that the residual did make a sizeable contribution to economic growth. However, this attack on (and, indeed, inversion of) the conventional wisdom was answered by Denison, whose own results were consistent with the prevailing wisdom.

Denison (1972) compared his procedures with those of Jorgenson and Griliches and found that part of the divergence was caused by a difference in the time periods covered by the two studies and that another part was due to a capacity-utilization adjustment based on electricity use. The latter indicated a secular increase between equivalent years in the business cycle; and when this was removed, and the two studies put in the same time frame, Denison found that the Jorgenson-Griliches residual was far from zero.

The debate between Denison (1972) and Jorgenson and Griliches (1967, 1972) focused attention on the bottom line of empirical growth analysis: how much output growth can be explained by total factor productivity (manna from heaven), and how much had to be paid for capital formation. However, the debate obscured the true contribution of the Jorgenson-Griliches study, which was to cement the one-to-one link between production theory and growth accounting. For Solow, the aggregate production function was a parable for the measurement of TFP; for Jorgenson and Griliches it was the blueprint. Implementing this blueprint led to a number of important innovations in the Solow residual—a sort of productivity improvement in the TFP model itself.

One of the principal innovations was to incorporate the neoclassical investment theory developed in Jorgenson (1963) into productivity analysis. The first step was to recognize that the value of output in the accounting identity in equation (1) is the sum of two components: the value of consumption goods produced, $p^C C$, and the value of investment goods produced, $p^I I$ (hence, $pQ = p^C C + p^I I = wL + rK$). The price of the investment good was then assumed to be equal to the present value of the rents generated by the investment (with an adjustment for the depreciation of capital). This present value is then solved to yield an expression for the user cost of capital, $r = (i + \delta)P^I - \Delta P^I$. The problem, then, is to find a way of measuring r or its components. Direct estimates of the user cost are available for only a small fraction of the universe of capital goods (those that are rented). The alternative is to estimate the components of r. The investment good price, P^I, can be obtained from national accounts data, and the depreciation rate, δ, can be based on the Hulten-Wykoff (1981) depreciation study. The rate of return, i, can be estimated in two ways. First, it can be estimated independently from interest rate or equity return data. This is somewhat problematic because of the multiplicity of candidates and the need to pick a rate that reflects the risk and opportunity

cost of the capital good. Jorgenson and Griliches suggest a second way: Impose constant returns to scale and find the implied i that causes the accounting equation $pQ = wL + rK$ to hold.[8] It is only at this point that constant returns are required for the measurement of TFP.

The quantity of capital, K_t, and the quantity of new investment, I_t, are connected (in this framework) by the perpetual inventory method, in which the stock is the sum of past investments adjusted for deterioration and retirement. The resulting concept of capital is thus defined net of deterioration, in contrast with the concept of undeteriorated "gross" stock used in some studies.

On the other hand, Jorgenson and Griliches recognized that output must be measured gross of depreciation if it is to conform to the accounting system implied by the strict logic of production theory. This put them in conflict with Denison, who advocated a concept of output net of depreciation, and Solow, who used gross output in his empirical work but preferred net output on the theoretical grounds that it is a better measure of welfare improvement arising from technical progress. The debate over this point with Denison thus seemed to pivot on the research objective of the study, not on technical grounds. However, as we have seen, the Potential Function Theorem, published after the 1967 Jorgenson and Griliches study, links their gross output approach to the A_t of conventional production theory, implying that the competing views of output cannot be simultaneously true (except in very special cases).

Another major contribution of the Jorgenson-Griliches study was to disaggregate capital and labor into their component parts, thereby avoiding the aggregation bias associated with internal shifts in the composition of the inputs (e.g., the compositional bias due to a shift from long-lived structures to shorter-lived equipment in the capital stock, or the bias due to the shift toward a more educated work force). The Divisia index framework was applied consistently to the aggregation of the individual types of capital and labor into the corresponding subaggregate, and applied again to arrive at the formulation in equation (7). However, because data are not continuous over time but come in discrete-time units, Jorgenson and Griliches introduced a discrete-time approximation to the Divisia derived from the Törnqvist index.[9]

In sum, Jorgenson and Griliches tied data development, growth ac-

8. The implied value of i is then $[P^C C + P^I I - \delta P^I K - \Delta P^I K]/P^I K$. When there are several types of capital goods, a different δ and P^I is estimated for each type, but arbitrage is assumed to lead to a common i for all assets. Hall and Jorgenson (1967) extended the user cost model to include parameters of the income tax system.

9. In the Tornqvist approximation, the continuous-time income shares s_t^K and s_t^L in equation (7) are replaced by the average between-period shares. Capital's discrete-time income share is $(s^K_t + s^K_{t-1})/2$. Continuous-time growth rates are also replaced with differences in the natural logarithm of the variable. The growth rate of capital, for example, is $\ln(K_t) - \ln(K_{t-1})$.

counting, and production theory firmly together. The three are mutually dependent, not an ascending hierarchy as is commonly supposed. These linkages were developed further by Christensen and Jorgenson (1969, 1970), who developed an entire income, product, and wealth accounting system based on the mutuality principle.

1.2.6 Diewert's Exact and Superlative Index Numbers

The continuous-time theory of the residual developed by Solow provides a simple yet elegant framework for productivity measurement. Unfortunately, data do not come in continuous-time form. One solution, noted earlier, is to find reasonable discrete-time approximations to the continuous-time model. In this approach, the choice among competing approximation methods is based largely on computational expediency, with the implication that the discrete-time approximation is not derived as an organic part of the theory, thereby weakening the link between theory and measurement.

Herein lies the contribution of Diewert (1976). He showed that the Tornqvist approximation to the Divisia index used by Jorgenson and Griliches was an exact index number if the production function shown in equation (4) had the translog form developed by Christensen, Jorgenson, and Lau (1973). In other words, the Tornqvist index was not an approximation at all, but was actually exact under the right conditions. Moreover, because the translog production function could also be regarded as a good second-order approximation to other production functions, the discrete-time Tornqvist index was a sensible choice even if the "world" was not translog. In this event, the degree of exactness in the index number depends on the closeness of the translog function to the true production function. Diewert used the term "superlative" to characterize this aspect of the index.

What Diewert showed, in effect, was that the translog specification of the production function served as a potential function for the discrete Tornqvist index in the same way that the continuous production function served as a potential function for the continuous Divisia index. One important consequence of this result is that the index number approach of the Solow residual is not entirely nonparametric. There is a parametric production function underlying the method of approximation if the discrete-time index is to be an exact measure of Hicksian efficiency. However, the values of the "inessential" parameters of the translog—that is, those other than the Hicksian efficiency parameter—need not be estimated if the Solow residual is used.

1.2.7 Dispelling the "Measure of Our Ignorance" with Econometrics

If a specific functional form of the technology must be assumed in order to obtain an exact estimate of the efficiency parameter, why not go ahead and estimate all the parameters of that function using econometric tech-

niques? That is, why not estimate the translog relation between Q_t, K_t, L_t, and A_t directly? For one thing, this avoids the need to impose the marginal productivity conditions of the index number approach.[10] Moreover, it gives a full representation of the technology: all the parameters (not just the efficiency term), and every possible path (not just the path actually followed). Moreover, noncompetitive pricing behavior, nonconstant returns, and factor-augmenting technical change can be accommodated, and embellishments like cost-of-adjustment parameters can be incorporated into the analysis to help "explain" the residual. Why settle for less when so much more can be obtained under assumptions that must be made anyway—for example, that the production function has a particular functional form like the translog?

The answers to these questions are familiar to practitioners of the productivity art. There are pitfalls in the econometric approach, just as there are with nonparametric procedures. For example, estimation of the translog (or another flexible) function can lead to parameter estimates that imply oddly shaped isoquants, causing practitioners to place a priori restrictions on the values of these parameters. There is often a question about the robustness of the resulting parameter estimates to alternative ways of imposing restrictions. Even with these restrictions, the abundance of parameters can press on the number of data observations, requiring further restrictions. Additionally, there is the question of the econometric procedures used to obtain the estimates. The highly complicated structure of flexible models usually requires nonlinear estimation techniques, which are valid only under special assumptions, and there are questions about the statistical properties of the resulting estimates. Finally, because the capital and labor variables on the right-hand side of the regression depend in part on the output variable on the left-hand side, there is the danger of simultaneous equations bias.

In other words, the benefits of the parametric approach are purchased at a cost. It is pointless to debate whether benefits outweigh those costs, simply because there is no reason that the two approaches should be viewed as competitors. In the first place, the output and input data used in the econometric approach are almost always index numbers themselves (there are simply too many types of output and input to estimate separately). Thus, the question of whether or when to use econometrics to measure productivity change is really a question of the stage of the analysis at which index number procedures should be abandoned. Secondly, there is no reason for there to be an either-or choice. Both approaches can be

10. The marginal productivity conditions can be avoided in the direct estimation of the production function. However, the marginal productivity conditions are used in the estimation of the "dual" cost and profit functions that form an essential part of the productivity econometrician's tool kit.

implemented simultaneously, thereby exploiting the relative simplicity and transparency of the nonparametric estimates to serve as a benchmark for interpreting the more complicated results of the parametric approach. The joint approach has an added advantage of forcing the analyst to summarize the parameters of the translog (or other) function in a way that illuminates their significance for TFP growth (i.e., for the dichotomy between the shift in the production function and factor-driven movements along the function).

Moreover, by merging the two approaches, econometrics can be used to disaggregate the TFP residual into terms corresponding to increasing returns to scale, the cost of adjusting the factor inputs, technical innovation, an unclassified trend productivity, and measurement error. Denny, Fuss, and Waverman (1981) were the first to start down this path, and it has grown in importance in recent years. The power of this approach is illustrated by the 1981 paper of Prucha and Nadiri on the U.S. electrical machinery industry. Their version of the TFP residual grew at an average annual rate of 1.99 percent in this industry from 1960 to 1980. Of this amount, 35 percent was attributed to technical innovations, 42 percent to scale economies, and 21 percent to adjustment cost factors, with only 2 percent left unexplained.

This development addresses the measure-of-our-ignorance problem posed by Abramovitz. It also provides a theoretically rigorous alternative to Denison, who attempted to explain the residual with informed guesses and assumptions that were above and beyond the procedures used to construct his estimates of the residual. It also speaks to the Jorgenson-Griliches hypothesis that the residual ought to vanish if all explanatory factors can be measured.

1.2.8 Digression on Research and Development Expenditures

Another contribution made by Jorgenson and Griliches (1967) was their recognition that aggregate measures of capital and labor included the inputs used in research and development programs to generate technical innovations. Thus, some part of the rate of innovation that drove the TFP residual was already accounted for in the data. As a result, if the social rate of return to the R&D expenditures buried in the input data is equal to the private return, the effect of R&D would be fully accounted for, and the innovation component of the residual should disappear. On the other hand, if there is a wedge between the social and private rates of return, then the innovation component of the residual should reflect the externality. This is a harbinger of the New Growth Theory view of endogenous technical innovation.

The important task of incorporating R&D expenditures explicitly into the growth accounting framework has, unfortunately, met with limited success. Griliches (1988) pointed out a key problem: Direct R&D spending is

essentially an internal investment to the firm, with no observable "asset" price associated with the investment "good" and no observable income stream associated with the stock of R&D capital. As a result, there is no ready estimate of the quantity of knowledge capital or its growth rate, nor of the corresponding share weight, which are needed to construct a Divisia index. Moreover, much of the R&D effort of any private firm goes toward improving the quality of the firm's products, not the productivity of its production process (more on this later).

There is, of course, a huge literature on R&D and the structure of production, but it is almost entirely an econometric literature (see Nadiri 1993 and Griliches 1994 for reviews). A satisfactory account of this literature is well beyond the scope of a biography of the nonparametric residual.

1.2.9 The Comparison of Productivity Levels

The TFP residual defined earlier is expressed as a rate of growth. The TFP growth rate is of interest for intertemporal comparisons of productivity for a given country or region at different points in time, but it is far less useful for comparing the relative productivity of different countries or regions. A developing country may, for example, have a much more rapid growth in TFP than a developed country, but start from a much lower level. Indeed, a developing country may have a more rapid growth in TFP than a developed country *because* it starts from a lower level and is able to import technology. This possibility is discussed in the huge literature on convergence theory.

The first translog nonparametric estimates of TFP levels were developed by Jorgenson and Nishimizu (1978) for the comparison of two countries. This innovation was followed by an extension of the framework to include the comparison of several countries simultaneously by Christensen, Cummings, and Jorgenson (1981) and Caves, Christensen, and Diewert (1982a). Moreover, in a contemporaneous paper, Caves, Christensen, and Diewert (1982b) apply a different approach—the Malmquist index—to the comparison of relative productivity levels.

The Malmquist index asks simple questions: How much output could country A produce if it used country B's technology with its own inputs? How much output could country B produce if it used country A's technology with its inputs? The Malmquist productivity index is the geometric means of the answers to these two questions. If, for example, the output of country A would be cut in half if it were forced to use the other country's technology, while output in country B would double, the Malmquist index would show that A's technology is twice as productive.[11] When the produc-

11. Formally, let $Q_A = F(X_A)$ be the production function in country A and $Q_B = G(X_B)$ in country B. The Malmquist approach estimates how much output Q_A^* would have been produced in A if the technology of B had been applied to A's inputs; that is, $Q_A^* = G(X_A)$. The ratio Q_A/Q_A^* is then a measure of how much more (or less) productive is technology A com-

tion functions differ only by the Hicks-neutral efficiency index, A_A and A_B, respectively, the Malmquist index gives the ratio A_A/A_B. This is essentially the Solow result in a different guise. Moreover, when the technology has the translog form, Caves, Christensen, and Diewert (1982b) show that the Tornqvist and Malmquist approaches yield the same result.

However, the two approaches may differ if efficiency differences are not Hicks neutral or if there are increasing returns to scale. In these situations, the relative level of technical efficiency will depend on the input levels at which the comparison is made. If, by some chance, other input levels had occurred, the Malmquist index would have registered a different value, even though the production functions in countries A and B were unchanged. This is the essence of the path dependence problem in index number theory.

Malmquist indexes have been used in productivity measurement mainly in the context of nonparametric frontier analysis (e.g., Färe et al. 1994). Frontier analysis is based on the notion of a best-practice level of technical efficiency that cannot be exceeded, and which might not be attained. An economy (or industry or firm) may be below its best-practice level for a variety of reasons: obsolete technology, poor management, constraints on the use of resources, and so on. A measured change in the level of efficiency may therefore reflect an improvement in the best-practice technology or in the management of the prevailing technology. Sorting out which is which is an important problem in productivity analysis.

Frontier analysis tackles this problem by using linear programming techniques to "envelope" the data and thereby locate the best-practice frontier. The main advantages of frontier analysis are worth emphasizing. First, frontier techniques allow the observed change in TFP to be resolved into changes in the best-practice frontier and in the degree of inefficiency. Second, the technique is particularly useful when there are multiple outputs, some of whose prices cannot be observed (as when, for example, negative externalities such as pollution are produced jointly with output). The principal drawback arises from the possibility that measurement errors may lead to data that are located beyond the true best-practice frontier. There is a danger that the outliers will be mistakenly enveloped by frontier techniques (though stochastic procedures may help here), resulting in an erroneous best-practice frontier.

1.2.10 Capital Stocks and Capacity Utilization

Production functions are normally defined as a relation between the flow of output on the one hand, and the flows of capital and labor services on

---

pared to technology B at A's input level. A similar calculation establishes the ratio Q_B^*/Q_B, which measures how much more productive technology B is when compared to that of A at the input level prevailing in country B. The Malmquist index is the geometric mean of the two ratios.

the other. If the residual is to be interpreted as the shift in an aggregate production function, the associated variables must be measured as flows. This is not a problem for output and labor because annual price and quantity data are available. Nor would it be a problem for capital goods if they were rented on an annual basis, in which case there would be little reason to distinguish them from labor input. Capital goods are, however, most often used by their owners. Thus, we typically observe additions to the stock of goods, but not to the stock itself or to the services flowing from the stock. Stocks can be imputed using the perpetual inventory method (the sum of net additions to the stock), but there is no obvious way of getting at the corresponding flow of services.

This would not be a problem if service flows were always proportional to the stock, but proportionality is not a realistic assumption. As economic activity fluctuates over the business cycle, periods of high demand alternate with downturns in demand. Capital stocks are hard to adjust rapidly, so periods of low demand are typically periods of low capital utilization. A residual calculated using capital stock data thus fluctuates procyclically along with the rate of utilization. These fluctuations tend to obscure the movements in the longer-run components of the residual and make it hard to distinguish significant breaks in trend. The dating and analysis of the productivity slowdown of the 1970s form an important case in point.

Jorgenson and Griliches address this problem by adjusting capital stock for a measure of utilization based on fluctuations in electricity use. The form of this adjustment became part of the controversy with Denison, but the real problem lay with the use of any externally imposed measure of capital utilization. Any such measure leads to a theoretical problem: How does a direct measure of capital utilization enter the imputed user cost? Indeed, shouldn't the opportunity cost of unutilized capital be zero?

Berndt and Fuss (1986) provide an answer to these questions. They adopt the Marshallian view that capital stock is a quasi-fixed input in the short run, the income of which is the residual after the current account inputs are paid off. In terms of the fundamental accounting identity, the residual return to capital is $rK = pQ - wL$, where K is the stock of capital (not the flow) and r is the ex post cost of using the stock for one period. Fluctuations in demand over the business cycle cause ex post returns to rise or fall relative to the ex ante user cost on which the original investment was based. The key result of Berndt and Fuss is that the ex post user cost equals the actual (short-run) marginal product of capital, and is thus appropriate for use in computing the TFP residual. Moreover, since the ex post user cost already takes into account fluctuations in demand, no separate adjustment is, in principle, necessary.

On the negative side, it must be recognized that the Berndt-Fuss revisions to the original Solow residual model fail, in practice, to remove the procyclical component of the residual. This failure may arise because

the amended framework does not allow for the entry and exit of firms over the business cycle (and, indeed, is only a partial theory of capital adjustment). Indeed, fluctuations in capital utilization are not just a nuisance factor in productivity measurement, but have an interesting economic life of their own (see Basu and Fernald, chapter 7 in this volume). Additionally, this approach to utilization does not generalize to multiple capital goods. However, the Berndt-Fuss insight into the nature of capital utilization, and its relation to the marginal product of capital, is a major contribution to productivity theory: It clarifies the nature of capital input and illustrates the ad hoc and potentially inconsistent nature of externally imposed utilization adjustments.[12]

1.3 Recent Developments and the Paths Not Taken

The 1980s were a high-water mark for the prestige of the residual, and a watershed for nonparametric productivity analysis as a whole. The Bureau of Labor Statistics (BLS) began publishing multifactor productivity (their name for TFP) estimates in 1983; major contributions also continued outside the government, with the articles already noted and with books by Jorgenson, Gollop, and Fraumeni (1987) and Baumol, Blackman, and Wolff (1989). There has also been an interest in applying growth accounting to explain international differences in growth (e.g., Dowrick and Nguyen 1989); the controversy triggered by Young (1992, 1995); and literature on infrastructure investment inspired by Aschauer (1989). However, the tide had begun to turn against the aggregative nonparametric approach pioneered by Solow, Kendrick, Jorgenson-Griliches, and Denison. Several general trends are discernible:

1. the growing preference for econometric modeling of the factors causing productivity change;
2. the shift in attention from the study of productivity at the aggregate and industry levels of detail to study at the firm and plant levels;
3. a shift in emphasis from the competitive model of industrial organization to noncompetitive models;
4. the effort to endogenize R&D and patenting into the explanation of productivity change; and
5. a growing awareness that improvements in product quality are potentially as important as process-oriented innovation that improve the productivity of capital and labor.

There were several reasons for this shift in focus. The explosion in computing power enabled researchers to assemble and analyze larger sets of

12. The dual approach to the Berndt-Fuss utilization model is explored in Hulten (1986). This papers clarifies the links between average cost, TFP, and the degree of utilization.

data. High-powered computers are so much a part of the current environment that it is hard to remember that much of the seminal empirical work done in the 1950s and early 1960s was done by hand or on mechanical calculating machines (or, later on, by early mainframe computers that were primitive by today's standards). Anyone who has inverted a five-by-five matrix by hand will know why multivariate regressions were not often undertaken. The growth of computing power permitted the estimation of more sophisticated, multiparametered production and cost functions (like the translog) and created a derived demand for large data sets like the U.S. Bureau of Census's Longitudinal Research Database (LRD), which came into play in 1982.

The arrival of the New Growth Theory was a more evident factor behind the shift in the research agenda of productivity analysis. New Growth Theory challenged the constant-returns and perfect-competition assumptions of the TFP residual by offering a view of the world in which (a) markets were noncompetitive; (b) the production function exhibited increasing returns to scale; (c) externalities among microunits were important; and (d) innovation was an endogenous part of the economic system. This shift in perspective gave an added push to the investigation of microdata sets and to the interest in R&D as an endogenous explanation of output growth.

These factors would have sufficed to redirect the research agenda of productivity analysis. However, it was the slowdown in productivity growth, which started sometime between the late 1960s and the 1973 OPEC oil crisis, that settled the matter. Or, more accurately, conventional productivity methods failed to provide a generally accepted explanation for the slowdown, which virtually guaranteed that the assumptions of the conventional analysis would be changed and that explanations would have to be sought elsewhere.[13] The residual was, after all, still the "measure of our ignorance," and the New Growth paradigm and the large-scale microproductivity data sets arrived just in time to fill the demand for their existence.

The directions taken by productivity analysis in recent years are not easy to summarize in a unified way. I will, however, offer some comments on recent developments in the field in the remaining sections. They reflect, to some extent, my own research interests and knowledge, and make no pretense of being an exhaustive survey.

13. The literature on the productivity slowdown is voluminous, and still growing (see, e.g., Denison 1979; Berndt 1980; Griliches 1980; Maddison 1987; Baily and Gordon 1988; Diewert and Fox 1999; and Greenwood and Jovanovic in chap. 6, this vol.). Many different explanations have been offered, from the failure to measure output correctly (particularly in the service sector) to the lag in absorbing and diffusing the IT revolution. No single explanation has decisively vanquished the others; nor has a consensus emerged about the relative importance of the various competing alternatives.

1.4 Productivity in the Context of Macrogrowth Models

1.4.1 The "Old" Growth Theory

The TFP model produces an explanation of economic growth based solely on the production function and the marginal productivity conditions. Thus, it is not a theory of economic growth because it does not explain how variables on the right-hand side of the production function—labor, capital, and technology—evolve over time. However, Solow himself provided an account of this evolution in a separate and slightly earlier paper (1956). He assumed that labor and technology were exogenous factors determined outside the model, and that investment is a constant fraction of output. Then, if technical change is entirely labor augmenting and the production function is well-behaved, the economy converges to a steady-state growth path along which both output per worker and capital per worker grow at the rate of technical change. Cass (1965) and Koopmans (1965) arrive at essentially the same conclusion using different assumptions about the saving-investment process.

Both of these "neoclassical" growth models produce a very different conclusion from that of the TFP model about the importance of technical change in economic growth. In the neoclassical growth models, capital formation explains *none* of the long-run, steady-state growth in output because capital is itself endogenous and driven by technical change: Technical innovation causes output to increase, which increases investment, which thereby induces an expansion in the stock of capital. This induced capital accumulation is the direct result of TFP growth and, in steady-state growth, *all* capital accumulation and output growth are due to TFP. While real-world economies rarely meet the conditions for steady-state growth, the induced-accumulation effect is present outside steady-state conditions whenever the output effects of TFP growth generate a stream of new investment.

What does this mean for the measurement of TFP? The residual is a valid measure of the shift in the production function under the Solow assumptions. However, because the TFP residual model treats all capital formation as a wholly exogenous explanatory factor, it tends to overstate the role of capital and understate the role of innovation in the growth process.[14] Since some part of the observed rate of capital accumulation is a TFP-induced effect, it should be counted along with TFP in any assessment of the impact of innovation on economic growth. Only the fraction of capital accumulation arising from the underlying propensity to invest

14. This was pointed out in Hulten (1975, 1978) in the context of the neoclassical model, and by Rymes (1971) and Cas and Rymes (1991) in a somewhat different context.

at a constant rate of TFP growth should be scored as capital's independent contribution to output growth.[15]

The distinction between the size of the residual on the one hand and its impact on growth on the other has been generally ignored in the productivity literature. This oversight has come back to haunt the debate over "assimilation versus accumulation" as the driving force in economic development. A number of comparative growth studies have found that the great success of the East Asian Tigers was driven mainly by the increase in capital and labor rather than by TFP growth (Young 1992, 1995; Kim and Lau 1994; Nadiri and Kim 1996; Collins and Bosworth 1996). With diminishing marginal returns to capital, the dominant role of capital implies that the East Asian Miracle is not sustainable and must ultimately wind down (Krugman 1994). However, these conclusions do not take into account the induced capital accumulation effect. The role played by TFP growth (assimilation) is actually larger, and the saving/investment effect is proportionately smaller.

Exactly how much larger is hard to say, because the induced-accumulation effect depends on several factors, such as the bias in technical change and the elasticity of substitution between capital and labor. I proposed a correction for this effect in my 1975 paper and estimated that the conventional TFP residual accounted for 34 percent of U.S. output growth over the period 1948 to 1966 (annual output growth was 4.15 percent and the residual was 1.42 percent). When the induced capital accumulation effect formation was taken into account, technical change was actually responsible for 64 percent of the growth in output. This is almost double the percentage of the conventional view of the importance of TFP growth.

A closely related alternative is to use a Harrod-Rymes variant of the TFP residual instead of the conventional Hicksian approach. The Harrodian concept of TFP measures the shift in the production function along a constant capital-output ratio, instead of the constant capital-labor ratio of the conventional Hicks-Solow measure (A_t) of the preceding sections. By holding the capital-output ratio constant when costless innovation occurs, the Harrodian measure attributes part of the observed growth rate of capi-

15. This point can be illustrated by the following example. Suppose that an economy is on a steady-state growth path with a Harrod-neutral rate of technical change of 0.06 percent per year. If capital's income share is one-third of GDP, a conventional TFP sources-of-growth table would record the growth rate of output per worker as 0.06 and allocate 0.02 to capital per worker and 0.04 to TFP. Observed capital formation seems to explain one-third of the growth in output per worker. However, its true contribution is zero in steady-state growth. The 0.06 growth rate of Q/L should be allocated in the following way: 0 to capital per worker and 0.06 to technical change.

A more complicated situation arises when technical change is also embodied in the design of new capital. In this case, the rate of investment affects the rate of technical change and creates a two-way interaction with TFP growth.

tal to the shift in the production function. Only capital accumulation in excess of the growth rate of output is counted as an independent impetus to output growth. The Harrodian approach thus allows for the induced-accumulation effect, and when the innovation happens to be of the Harrod-neutral form, the accounting is exact (Hulten 1975). Otherwise, the Harrodian correction is approximate.

When applied to the East Asian economies studied by Young, the Harrodian correction gives a very different view of the role of TFP growth (Hulten and Srinivasan 1999). Conventional Hicksian TFP accounts for approximately one-third of output growth in Hong Kong, South Korea, and Taiwan over the period 1966–1990/91. With Harrodian TFP, this figure rises to nearly 50 percent. Again, although the conventional Hicksian TFP residual is a valid measure of the shift in the production function, a distinction must be made between the magnitude of the shift and its importance for output growth.

1.4.2 The New Growth Models

Neoclassical growth models assume that innovation is an exogenous process, with the implication that investments in R&D have no systematic and predictable effect on output growth. But, can it really be true that the huge amount of R&D investment made in recent years was undertaken without any expectation of gain? A more plausible approach is to abandon the assumption that the innovation is exogenous to the economic system and to recognize that some part of innovation is, in fact, a form of capital accumulation.

This is precisely the view incorporated in the endogenous growth theory of Romer (1986) and Lucas (1988). The concept of capital is expanded to include knowledge and human capital and is added to conventional fixed capital, thus arriving at total capital. Increments of knowledge are put on an equal footing with all other forms of investment, and therefore the rate of innovation is endogenous to the model. The key point of endogenous growth theory is not, however, that R&D and human capital are important determinants of output growth. What is new in endogenous growth theory is the assumption that the marginal product of (generalized) capital is constant—not diminishing as in the neoclassical theories. It is the diminishing marginal returns to capital that bring about convergence to steady-state growth in the neoclassical theory; and, conversely, it is constant marginal returns that cause the induced-accumulation effect on capital to go on ad infinitum.[16]

Endogenous growth theory encompasses a variety of different models.

16. Barro and Sala-i-Martin (1995) provide a good overview of the various growth models (see also Easterly 1995). Not all relevant models involve increasing returns to scale, since technical change is endogenized by investment in R&D per se.

We will focus here on one that is perhaps the main variant in order to illustrate the implications of endogeneity for the measurement and interpretation of the productivity residual. Suppose that the production function in equation (4) has the Cobb-Douglas production function prevalent in that literature, and that (generalized) capital has two effects: Each 1 percent increase in capital raises the output of its owner-users by β percent but also spills over to other users, raising their output by a collective α percent. Suppose, also, that $\alpha + \beta = 1$, implying constant returns to scale in the capital variable across all producers, while labor and private capital are also subject to constant returns ($\beta + \gamma + 1$). This leads to

$$(8) \qquad Q_t = A_0 K_t^{\alpha}[K_t^{\beta} L_t^{\gamma}], \quad \alpha + \beta = 1, \quad \beta + \gamma = 1.$$

This production function exhibits increasing returns to scale overall, but it is consistent with equilibrium because each producer operates under the assumption of constant returns to the inputs that the producer controls.

What does this new formulation imply for the residual, computed as per the "usual" equation (7)? The residual is derived from the Hicksian production function shown in equation (4), and the formulation in equation (8) is a special case of this function in which the output elasticities are constant (Cobb-Douglas) and the efficiency term $A_0 K_t^{\alpha}$ replaces the Hicksian efficiency parameter A_t. The associated residual, analogous to equation (7), is thus equal to the growth rate of capital weighted by the spillover effect. The endogenous TFP residual continues to measure costless gains to society—the "manna from heaven"—from innovation. But now this manna is associated with the externality parameter α instead of the Hicksian efficiency parameter A_t. Thus, in the New Growth view, the residual is no longer a nonparametric method for estimating a fixed parameter of the production function, but is actually the reflection of a process. Moreover, there is no reason for the residual to disappear.[17]

The endogenous growth residual adds structure to the problem of interpreting the TFP residual, but does this new interpretation help explain the productivity slowdown? The endogenous growth view, in the increasing returns form set out previously, points either to a slowdown in the growth rate of (comprehensive) capital or to a decline in the degree of the externality α as possible causes of the slowdown. Unfortunately, neither possibility is supported by the available evidence. Investment in R&D as a percent of GDP has been relatively constant, and the proportion of industrial R&D has increased. The growth in fixed capital does not correlate with

17. These conclusions assume that the spillover externality augments the return to labor and "private" capital equally (an implication of the Cobb-Douglas form). All is well if labor continues to be paid the value of its marginal product. However, endogenous growth theory is part of a more general view of growth that stresses the importance of imperfect competition, and it is possible that the presence of spillover externalities may lead to a wedge between output elasticities and factor shares.

the fall in the residual. Moreover, the evidence does not provide support for a decline in the externality or spillover effect (Nadiri 1993; Griliches 1994), although this is debatable. It therefore seems reasonable to conclude that we must look elsewhere in the emerging growth literature, perhaps at the learning and diffusion mechanisms described in the Greenwood-Jovanovic survey, to explain fluctuations in the rate of productivity change.

1.4.3 Data on Quality and the Quality of Data

The production function–based models of TFP described in the preceding sections are based on a *process-oriented* view of technical change, one in which productivity growth occurs through improvements in transforming input into output. No explicit mention has been made of another important dimension of innovation: improvements in the quality of products and the introduction of new goods. Both present consumers and producers with a new array of products and, over time, completely transform the market basket (automobiles replace horses, personal computers replace typewriters, etc.). Much of the welfare gain from innovation comes from the production of better goods, not just from the production of more goods (i.e., by moving up the "quality ladder" [Grossman and Helpman 1991]). Unfortunately, the TFP residual is intended to measure only the production of more goods—this is what a shift in the production function means—and only the costless portion at that. Innovation that results in better goods is not part of the TFP story.

One way to handle this issue is to treat the two types of innovation as separate measurement problems and restrict use of the TFP residual to its proper domain. Unfortunately, the two types of innovation are not easily segregated, as the following example shows. First, imagine two economies, both of which have the same technology and start with 100 units of input, so that both produce 100 physical units of output. Suppose, now, that some ingenious person in economy A discovers a way to double the amount of output that the 100 units of input can produce. At the same time, an innovator in economy B discovers a way to double the utility of the 100 physical units of output that are produced (that is, inhabitants of B gladly exchange two units of the old output for one unit of new output). A measure of TFP based entirely on physical units will double in A but remain flat in B, even though the inhabitants of both countries are equally well off as a result of their respective innovations.

Is this the right result? In a sense, it is. The production function for physical units of output shifted in economy A but not in B. However, this judgment reflects a particular conception of output—that is, that physical units are the appropriate unit of measure. This convention obviously provides an unfavorable view of economy B because it defines away the true gains made in B. An alternative approach would be to measure output in units of consumption efficiency—that is, in units that reflect the marginal

rate of substitution between old and new goods. In this efficiency-unit approach, both A and B experience a doubling of output, albeit for different reasons, and measured TFP reflects the increase. In other words, the TFP model does service in measuring both process and product innovation when output is measured in efficiency units.

The efficiency approach to productivity measurement has proceeded along two general lines. First, the 1950s saw the theoretical development of the model of capital-embodied technical change (Johansen 1959; Salter 1960; Solow 1960). In this model, technical innovation is expressed in the design of new machines, with the implication that different vintages of capital may be in service with different degrees of marginal productivity. When expressed in efficiency units, one physical unit of new capital represents more capital than one physical unit of an older vintage. The total "size" of this capital stock is the number of efficiency units it embodies, and the growth in this stock is the results of two factors: the arrival of more investment and the arrival of better investment. Moreover, the implied rate of productivity growth depends on the rate of investment.

Though theoretically plausible, the capital-embodiment model met initially with limited empirical success. Moreover, it was dismissed as unimportant by one of the leading productivity analysts, Denison (1964). However, the issue did not disappear entirely and has returned to prominence with the hedonic price study by Cole et al. (1986), who used price data to show that official investment-price statistics had essentially missed the computer revolution, overstating price and understating quantity (measured in efficiency units).[18] This finding led the Bureau of Economic Analysis (BEA) to switch to an efficiency-unit convention for investment in computers in the U.S. national income and product accounts (but only for computers). This analysis was extended by Gordon (1990), who adjusted the prices of a wide range of consumer and producer equipment for changes in quality. Gordon also found systematic overstatement of official price statistics and a corresponding understatement of efficiency-adjusted quantity investment output and the resulting capital input.

The CPI is another area in which price data are routinely adjusted for quality change. A variety of procedures is used in the adjustment process, including price hedonics, but the Advisory Commission (1996) concluded that they were not adequate and that the CPI was biased upward by 0.6

18. In the hedonic price model, a product is viewed as a bundle of constituent characteristics. The more there is of each characteristic, the more there is of the good. Computers, for example, are seen in terms of CPU speed, memory speed and capacity, storage capacity, and so on. The hedonic model estimates a "price" for each characteristic and thereby derives an implied price for the whole bundle. This also yields a "quantity" of the good measured in terms of efficiency. Embodied technical change is naturally seen as an increase in the efficiency units via an increase in the characteristics. See Triplett (1983, 1987) for more on the hedonic price model.

percentage points per year. In other words, the growth in efficiency price of consumption goods was overstated, and the corresponding quantity was understated. The BLS is currently undertaking revisions in its procedures, including increased reliance on price hedonics, to address the quality problem.

The fundamental problem with the efficiency approach is that improvements in product quality, or the advent of entirely new consumer goods, are essentially subjective. Physical units can be observed, however imperfectly, but when characteristic-efficiency units are involved, there is no direct observational check to the imputed amount of product. It is all too easy to misstate the true quantity of efficiency units, and there is little intuitive basis for rejecting the misstatement (exactly how much more utility do you feel you get from a Pentium III processor?).[19] It is worth recalling the words of Adam Smith, "Quality . . . is so very disputable a matter, that I look upon all information of this kind as somewhat uncertain."

The subjective nature of the efficiency approach leads to a more subtle problem. Because the quantity of efficiency units is determined by imputation of the relative marginal utility between old and new products, the very definition of product quantity becomes a matter of utility and consumer choice (Hulten 2000). This tends to blur the boundary between the supply-side constraint on growth, the production function, and the objective of growth, which is the province of the utility function. We will return to such boundary issues in the following sections.

1.4.4 Quality Change and the TFP Residual

Most of the TFP studies that have incorporated product-oriented innovation into the residual have focused on capital-embodied technical change. Nelson (1964) expressed the residual as a function of the rate of embodiment and the average age of the capital stock. Domar (1963) and Jorgenson (1966) observed that capital is both an input and an output of the production process, and the failure to measure capital in efficiency units causes two types of measurement error: one associated with the mismeasurement of capital input and one with the mismeasurement of invest-

19. The mismeasurement of quality in improved products is particularly difficult regarding nondurable consumer goods, where reliable overlapping prices of old and new models are harder to obtain. Moreover, the measurement problems posed by "quality" are not limited to product-oriented innovation. There are also myriad problems in the definition of output that involves a quality dimension without reference to innovation. Griliches (1994) speaks of the "hard to measure" sectors of the economy—largely the service sector—and notes that these sectors in particular have grown over time. For example, the bank revenues can be measured with some precision, but what exactly are the units of output? How would one measure these units in principle and account for differences in the quality of service that is characteristic of competition among banks? Unless the nature of the output can be defined precisely, it is impossible to determine its rate of growth and to confront questions about the impact of quality-enhancing innovations like automatic teller machines.

ment good output. Surprisingly, the two errors exactly cancel in Golden Rule steady-state growth, leaving the residual unbiased.[20]

The actual size of the input and output embodiment errors depends on the rate at which embodied efficiency increases and on the average embodied efficiency of the older vintages of capital stock. These cannot be estimated within the residual's index number framework, but in an earlier paper (1992b), I use data from Gordon (1990) to estimate the net embodiment effect for the U.S. manufacturing industry. The net embodiment effect was found to account for about 20 percent of the TFP residual over the time period 1949–83. Wolff (1996) reports an effect that is roughly twice as large for the economy as a whole for the same years. Greenwood, Hercowitz, and Krusell (1997) propose a variant of the embodiment model in which the total value of investment is deflated by the price of consumption rather than investment. The resulting estimate of the embodiment effect accounts for 58 percent of the aggregate residual, per the period 1954–90.

These studies deal with capital-embodied technical change. Productivity analysis has paid less attention to quality change in consumption goods. The example of the economies A and B from the preceding section suggests that this neglect results in an understatement of true output and TFP growth (recall the situation in economy B). However, the problem is even more complicated than that example suggests, because of another problem that has lurked in the background of productivity analysis: the cost of achieving technical innovations. A variant of our example illustrates the problem. Economies A and B each start with 100 units of input and the same technology, and produce 100 physical units of output. Economy A now invests half its workforce in research and is able to quadruple the output of the remaining 50 workers. Output and TFP thus double. In economy B, on the other hand, the 50 are diverted to research and manage to invent a new good that is four times as desirable (that is, inhabitants of B gladly exchange four units of the old output for one unit of new), but only 50 units of physical output are produced. Physical output and TFP fall by half in B, even though innovation has made the inhabitants of B as well off as those in A. The failure to measure output in efficiency units

20. This point is often overlooked in econometric studies of embodied technical change. If both capital input and investment output are correctly measured in efficiency units, the economy-wide TFP residual should be invariant to changes in the rate of capital embodiment. If input and output are not adjusted for quality, aggregate TFP is still invariant along the optimal growth path. Off the optimal path, there is the Hall (1968) identification problem: the exponential part of capital-embodied technical change cannot be distinguished from the equivalent rate of disembodied technical change given price or quantity data on age, vintage, and time. Only deviations from the exponential path can be identified. Finally, it is well to remember that the residual can only measure the costless part of innovation, embodied or otherwise.

thus gives the appearance of technical regress even though progress has occurred.

These considerations can be parameterized and embedded into the standard TFP model by introducing a simple type of quality ladder (Hulten 1996, 2000). Suppose that product-oriented technical change proceeds at a rate θ (essentially the marginal rate of substitution between old goods and new goods of superior quality), and the cost of achieving this rate of quality change is $\mu\theta$. Costless innovation occurs when μ equals zero. In a simplified world in which capital and labor are fixed, it can be shown that the TFP residual falls at the rate $\mu\theta$ when output is measured in physical units, but grows at a rate of $(1 - \mu)\theta$ when efficiency units are used. In the first case, an increase in the rate of innovation θ will actually cause the residual to decrease, resonating with the New Economy critique that the problem with productivity statistics is its failure to count improvements in product quality.[21]

1.4.5 Capacity versus Welfare Interpretations of the Residual: The Problem of Sustainable Consumption

Once it is recognized that product quality adjustments allow consumer welfare parameters to creep into the TFP residual, the boundary between the supply-side conception of the residual and the demand-side interpretations is blurred. If welfare considerations are permitted inside one region of the supply-side boundary (and they must be, if the quality dimension of output is to make sense), perhaps they should be permitted in other boundary areas, such as the net-versus-gross output controversy, where welfare arguments have also been made. After all, a high rate of real GDP growth, and hence a large gross-output productivity residual, can be sustained in the short run by depleting unreproducible resources at the expense of long-run welfare. Net output solves this problem by controlling for depreciation and environmental damage; some believe that it thus provides a more accurate picture of sustainable long-run economic growth. Does it not follow that a separate TFP residual based on net output is the appropriate indicator of the contribution of costless technical innovation to sustainable growth?

The short answer is "no." Changes in social welfare can be shown to depend on the standard gross-output concept of TFP, with no need to define a net-output variant of TFP. The result follows from the optimal growth model studied by Cass (1965) and Koopmans (1965), as augmented by Weitzman (1976), in which the intertemporal utility function

21. There is another possibility. Even if output is correctly measured in quality units, the residual can fall if the rate of innovation θ is pushed beyond its cost-effective optimum. In other words, research "booms" can lower TFP if pushed too far.

$U(C_0, \ldots, C_t)$ is maximized (C_t is the amount of consumption t years from the present time). For present purposes, it is useful to assume that prices are proportional to marginal utilities and to express the intertemporal welfare problem as one of maximizing the present value equation

$$W_0 = \sum_{t=0}^{\infty} \frac{p_t C_t}{(1 + i)^{t+1}}, \tag{9}$$

subject to the production function $C_t + I_t = A_t F(K_t, L_t)$ and the accumulation condition $K_t = I_t + (1 - \delta)K_{t-1}$ (here, we revert to the assumption that Hicksian efficiency and labor growth are exogenously determined). The economic problem of contemporary society, at each point in time, is to determine the optimal division of current output between consumption and investment.

This problem was studied by Weitzman (1976), who demonstrated that the optimal consumption path (C_t^*) satisfies the condition $p_t C_t^* + p_t \Delta K_t^*$. But this is really nothing more than the Hicksian definition of income: the maximum amount of output that could be consumed each year without reducing the original amount of capital, or, equivalently, "sustainable" consumption. This is the welfare indicator appropriate for the annualized measurement of increments to consumer welfare.

This welfare indicator of output is not the same as GDP. According to the fundamental accounting identity in equation (1), GDP is equal to the gross payments to capital and labor (as well as $p_t Q_t$). With some algebraic manipulation based on the Hall-Jorgenson user cost formula, it can be shown that Hicksian income is equal to net factor income or net national product in nominal prices, which differs from gross output by the amount of depreciation (Hulten 1992a):

$$p_t C_t^* + p_t \Delta K_t^* = i_t p_t K_t + w_t L_t < p_t Q. \tag{10}$$

This identity may encourage some to suppose that net national product (NNP) should be used in productivity analysis instead of GDP because it is associated with maximum intertemporal welfare. However, the two output concepts are complements, not substitutes. The growth in real GDP indicates the expansion of the supply-side constraint in any year, and the residual computed using real GDP measures the change in the efficiency of production as represented by A_t (the shift in production constraint). The growth in NNP cum Hicksian income reveals the extent to which growth has improved society's welfare. These are separate issues and must be kept separate, and it is important to recognize that the gross-output TFP residual fits into the welfare-maximization problem via the production constraint.

This result does raise the question of how the gross-output residual is related to changes in economic welfare. This is a complicated issue that

involves treating capital as an intertemporal intermediate product, and linking labor input and technology directly to the attainable consumption path (Hulten 1979). If the optimal consumption path (C_t^*) is chosen—that is, the one that maximizes equation (9)—an intertemporal consumption-based residual can be derived that is the weighted sum of the TFP residuals:

$$\Omega_{0,T} = \sum_{t=0}^{T} \omega_t \frac{\dot{A}_t}{A_t}. \tag{11}$$

The individual weights in this expression, ω_t, are the respective annual ratios of GDP to total wealth, W_0. They are the intertemporal counterparts of the weights used by Domar (1961) and Hulten (1978) to aggregate the sectoral gross-output residuals in the presence of intermediate inputs.

The $\Omega_{0,T}$ residual indicates the increase in optimal consumption associated with changes in the annual (gross-output) TFP residuals. It is not a substitute for these residuals, but a complement. It is clear, once again, that the appropriate welfare-based analysis is separate from, and complementary to, the GDP-based analysis of productive efficiency.

1.4.6 The Boundaries of Productivity Analysis

We have seen that the boundary between welfare and capacity is not as straightforward as one might wish. However, two general boundary principles are clear enough: A distinction must be maintained between ends (welfare improvement) and means (production); and a distinction must also be maintained between the impulse to save (i.e., defer consumption) and the impulse to invent (productivity). This section deals with yet another boundary: the line between what should be counted as output and input and what should not. This "comprehensiveness" boundary is central to the debate about the desirability of a "Green GDP" raised by environmentalists and discussed in Nordhaus and Kokkelenberg (1999).

A complete set of economic accounts would include information on the price and quantity of every variable that enters into the production or utility function of every agent in the economy. The required list of variables would extend far beyond the boundaries of the market economy. Goods produced in the household sector would be an important part of the complete accounting system, including work around the home, leisure, and education. Those public goods produced in the government sector and distributed free of direct charge (or at a price that does not reflect marginal cost) must also be part of the accounts, including national defense, public infrastructure, and so on. Also necessary are goods held in common for private use (such environmental variables as clean air and water, parks, forests, and mineral deposits), as well as spillover externalities, such as knowledge and congestion, and so on.

This is an impossibly large order to fill. The boundaries of a complete accounting system would include everything that correlates with the production of goods and services and affects economic welfare. Thus, for example, the effects of urbanization and materialism that are alleged correlates of the modern capitalist system could force their way into the complete accounts on the grounds that the breakdown of welfare-enhancing institutions (such as family and religion) are the results of these effects. The boundaries of a complete set of economic accounts may thus be extended to include statistics on crime, drug abuse, divorce, and so on.

Boundaries drawn this broadly go far beyond the limits of the current national economic accounts, and probably far beyond the comfort limits of most economists. This reinforces the current national income-accounting practice of relying primarily on market transactions to generate data. Market transactions, though flawed and incomplete, do provide an objective yardstick for measuring the volume of economic activity, as well as prices and quantities. Market data are also relatively easy to collect. These benefits are, unfortunately, purchased at a price: Narrow focus on products exchanged for money leads to the exclusion of many goods the data for which are harder to obtain. This, in turn, can lead to a distorted picture of the true production possibilities facing an economy. Productivity, in any of its many forms, is essentially a ratio of output to input and will be affected by the omission of any element of the numerator or denominator.

This dilemma can be illustrated by the following simplified example. Suppose that an industry produces a good Q_t, which it sells at marginal cost in the marketplace for price P_t. It produces the good using input X_t, which it purchases in the factor market for w_t, but also uses a good Z_t which is available without cost to the firm. The item Z_t might be a common good (e.g., clean air), an externality associated with another agent's behavior (e.g., technical knowledge appropriated from other firms in the industry), or self-constructed capital produced in an earlier year (the firm's stock of technical know-how). In any event, the statistician who looks only at market data will record the accounting identity $P_tQ_t = w_tX_t$, and the analyst will reckon productivity to be Q_t/X_t. The true nature of things is, of course, different. The correct accounting identity is $P_t^*Q_t = w_tX_t + \rho_tZ_t$, where P^* is the marginal social cost of the good, as opposed to the private cost, P_t, and ρ_t is the implicit cost to using the "free" input Z_t. The true productivity ratio is $Q_t/F(X_t, Z_t)$. The example could be complicated further by supposing that the firm generates an externality as it produces Q_t.[22]

22. There are many candidates for the role of "significant omitted variable." One in particular deserves mention because of its relation to the productivity of the computer revolution. The advent of computers has allowed firms to reduce the number of employees, often resulting in productivity gains to the firm. But this has often come at the expense of the consumer, who must substitute his/her own time for that of the departed employee. Anyone who has been on hold for a telephone connection to a human voice, or suffered through seemingly interminable menu-driven options, understands this problem.

In order for the statistician to "get it right," the variable Z_t, must be recognized and measured, and imputations must be made for the shadow prices P^* and ρ_t. The latter is particularly hard. Some imputations can be made using technical procedures like price hedonics, but many must be approached with controversial techniques such as "willingness-to-pay" criteria (e.g., see the discussion in Nordhaus and Kokkelenberg 1999). It is even harder for the statistician to proceed when imputation involves a politically sensitive issue such as the public's health, the preservation of the environment, or worker or product safety. Partisans with different points of view often impute vastly different amounts to the value of life or protection of the environment. In these cases, the imputation process is thus as likely to reflect partisan agendas as to reflect the true nature of productivity growth.

Some imputations are made in practice in the national accounts (e.g., owner-occupied housing), and quasi-imputations for government "output" are used. However, the bulk of unpriced goods is not included. This seems the safe path to follow, at least for the time being. Although the omission of important variables may limit the generality of conclusions that can be drawn from productivity statistics, at least the results are not subject to the changing winds of ideology or special interests. Nor is the direction of the "boundary bias" clear.[23]

1.5 The Total Factor Productivity Residual for Firms and Industries

1.5.1 The View from the Top Down

A TFP residual can, in principle, be computed for every level of economic activity, from the plant floor to the aggregate economy. These residuals are not independent of each other because, for example, the productivity of a firm reflects the productivity of its component plants. Similarly, industry residuals are related to those of the constituent firms, and productivity in the aggregate economy is determined at the industry level. As a result, productivity at the aggregate level will increase if productivity in each constituent industry rises, or if the market share of the high productivity industry increases (and so on, down the aggregation hierarchy).[24] A

23. The debate over boundaries has generally failed to recognize that the omission of environmental spillovers from official data does not necessarily mean that they are unnoticed. The public feel their effects regardless of whether they appear in the data, and, indeed, rational citizens should make their own corrections to flawed data (Hulten 2000). A great deal of pro-environment legislation has been informed by the "biased" statistics, and it is unclear whether fixing the bias would have led to a superior outcome.

24. The significance of shifting sectoral shares for explaining productivity growth has received much attention (see particularly Denison 1967 and Baumol 1967). The shift in resources out of agriculture is often held to be a cause of accelerating productivity growth, and the shift out of manufacturing industry into the service sectors is a potential explanation of slowing productivity. The Baumol stagnation hypothesis holds that a slowdown is inevi-

complete picture of the industrial dynamics of an economy would include a *mutually consistent* measure of the TFP residuals at each level in the hierarchy and of the linkages used to connect levels.

The task of constructing this hierarchy of residuals can be approached from the top down, in a process that can be likened to unpeeling an onion in order to reach lower layers of structure. Domar (1961) was the first to work out the problem of "unpeeling" the TFP residual, and to recognize the complication introduced by the presence of intermediate goods. This complication arises because plants and firms in each sublayer produce goods and services that are used as inputs in the production processes of the plants and firms. As each layer is unpeeled, the magnitude of these intermediate deliveries grows. For example, there are no intermediate goods in the aggregate economy because there is only one industry at this level of aggregation, and all interindustry flows cancel out.

However, these interindustry flows "uncancel" in passing to the one-digit industry level of detail. The iron ore delivered to the steel industry is counted in the gross output of the extractive industries, and is counted again as part of the gross output of the manufacturing industry. The sum of the one-digit industry gross output is therefore larger than total aggregate output.

The nature of this problem can be made more precise by observing that the total output of an industry (plant, firm) is composed of deliveries to final demand plus deliveries of the industry's output to the other industries that use the good. On the input side, the firm uses not only labor and capital, but also intermediate goods purchased from other industries. This leads to the accounting identity

$$p_i D_i + p_i \Sigma_i M_{i,j} = w_i L_i + r_i K_i + \Sigma_j p_{j,i} M_{j,i}. \tag{12}$$

The summation term on the left-hand side of this expression is the value of the deliveries of the *i*th industry's output, and D_i denotes deliveries to final demand (time subscripts have been omitted for clarity of exposition). The summation on the right-hand side is the value of intermediate goods purchased from other industries, and the remaining terms on the right-hand side constitute the value added by the industry, $w_i L_i + r_i K_i$.

There is an expression like equation (12) for each industry (firm, etc.) in the economy. Summing them all up to the aggregate level gives the identity

$$\Sigma_i p_i D_i + \Sigma_i w L_i + \Sigma_i r K_i = wL + rK. \tag{13}$$

(It is assumed here that competition equates wages and capital cost across sectors.) This is a variant of the fundamental accounting identity with

---

table in an economy in which the output demand for the low-productivity growth sector is inelastic. A large literature on this subject has evolved, but space limitations prevent a more detailed treatment of the various strands and criticisms.

which we started, but here we have total deliveries to final demand as the output measured on the left-hand side, and total value added on the right-hand side.

Total factor productivity residuals can be obtained from both expressions—industry residuals from equation (12) and the aggregate residual from equation (13) cum equation (1). Domar (1961) showed that the aggregate residual is the weighted sum of the industry residuals, where the weights are the ratio of industry gross output to total deliveries to final demand (GDP). His results are generalized in Hulten (1978) to

$$\frac{\dot{A}_t}{A_t} = \sum_{i=1}^{N} \frac{p_{i,t} Q_{i,t}}{\sum_i p_{i,t} D_{i,t}} \frac{A_{i,t}}{A_{i,t}}. \tag{14}$$

The unusual feature of this expression is that the weights sum to a quantity greater than 1 to account for the presence of the intermediate goods. Thus, for example, a uniform 1 percent rate of increase in productivity at the industry level may translate into, say, a 1.5 percent increase in productivity at the aggregate level. This inflation in the aggregate number is needed to account for the fact that, although an increase in industry-level productivity augments the production of intermediate goods, these intermediate goods have subsequently disappeared in the process of aggregation.[25]

The production function underlying the residual in equation (14) is the second unusual feature of the analysis. Whereas Solow assumed that the aggregate production function could be expressed as $Q = AF(K, L)$, the technology underlying equation (14) is a production possibility frontier of the following form: $F(Q_1, \ldots, Q_n; K, L, A_1, \ldots, A_n) = 0$. The left-hand side of equation (14) is the shift in the frontier, holding capital and labor constant. The right-hand side indicates that this shift can be "unpeeled" into separate components: the growth rates of industry-level productivity (A_i), and the sectoral share weights, which may change with the reallocation of GDP among sectors with different TFP levels and growth rates. There is no guarantee that the aggregate productivity index is path independent when the component A_i grow at different rates.

The chief difficulty with this unpeeling process lies in the nature of intermediate goods. The quantity gross output and intermediate goods in any industry are greatly affected by mergers and acquisitions. The merger of firms can transform what were once interfirm flows of goods into intrafirm flows, thereby extinguishing some amount of gross output. This has led some researchers to use real value added, a concept of industry output that is immune to this problem.

25. It is no accident that equation (14) looks very much like equation (11), the welfare equivalent of the Solow residual. The welfare residual is based on the intertemporal optimization of consumption, and capital is treated as an intermediate good in that model. Moreover, "years" are formally equivalent to industries in the conventional intermediate goods model described in this section.

The productivity analyst's job would be made easier if intermediate goods could be netted out directly in the identity shown in equation (12), leaving industry final demand equal to value added (i.e., $p_i D_i = w_i L_i + r_i K_i$). However, this will generally not happen, since the value of intermediate goods produced in an industry need not equal the amount used. One solution is to focus on the right-hand side of this expression and define industry output as the "real," or constant-price, part of $w_i L_i + r_i K_i$. Industry value added sums to total value added (GDP), and the relation between the two is not affected by intermediate goods. A variant of the TFP residual can be based on this concept of industry "output" by applying the original Solow formula. The result can be weighted up to the aggregate level using value added weights.

There are, however, two problems with this approach. First, there is nothing in the real world that resembles real value added. Do plants actually make things in units of real value added? Second, it is well known that real value added works only when innovation enhances the productivity of capital and labor but not intermediate inputs—that is, the industry-level production function has the form $Q_i = F[M_i, A_i G(K_i, L_i)]$. Thus, the productivity analyst is confronted with a dilemma: Use the gross output approach and become a prisoner of the degree of vertical and horizontal industrial integration, or use the implausible value added approach. Moreover, there is no guarantee that the production functions underlying either approach are suitable potential functions for the path-independent line integration required in equation (14), and many other problems are encountered at the industry level of analysis (Gullickson and Harper 1999).

1.5.2 The View from the Bottom Up

The preceding remarks take the top-down view of sectoral productivity analysis, in which the aggregate TFP residual is the point of reference. The bottom-up approach to productivity measurement starts from a very different perspective. It takes the universe of plants or firms as the fundamental frame of reference and does not impose the restrictive aggregation assumptions needed to achieve a consistent measure of overall productivity. Instead, it stresses the basic heterogeneity of the microproduction units. An important goal of this approach is to explain the observed heterogeneity of plant productivity in terms of factors such as R&D spending or patenting, or of differences in the financial or industrial structure.[26]

The literature on this approach is huge and can be treated with only a cursory overview. Early contributions were made by Griliches, Mansfield, and others (see Griliches 1994), and the work of Nelson and Winter explic-

26. See Bartelsman and Doms (2000) for a recent survey of this area, and the paper by Foster, Haltiwanger, and Krizan (chapter 8 in this volume).

itly focused on heterogeneity. This line of investigation was greatly aided by the development of microdata sets like the LRD in 1982, and by the enormous increase in computing power, which enabled researchers to analyze increasingly large data sets with ever more sophisticated econometric techniques. The R&D work of Griliches and Mairesse (1984) and B. Hall (1993) is noteworthy in this regard, as are the seminal contributions of Davis and Haltiwanger (1991).

The heterogenous plant/firm approach has much to recommend it because it permits a detailed examination of the factors that actually determine microproductivity. However, its very success is also its chief problem: It is hard to generalize the lessons learned from the microanalysis. This is due in part to the inherent heterogeneity of the data, but it is also due to the diverse (and often contradictory) findings of different econometric studies, although this is not an uncommon problem with large and complex data sets.

Several studies have attempted to link the micro and macro levels of analysis. Baily, Hulten, and Campbell (1992) used data from the LRD to examine the internal dynamics of industry-level residuals. This study found, among other things, that the representative agent model, which is often considered the conceptual link between macro and micro levels of analysis, is not supported by the data. When industry-level residuals were resolved into the weighted sum of the plant-level residuals, it was found that the plants with rising TFP levels and plants with high preexisting TFP levels were the main contributors to productivity growth. Firms with low preexisting TFP levels and declining firms were a drag on productivity. The persistence of firms with both high and low levels of productivity suggests a more complex view of industrial organization than the simple representative agent model used to motivate the aggregate TFP residual. The microdata also suggest a more complex productivity dynamic in which the entry and exit of firms, as well as their expansion and contraction, are important dimensions.

Many advances have been made in subsequent research. However, it remains true that a compelling link between the micro and macro levels has yet to be forged. This is one of the greatest challenges facing productivity analysts today. This challenge is all the more daunting because it must confront this problem: Industries are composed of heterogenous firms operated under conditions of imperfect competition, but the theoretical aggregation conditions required to proceed upward to the level of macroeconomy rely on perfect competition.

1.6 Conclusion

Any respectable biography must end with a summary judgment of the subject at hand; and, above all, the true character of the subject should be

revealed. This is particularly important in the case of the TFP residual, the true character of which has often been misunderstood by friends and critics alike. The portrait painted in this paper reveals these essential features:

1. The TFP residual captures changes in the amount of output that can be produced by a given quantity of inputs. Intuitively, it measures the shift in the production function.
2. Many factors may cause this shift: technical innovations, organizational and institutional changes, shifts in societal attitudes, fluctuations in demand, changes in factor shares, omitted variables, and measurement errors. The residual should *not* be equated with technical change, although it often is.
3. To the extent that productivity is affected by innovation, it is the costless part of technical change that it captures. This "manna from heaven" may reflect spillover externalities thrown off by research projects, or it may simply reflect inspiration and ingenuity.
4. The residual is a nonparametric index number designed to estimate one parameter in the larger structure of production, the efficiency shift parameter. It accomplishes this by using prices to estimate marginal products.
5. The various factors comprising TFP are not measured directly but are lumped together as a "left-over" factor (hence the name "residual"). They cannot be sorted out within the pure TFP framework, and this is the source of the famous epithet, "a measure of our ignorance."
6. The Divisia index must be path independent to be unique. The discrete-time counterpart of the Divisia index, the Tornqvist approximation, is an exact index number if the underlying production function has the translog form. The problem of path dependence is one of uniqueness, and this is not the same thing as measurement bias.
7. The conditions for path independence are (a) the existence of an underlying production function and (b) marginal productivity pricing. Neither constant returns to scale nor Hicksian neutrality are absolutely necessary conditions, although they are usually assumed for convenience of measurement.
8. When the various assumptions are met, the residual is a valid measure of the shift in the production function. However, it generally understates the importance of productivity change in stimulating the growth of output because the shift in the function generally induces further movements along the function as capital increases.
9. The residual is a measure of the shift in the supply-side constraint on welfare improvement, but it is not intended as a direct measure of this improvement. To confuse the two is to confuse the constraint with the objective function.

This is the essential character of our subject. As with any portrait that is examined closely, flaws are detected and the final judgment is usually mixed with praise and criticism.

Much of the praise is deserved, but so is much of the criticism. The assumptions needed for the TFP model to work perfectly are stringent; much is left out of the analysis, and the pure TFP approach did not provide a consensus explanation of the productivity slowdown. However, alternative approaches are not immune to these criticisms, and a fair judgment must go beyond these criticisms and address a more fundamental question: To what extent are the perceived failures inherent in the character of the residual, and to what extent are the problems inherent in the data to which the residual technique is applied? If data on prices and quantities do not accurately reflect quality improvement, or if the boundaries of the data set are drawn too closely, attacking TFP is rather like shooting the messenger because of the message. If the data are the real source of complaint, other methods (e.g., econometrics) will not fare much better than the simple residual. Bad data are bad data regardless of how they are used.

The positive value of the TFP residual greatly outweighs the negatives. The residual has provided a simple and internally consistent intellectual framework for organizing data on economic growth, and has provided the theory to guide economic measurement. Moreover, it teaches lessons that are still not fully appreciated by mainstream economics and national income accounting: An empirically testable theory places restrictions on the way data must be collected and organized, and choices about the measurement procedures are often implicit choices about the underlying theory.

The residual is still, after more than forty years, the workhorse of empirical growth analysis. For all the residual's flaws, real and imagined, many researchers have used it to gain valuable insights into the process of economic growth. Thousands of pages of research have been published, and more are added every year (for, example, the TFP residual is central to the recent debate over the role of computers in stimulating economic growth). Total factor productivity has become a closely watched government statistic. Not bad for a forty-year-old.

References

Abramovitz, Moses. 1956. Resource and output trends in the United States since 1870. *American Economic Review* 46 (2): 5–23.

Aschauer, David A. 1989. Is public expenditure productive? *Journal of Monetary Economics* 23:177–200.

Baily, Martin N., and Robert J. Gordon. 1988. Measurement issues, the slowdown, and the explosion of computer power. *Brookings Papers on Economic Activity,* issue no. 2: 347–420. Washington, D.C.: Brookings Institution.

Baily, Martin N., Charles R. Hulten, and David Campbell. 1992. Productivity dynamics in manufacturing plants. *Brookings Papers on Economic Activity, Microeconomics:* 187–249. Washington, D.C.: Brookings Institution.

Barro, Robert J., and Xavier Sala-i-Martin. 1995. *Economic growth.* New York: McGraw-Hill.

Bartelsman, Eric J., and Mark Doms. 2000. Understanding productivity: Lessons from longitudinal micro data. *Journal of Economic Literature* 37 (3): 569–94.

Baumol, William J. 1967. Macroeconomics of unbalanced growth: The anatomy of urban crisis. *Journal of Political Economy* 57:415–26.

Baumol, William J., Sue A. B. Blackman, and Edward N. Wolff. 1989. *Productivity and American leadership: The long view.* Cambridge, Mass.: MIT Press.

Berndt, Ernest R. 1980. Energy price increases and the productivity slowdown in U.S. manufacturing. *The decline in productivity growth,* Conference Series no. 22, 60–89. Boston: Federal Reserve Bank of Boston.

Berndt, Ernest R., and Melvyn A. Fuss. 1986. Productivity measurement with adjustments for variations in capacity utilization, and other forms of temporary equilibrium. *Journal of Econometrics* 33:7–29.

Cas, Alexandra, and Thomas K. Rymes. 1991. *On concepts of multifactor productivity.* New York: Cambridge University Press.

Cass, David. 1965. Optimum growth in an aggregative model of capital accumulation. *Review of Economic Studies* 32 (July): 233–40.

Caves, Douglas W., Laurits R. Christensen, and W. Erwin Diewert. 1982a. Multilateral comparisons of output, input, and productivity using superlative index numbers. *Economic Journal* 92 (March): 73–86.

———. 1982b. The economic theory of index numbers and the measurement of input, output, and productivity. *Econometrica* 50 (6): 1393–1414.

Christensen, Laurits R., Diane Cummings, and Dale W. Jorgenson. 1981. Relative productivity levels, 1947–1973: An international comparison. *European Economic Review* 16:61–94.

Christensen, Laurits R., and Dale W. Jorgenson. 1969. The measurement of U.S. real capital input, 1929–1967. *Review of Income and Wealth* 15 (December): 293–320.

———. 1970. U.S. real product and real factor input, 1929–1969. *Review of Income and Wealth* 16 (March): 19–50.

Christensen, Laurits R., Dale W. Jorgenson, and Lawrence J. Lau. 1973. Transcendental logarithmic production frontiers. *Review of Economics and Statistics* 55 (February): 28–45.

Cobb, Clifford, Ted Halstead, and Jonathan Rowe. 1995. If the GDP is up, why is America down? *Atlantic* 276 (4): 59–78.

Cole, Rosanne, Y. C. Chen, J. A. Barquin-Stolleman, E. Dullberger, N. Helvacian, and J. H. Hodge. 1986. Quality-adjusted price indexes for computer processors and selected peripheral equipment. *Survey of Current Business* 66 (January): 41–50.

Collins, Susan M., and Barry P. Bosworth. 1996. Economic growth in East Asia: Accumulation versus assimilation. *Brookings Papers on Economic Activity,* issue no. 2: 135–91.

Copeland, Morris A. 1937. Concepts of national income. *Studies in Income and Wealth.* Vol. 1, 3–63. New York: National Bureau of Economic Research.

Copeland, Morris A., and E. M. Martin. 1938. The correction of wealth and income estimates for price changes. *Studies in Income and Wealth.* Vol. 2, 85–135. New York: National Bureau of Economic Research.

Davis, Steven J., and John C. Haltiwanger. 1991. Wage dispersion between and

within U.S. manufacturing plants, 1963–86. *Brookings Papers on Economic Activity,* issue no. 1: 115–20.

Denison, Edward F. 1962. *The sources of economic growth in the United States and the alternatives before us.* New York: Committee for Economic Development.

———. 1964. The unimportance of the embodiment question. *American Economic Review* 79 (5): 90–94.

———. 1967. *Why growth rates differ: Postwar experiences in nine western countries.* Washington, D.C.: Brookings Institution.

———. 1972. Some major issues in productivity analysis: An examination of the estimates by Jorgenson and Griliches. *Survey of Current Business* 49 (5, part 2): 1–27.

———. 1979. Explanations of declining productivity growth. *Survey of Current Business* 59 (August): 1–24.

Denny, Michael, Melvyn Fuss, and Leonard Waverman. 1981. The measurement and interpretation of total factor productivity in regulated industries, with an application to Canadian telecommunications. In *Productivity measurement in regulated industries,* ed. T. Cowing and R. Stevenson, 179–218. New York: Academic Press.

Diewert, W. Erwin. 1976. Exact and superlative index numbers. *Journal of Econometrics,* 4:115–45.

———. 1980. Aggregation problems in the measurement of capital. In *The measurement of capital,* ed. Dan Usher, 433–528. Studies in Income and Wealth, vol. 45. Chicago: University of Chicago Press.

Diewert, W. Erwin, and Kevin J. Fox. 1999. Can measurement error explain the productivity paradox? *Canadian Journal of Economics* 32 (April): 251–80.

Domar, Evsey D. 1961. On the measurement of technical change. *Economic Journal* 71:710–29.

———. 1963. Total factor productivity and the quality of capital. *Journal of Political Economy* 71 (December): 586–88.

Dowrick, Steven, and Duc-Tho Nguyen. 1989. OECD comparative economic growth 1950–85: Catch-up and convergence. *American Economic Review* 79 (5): 1010–30.

Easterly, William. 1995. The mystery of growth: Shocks, policies, and surprises in old and new theories of economic growth. *Singapore Economic Review* 40 (1): 3–23.

Färe, Rolf, Shawna Grosskopf, Mary Norris, and Zhongyang Zhang. 1994. Productivity growth, technical progress, and efficiency change in industrialized countries. *American Economic Review* 84 (1): 66–83.

Fisher, Franklin. 1965. Embodied technical change and the existence of an aggregate capital stock. *Review of Economic Studies* 32:326–88.

Gallman, Robert E. 1987. Investment flows and capital stocks: U.S. experience in the 19th century. In *Quantity and quiddity: Essays in U.S. economic history in honor of Stanley Lebergott,* ed. Peter Kilby, 214–54. Middletown, Conn.: Wesleyan University Press.

Gordon, Robert J. 1990. *The measurement of durable goods prices.* Chicago: University of Chicago Press.

Greenwood, Jeremy, Zvi Hercowitz, and Per Krusell. 1997. Long-run implications of investment-specific technical change. *American Economic Review* 87 (3): 342–62.

Griliches, Zvi. 1980. R&D and the productivity slowdown. *American Economic Review* 70 (2): 343–48.

———. 1988. Research expenditures and growth accounting. In *Technology, education, and productivity,* ed. Zvi Griliches, 249–67. New York: Blackwell.
———. 1992. The search for R&D spillovers. *Scandinavian Journal of Economics* 94:29–47.
———. 1994. Productivity, R&D, and the data constraint. *American Economic Review* 84 (1): 1–23.
———. 1996. The discovery of the residual: A historical note. *Journal of Economic Literature* 34 (September): 1324–30.
Griliches, Zvi, and Jacques Mairesse. 1984. Productivity and R&D growth at the firm level. In *R&D, patents, and productivity,* ed. Zvi Griliches, 339–73. Chicago: University of Chicago Press.
Grossman, Gene M., and Elhanan Helpman. 1991. *Innovation and growth in the global economy.* Cambridge, Mass.: MIT Press.
Gullickson, William, and Michael J. Harper. 1999. Possible measurement bias in aggregate productivity growth. *Monthly Labor Review* 122 (2): 47–67.
Hall, Bronwyn H. 1993. Industrial research during the 1980s: Did the rate of return fall? *Brookings Papers on Economic Activity, Microeconomics:* 289–330.
Hall, Robert E. 1968. Technical change and capital from the point of view of the dual. *Review of Economic Studies* 35:34–46.
———. 1988. The relation between price and marginal cost in U.S. industry. *Journal of Political Economy* 96:921–47.
Hall, Robert E., and Dale W. Jorgenson. 1967. Tax policy and investment behavior. *American Economic Review* 57:391–414.
Hicks, John. 1946. *Value and capital.* London: Oxford University Press.
Hulten, Charles R. 1973. Divisia index numbers. *Econometrica* 41:1017–25.
———. 1975. Technical change and the reproducibility of capital. *American Economic Review* 65 (5): 956–65.
———. 1978. Growth accounting with intermediate inputs. *Review of Economic Studies* 45 (October): 511–18.
———. 1979. On the "importance" of productivity change. *American Economic Review* 69:126–36.
———. 1986. Productivity change, capacity utilization and the source of efficiency growth. *Journal of Econometrics* 33:31–50.
———. 1990. The measurement of capital. In *Fifty years of economic measurement,* ed. Ernst R. Berndt and Jack E. Triplett, 119–52. Studies in Income and Wealth, vol. 54. Chicago: University of Chicago Press.
———. 1992a. Accounting for the wealth of nations: The net versus gross output controversy and its ramifications. *Scandinavian Journal of Economics* 94 (supplement): S9–S24.
———. 1992b. Growth accounting when technical change is embodied in capital. *American Economic Review* 82 (4): 964–80.
———. 1996. Quality change in capital goods and its impact on economic growth. NBER Working Paper no. 5569. Cambridge, Mass.: National Bureau of Economic Research, May.
———. 1997. Comment on "Do real output and real wage measures capture reality? The history of lighting suggests not." In *The Economics of New Goods.* Vol. 58, Studies in Income and Wealth, ed. Timothy Bresnahan and Robert J. Gordon, 66–70. Chicago: University of Chicago Press.
———. 2000. Measuring innovation in the New Economy. University of Maryland, Manuscript.
Hulten, Charles R., and Sylaja Srinivasan. 1999. Indian manufacturing industry: Elephant or tiger? NBER Working Paper no. 5569. Cambridge, Mass.: National Bureau of Economic Research, October.

Hulten, Charles R., and Frank C. Wykoff. 1981. The estimation of economic depreciation using vintage asset prices. *Journal of Econometrics* 15:367–96.
Johansen, Leif. 1959. Substitution versus fixed production coefficients in the theory of economic growth: A synthesis. *Econometrica* 27 (April): 157–76.
Jones, Charles I. 1995a. R&D-based models of economic growth. *Journal of Political Economy* 103 (August): 759–84.
———. 1995b. Times series tests of endogenous growth models. *Quarterly Journal of Economics* 110 (2): 495–525.
Jorgenson, Dale W. 1963. Capital theory and investment behavior. *American Economic Review* 53 (2): 247–59.
———. 1966. The embodiment hypothesis. *Journal of Political Economy* 74 (February): 1–17.
Jorgenson, Dale W., Frank M. Gollop, and Barbara M. 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 (July): 349–83.
———. 1972. Issues in growth accounting: A reply to Edward F. Denison. *Survey of Current Business* 52:65–94.
Jorgenson, Dale W., and Mieko Nishimizu. 1978. U.S. and Japanese economic growth, 1952–1974: An international comparison. *Economic Journal* 88 (December): 707–26.
Jorgenson, Dale W., and Kevin J. Stiroh. 2000. Raising the speed limit: U.S. economic growth in the information age. *Brookings Papers on Economic Activity,* issue no. 2: pp. 125–211.
Kendrick, John. 1961. *Productivity trends in the United States.* New York: National Bureau of Economic Research.
Kim, Jong-Il, and Lawrence J. Lau. 1994. The sources of economic growth of the East Asian newly industrialized countries. *Journal of Japanese and International Economies* 8:235–71.
Koopmans, T. C. 1965. On the concept of optimal economic growth. *Pacifica Academia Scientiarus* (Rome):276–79.
Krugman, Paul. 1994. The myth of Asia's miracle. *Foreign Affairs* 73 (6): 62–77.
Lucas, Robert E., Jr. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22:3–42.
Maddison, Angus. 1987. Growth and slowdown in advanced capitalist economies: Techniques and quantitative assessment. *Journal of Economic Literature* 25 (2): 649–98.
Nadiri, M. Ishaq. 1970. Some approaches to the theory and measurement of total factor productivity: A survey. *Journal of Economic Literature* 8 (December): 1137–77.
———. 1993. *Innovations and technological spillovers.* NBER Working Paper no. 4423. Cambridge, Mass.: National Bureau of Economic Research, August.
Nadiri, M. Ishaq, and Seongjun Kim. 1996. R&D, production structure, and productivity growth: A comparison of U.S., Japan, and Korean manufacturing sectors. RR no. 96-11, C. V. Starr Center for Applied Economics, New York University, March.
Nelson, Richard R. 1964. Aggregate production functions and medium-range growth projections. *American Economic Review* 54:575–606.
Nordhaus, William D. 1997. Do real output and real wage measures capture reality? The history of lighting suggests not. In *The economics of new goods,* ed. Timothy Bresnahan and Robert J. Gordon, 29–66. Studies in Income and Wealth, vol. 58. Chicago: University of Chicago Press.
Nordhaus, William D., and Edward C. Kokkelenberg, eds. 1999. *Nature's numbers:*

Expanding the national economic accounts to include the environment. Washington, D.C.: National Research Council, National Academy Press.

Oliner, Stephen D., and Daniel E. Sichel. 2000. The resurgence of growth in the late 1990s: Is information technology the story? *Journal of Economic Perspectives,* 14 (4), 3–22.

Prucha, Ingmar R., and M. Ishaq Nadiri. 1981. Endogenous capital utilization and productivity measurement in dynamic factor demand models: Theory and an application to the U.S. electrical machinery industry. *Journal of Econometrics* 15:367–96.

Richter, Marcel K. 1966. Invariance axioms and economic indexes. *Econometrica* 34:739–55.

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

Rymes, Thomas K. 1971. *On concepts of capital and technical change.* Cambridge, Mass.: Cambridge University Press.

Salter, W. E. G. 1960. *Productivity and technical change.* Cambridge: Cambridge University Press.

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

———. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39 (August): 312–20.

———. 1960. Investment and technical progress. In *Mathematical methods in the social sciences 1959,* ed. K. Arrow, S. Karlin, and P. Suppes, 89–104. Stanford: Stanford University Press.

Stigler, George J. 1947. *Trends in output and employment.* New York: National Bureau of Economic Research.

Tinbergen, Jan. 1942. Zur theorie der langfristigen wirtschaftsentwick lung. *Weltwirtschaftliches Archiv* 55 (1): 511–49.

Törnqvist, L. 1936. The Bank of Finland's consumption price index. *Bank of Finland Monthly Bulletin* 10:1–8.

Triplett, Jack E. 1983. Concepts of quality in input and output price measures: A resolution of the user value-resource cost debate. In *The U.S. national income and product accounts: Selected topics,* ed. Murray F. Foss, 296–311. Studies in Income and Wealth, vol. 47. Chicago: University of Chicago Press.

———. 1987. Hedonic functions and hedonic indexes. In *The new Palgrave dictionary of economics,* vol. 2, ed. John Eatwell, Murray Milgate, and Peter Newman, 630–34. New York: Macmillan Press Limited.

———. 1996. Depreciation in production analysis and economic accounts. *Economic Inquiry* 31 (1): 93–115.

U.S. Advisory Commission to Study the Consumer Price Index. 1996. *Toward a more accurate measure of the cost of living.* Final Report to the Senate Finance Committee. 4 December.

U.S. Department of Commerce, Bureau of the Census. 1975. *Historical statistics of the United States, colonial times to 1970.* Washington, D.C.: GPO.

U.S. Department of Labor, Bureau of Labor Statistics. 1983. *Trends in multifactor productivity, 1948–81,* Bulletin 2178. Washington, D.C.: GPO, September.

Weitzman, Martin L. 1976. On the welfare significance of national product in a dynamic economy. *Quarterly Journal of Economics* 90:156–62.

Wolff, Edward N. 1996. The productivity slowdown: The culprit at last? Follow-up on Hulten and Wolff. *American Economic Review* 86 (5): 1239–52.

Young, Alwyn. 1992. A tale of two cities: Factor accumulation and technical change in Hong Kong and Singapore. In *NBER macroeconomics annual*

1992, ed. Olivier Blanchard and Stanley Fischer, 13–53. Cambridge, Mass.: MIT Press.
———. 1995. The tyranny of numbers: Confronting the statistical realities of the East Asian experience. *The Quarterly Journal of Economics* 110 (3): 641–80.

Comment

Jack E. Triplett

Charles R. Hulten has given us what he calls "a biography of an idea." It is, as we expect from Hulten, a valuable biography, a contribution that one would assign to graduate students as an introduction to the productivity literature. My comments amplify some of Hulten's points.

Measuring Productivity and Explaining It

Hulten states as an organizing principle that the productivity paradigm seeks to determine how much economic growth originates from productivity improvement (improvements in technology) and how much from increasing inputs (he says capital inputs, but per capita growth can also increase because of improvements in labor quality). Hulten uses this input-productivity or input-technology dichotomy as an organizing principle not only because it is theoretically appropriate (the theory of capital, for example, provides a framework for thinking about the capital input), but also because the available data and the relevant empirical work are both organized around the same dichotomy.

Useful as this dichotomy is, sometimes it cannot be implemented. One example, discussed in section 1.4, occurs when new technology is embodied in new machinery. A second, related problem is the distinction between innovations that are costly (brought about by investment in R&D, for example), and those that are in some sense or other "costless." I would add another: At the margins, the dichotomy depends on whether a particular innovating activity is paid for, and not just whether the innovation is costly. An anecdote illustrates.

A number of years ago I toured a machine-tool manufacturing plant. This establishment made very high-tech, advanced machine tools, but the factory in which these machines were made had been built in the nineteenth century and was originally water powered. Its manager told me that the employees had always brought the purchased materials in on the ground floor, carried out subassemblies on the second, and completed final assembly on the third floor. As the machines became larger and more complex, it proved ever more difficult to get them down from the third floor. Someone suggested sending the materials to the third floor so final as-

Jack E. Triplett is a visiting fellow at the Brookings Institution.

sembly could take place on the ground, an idea that resulted in an immediate improvement in the plant's productivity.

How does the input-productivity dichotomy deal with such new ideas? Suppose the suggestion had come from a (paid) management consulting firm, and then suppose the contract had called for the consulting firm to be paid the discounted value of the expected stream of marginal cost savings from its suggestion. Then, the change in the plant's productive arrangements would be fully attributed to an input, and we would record no multifactor productivity (MFP).

Suppose, on the other hand, that the suggestion came from an employee and that the company did not pay the employee the full marginal product of the suggestion. Then there is no compensated input. The dichotomy attributes the improvement entirely to MFP.[1]

Few suggestions, I suspect, will be paid for fully, because of uncertainty about their ultimate value if for no other reason. Many real productive improvements bridge, uncomfortably, the input-productivity dichotomy, especially when we try to implement the dichotomy empirically with fairly aggregative data. The example suggests that the conceptual framework that divides economic growth into input growth and MFP—often called "the residual"—carries us only so far, useful as the framework is. Hulten makes related points; I am emphasizing this problem only because others have overlooked it.

The Productivity Slowdown and Mismeasurement of Economic Variables

Hulten notes, now only in passing, that the "mismeasurement hypothesis" is a very popular one among economists for explaining the post-1973 productivity slowdown. Though the hypothesis may ultimately be confirmed, there is enormous confusion within the profession about the hypothesis. When the mismeasurement hypothesis is properly understood, there is very little evidence in its behalf.

As I have noted elsewhere (Triplett 1997, 1998) the mismeasurement hypothesis is a hypothesis about *differential* mismeasurement. It is a statement that mismeasurement is more severe after 1973 than before.[2]

The evidence most often cited in behalf of the mismeasurement hypothesis consists of findings that some variable or other is currently mismeasured. For example, the Boskin Commission estimated that the Consumer Price Index (CPI) was upwardly biased by about 1.1 percentage points per year in the 1990s.

1. If the employee got the idea from some other firm, which, in turn, had paid the management consultant, this case would parallel spillovers from R&D.

2. It is also a statement that the effects of mismeasurement go predominantly in the same direction—that price increases are overstated by mismeasurement, and that growth rates of real variables, especially those of output, are understated.

However, the Boskin Commission provided no evidence that the CPI has been *differentially* mismeasured. For Boskin-type bias to explain the productivity slowdown, the CPI must have been measured more accurately in "the old days," before 1973. Yet, in 1961 the Stigler Committee pointed to almost exactly the same list of CPI defects that were identified by the Boskin Commission, and it recorded a professional consensus that the CPI was upwardly biased *then* (though the Stigler Committee never made a point estimate of the CPI bias). It should be evident that CPI measurement error did not begin in 1973; neither did the defects in the CPI cited by the Boskin Commission commence in the post-1973 period. The Boskin Commission estimate, therefore, does not by itself provide any evidence in favor of the mismeasurement hypothesis.

Indeed, convincing evidence is lacking that a major part of the productivity slowdown has its origins in differential mismeasurement. Differential mismeasurement implies one or more of several things: that statistical agencies are now worse than they used to be at adjusting for quality change and measuring the hard-to-measure services; that the amount of quality change is greater than it used to be; that measuring services is for some reason more difficult than it used to be (perhaps because the nature of services has changed); or that the sectors where mismeasurement exists have become more important than they were before 1973.

Additionally, the mismeasurement hypothesis implies that the measurement changes must have been abrupt because the productivity slowdown was abrupt. Though there is some debate about whether it really started in 1973, or whether signs of it were visible in the United States around 1968, the slowdown was not a gradual reduction in the rate of productivity improvement. If mismeasurement is to account for the productivity slowdown, then we must find some fairly abrupt change in measurement practices, or an abrupt increase in measurement problems, or an abrupt increase in the size of the poorly measured sectors of the economy. There is little evidence on this, but introspection weighs against abruptness in these changes.

Finally, the mismeasurement hypothesis implies measurement changes in many countries, because the productivity slowdown affected most industrialized economies at about the same time and in roughly similar magnitudes. Even if one thought that the U.S. Bureaus of Labor Statistics (source of U.S. price indexes and productivity measures) and Economic Analysis (the compilers of GDP) did things better in the "old days" (which seems implied by the views of some U.S. economists who subscribe to the mismeasurement hypothesis), how could economic statisticians in all countries "forget" in concert?[3]

3. I have discussed the evidence on the mismeasurement hypothesis in Triplett (1997, 1998, 1999).

Table 1C.1 Top Computer-Using Industries, 1992 Capital Flow

	Computers ($ millions)	Computers and Peripherals ($ millions)
Financial services	2,270	6,677
Wholesale trade	1,860	4,874
Business services[a]	1,383	3,598
Miscellaneous equipment rental and leasing	1,233	3,200
Communications services	873	2,299
Insurance services	738	1,875
Top four businesses	6,746	18,349
Percentage of top four industries of total	42.6	42.1
Top six industries	8,357	22,523
Percentage of top six industries of total	52.8	51.7

Source: Bonds and Aylor (1998).

[a]Excludes miscellaneous equipment rental and leasing.

I believe that the productivity slowdown is real, that it is not primarily a chimera caused by mismeasurement.

A Different Mismeasurement Story

Mismeasurement in economic statistics exists, however, and it is a problem for understanding productivity and technical change in exactly the portions of our economy—high technology and services—that are its rapidly expanding and dynamic sectors.

Computers are nearly the essence of the technology of our time. Consider where computers go, and where they are most used. Four industrial sectors—financial services, wholesale trade, miscellaneous equipment renting and leasing, and business services—account for more than 40 percent of computer investment in the 1992 capital flow table (Bonds and Aylor 1998). Add in two more sectors—insurance and communications—and the share exceeds 50 percent (see table 1C.1). Only in miscellaneous renting and leasing does the share of computer investment in total equipment investment approach half; these computer-using sectors are not necessarily computer intensive.

These six computer-using industries share several important characteristics. First, they are all services industries, broadly defined.

Second, *measured* productivity in these computer-using industries has been declining. Table 1C.2 presents the available numbers.[4]

4. Data in table 1C.2 do not incorporate the revisions to the industry accounts released in mid-2000.

Table 1C.2 **Multifactor Productivity and Labor Productivity, Selected Service Industries**

	Multifactor Productivity		Labor Productivity (GPO per hour)	
	1947–63	1977–93	1960–73	1973–97
Financial services				
Banks (SIC 60, 61)	n.a.	−2.9[a]	0.2	−0.3
Insurance services				
Insurance carriers	n.a.	−2.2	1.9	−0.1
Insurance agents	n.a.	−2.7	0.2	−0.8
Wholesale trade	n.a.	1.3	3.2	2.9
Business services (SIC 73)	n.a.	−0.4[b]	−0.2[c]	−0.4[c]
Communications services	2.5	1.8	5.0	3.9

Sources: Multifactor productivity figures are from Gullickson and Harper (1999). Labor productivity figures are from Triplett and Bosworth (2001).

Note: n.a. = not available.

[a]Also includes holding companies.

[b]Includes miscellaneous repair services (SIC 76).

[c]Also includes professional services (SIC 87).

New MFP estimates for services industries are in a BLS study by Gullickson and Harper (1999). Multifactor productivity is the ratio of gross output to capital and labor inputs. Additionally, value added per hour can be computed from BEA's gross product originating (GPO) series. Statistical information for services industries is often less complete than for the goods-producing sectors, as the "n.a." entries in table 1C.2 indicate.

Even though gross output MFP and value-added labor productivity do not always agree—and indeed, they shouldn't—the general picture for these computer-using services industries is the same, no matter which measure is used: Productivity growth has slowed remarkably since 1973, compared with the earlier postwar years. Additionally, table 1C.2 is filled with negative productivity numbers. In fact, among the computer-intensive services industries, only communications and wholesale trade show upward trends. Negative productivity numbers are always puzzling.

Third, with the possible exception of communications, the outputs of all these computer-intensive services industries are hard to measure.[5] As Zvi Griliches (1994, 1997) has repeatedly emphasized, if we do not know how to measure the output of an industry, then we do not know how to measure its productivity. And if the available productivity numbers, measured as best the statistical agencies can, show negative productivity, per-

5. How does one measure the output of banking and finance? This is an old, contentious issue in national accounts (see Triplett 1992 for a summary). A similar controversy concerns the output of the insurance industry. Furthermore, how do we measure the output of business services? For example, what is the output of an economics consulting firm? What is its price index? How would we compute its productivity?

haps the reason is that economic statistics are missing part of the output that these industries produce.

The relevance of this mismeasurement point is underscored by communications, which has positive productivity growth. Communications output is probably measured better than is the output of the computer-using services industries that have negative productivity. For example, even though evidence suggests that new communications products, such as cellular phones (Hausman 1997), do not get into the data fast enough, economic statistics are probably better at measuring telephone calls than consulting services. It may be no coincidence that communications is the computer-intensive industry with the strongest positive productivity growth. Those other negative productivity numbers might be suspicious.

Even if the output of computer-intensive services industries is mismeasured, this is not evidence for mismeasurement of *aggregate* productivity. Most of the output of these computer-using industries is intermediate, not final. By definition, all of business services (except for exports) and all of wholesale trade are intermediate products. Equipment renting and leasing is also largely an intermediate activity (consumer renting is in the retail sector in the old U.S. SIC system, and computer, aircraft, and vehicle leasing are not classified in this industry). Although finance, insurance, and communications contribute to final output in their sales to consumers (and in contributions to net exports),[6] much of their output goes to other business. Roughly two-thirds of communications and half of insurance are intermediate inputs to other industries. Thus, half of computer investment in the United States goes to six industries that primarily produce intermediate output.

The outputs of intermediate products net out in aggregate productivity measures, such as BLS's private nonfarm MFP. If computers are revolutionizing wholesale trade, as anecdotes suggest, their impact on wholesale trade will show up in the aggregate productivity numbers in the downstream industries that consume the output of the wholesale trade sector, mainly retail trade. If U.S. economic statistics measure correctly the price indexes and output of the retail trade sector (and that is a big "if"), then the contribution of computer investment in wholesale trade will already be incorporated into the aggregate productivity numbers, no matter how wholesale trade output is measured. Similarly, the causes of the great expansion of business services in the U.S. economy are not clear; but if business services are doing something to raise aggregate productivity, then their contribution is to the downstream-using industries.[7] Even if productivity growth in these computer-using industries were tremendous, it could not affect aggregate productivity directly, because in aggregate productivity,

6. Insurance has negative net exports.
7. Except for exports of business services, which have been growing rapidly.

as in GDP, the contributions of intermediate-producing industries cancel out in the totals.

Having no effect on aggregate productivity numbers does not mean, however, that possible mismeasurement in computer-intensive, intermediate services industries is unimportant. To understand the role of technology in a high-tech economy, to understand the impact of the computer on the U.S. economy, we ought to be looking at the impact of the computer at the industry level, to ask how computers have been contributing to industry growth and productivity, and how those industry growth patterns affect other industries and their uses of resources. At the industry level, however, our economic statistics do not appear adequate to analyze the effect of the computer, because much computer investment goes to sectors of the economy where even the concept of output is not well defined, and the existing measures of output in these computer-using sectors seem questionable. If the output measures and the productivity measures are inadequate, we lack the statistical basis on which to determine the impact of technology on industry performance. For a technological country, that is a great informational lacuna.

I conclude by stating that this is a good paper that deserves wide readership.

References

Bonds, Belinda, and Tim Aylor. 1998. Investment in new structures and equipment in 1992 by using industries. *Survey of Current Business* 78 (12): 26–51.

Griliches, Zvi, ed. 1992. Output measurement in the service sector. Studies in Income and Wealth, vol. 56. 71–108. Chicago: University of Chicago Press.

———. 1994. Productivity, R&D, and the data constraint. *American Economic Review* 84 (1): 1–23.

———. 1997. Paper read at the Simon Kuznets Memorial Lectures, 30 October, at Yale University, New Haven, Connecticut.

Gullickson, William, and Michael J. Harper. 1999. Possible measurement bias in aggregate productivity growth. *Monthly Labor Review* 122 (2): 47–67.

Hausman, Jerry. 1997. Valuing the effect of regulation on new services in telecommunications. *Brookings Papers on Economic Activity, Microeconomics:* 1–38.

Triplett, Jack E. 1992. Banking output. In *The new Palgrave dictionary of money and finance,* vol. 1, ed. Peter Newman, Murray Milgate, and John Eatwell, 143–46. New York: Stockton.

———. 1997. Measuring consumption: The post-1973 slowdown and the research issues. *Review of the Federal Reserve Bank of St. Louis* 79 (3): 9–42.

———. 1998. The mismeasurement hypothesis and the productivity slowdown. Paper presented at *International Conference on Information and Communications Technologies, Employment and Earnings.* 22–23 June, *Sophia Antipolis, France.*

———. 1999. The Solow productivity paradox: What do computers do to productivity? *Canadian Journal of Economics* 32 (2): 309–34.

Triplett, Jack E., and Barry Bosworth. 2001. Productivity in the services sector. In *Services in the International Economy,* ed. Robert M. Stern. Ann Arbor: University of Michigan Press. Forthcoming.