NBER WORKING PAPER SERIES

AGGREGATE PRODUCTIVITY AND THE PRODUCTIVITY OF AGGREGATES

Susanto Basu
John G. Fernald

Working Paper 5382

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 1995

We thank Martin Eichenbaum, Robert Hall, Michael Horvath, and Miles Kimball for valuable suggestions. Seminar participants at Boston University, the Federal Reserve Board, the Federal Reserve Bank of San Francisco, Johns Hopkins University, the NBER, New York University, Stanford University, UC-Berkeley, UC-San Diego, the University of Michigan, and the University of Pennsylvania provided helpful comments. The first draft of this paper was completed while Basu was a National Fellow at the Hoover Institution, which he thanks for its hospitality. Basu gratefully acknowledges financial support from the National Science Foundation. This paper is part of NBER's research programs in Economic Fluctuations and Monetary Economics. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Working Paper 5382
December 1995

AGGREGATE PRODUCTIVITY AND THE PRODUCTIVITY OF AGGREGATES

ABSTRACT

Explanations of procyclical productivity play a key role in a variety of business-cycle models. Most of these models, however, explain this procyclicality within a representative-firm paradigm. This procedure is misleading. We decompose aggregate productivity changes into several terms, each of which has an economic interpretation. However, many of these terms measure composition effects such as reallocations of inputs across productive units. We apply this decomposition to U.S. data by aggregating from roughly the two-digit level to the private economy. We find that the compositional terms are significantly procyclical. Controlling for these terms virtually eliminates the evidence for increasing returns to scale, and implies that input growth is uncorrelated with technology change.

Susanto Basu
Department of Economics
University of Michigan
611 Tappan Street
Ann Arbor, MI 48109-1220
and NBER

John G. Fernald
Division of International Finance
Federal Reserve Board
Mailstop 20
Washington, DC 20551

Aggregate productivity is procyclical. Explaining this stylized fact is crucial for explaining business cycles; indeed, choosing among the explanations for cyclical productivity is almost tantamount to choosing among the major hypotheses in business-cycle theory.

There are three prominent explanations for procyclical productivity. First, observed changes in productivity may reflect exogenous changes in efficiency — technology shocks — that drive the cycle. Second, cyclical productivity may reflect endogenous changes in efficiency that occur because the economy operates with increasing returns to scale; with increasing returns, productivity rises whenever inputs rise.[1] Third, changes in measured productivity may in fact be caused by systematic, unmeasured changes in capacity utilization or labor effort. In booms, actual input use then rises more than we observe.

These three explanations have different implications for the impulses and propagation mechanisms driving business cycles. Standard real-business-cycle models show how shocks to technology can serve as impulses leading to fluctuations even in a competitive economy. Models based on imperfect competition and increasing returns, by contrast, can potentially generate endogenous fluctuations driven by sunspots or self-fulfilling changes in expectations if increasing returns are sufficiently large. In any case, the existence of imperfect competition and increasing returns helps propagate exogenous shocks throughout the economy. On the other hand, measured changes in productivity driven by variations in utilization are not equivalent to procyclical efficiency, however defined. Nevertheless, cyclical utilization can also serve as a propagation mechanism for exogenous shocks. These models complement traditional Keynesian theories of money-driven fluctuations by explaining why such fluctuations would induce measured changes in productivity.

Macroeconomic models embodying one or more of these interpretations generally share one common feature, however: they assume the existence of a representative producer. Of course, no one assumes that exact aggregation is possible, but one generally hopes that failures of aggregation will not lead to first-order problems in estimating and calibrating macro models. We show that this hope

---

[1] Imperfect competition without increasing returns can also lead to procyclical productivity since productivity calculations incorrectly weight the contribution of different inputs (Hall, 1988). Given the absence of large pure profits, however, we view a significant degree of imperfect competition as possible only with increasing returns. This seems to be the view of Hall (1990).

is in vain. Aggregation bias affects the measurement of many parameters that are critical inputs to recent business-cycle models.

The intuition for this conclusion is straightforward. We find that sectors of the U.S. economy are characterized by significant heterogeneity in productivity levels, factor prices, and returns to scale. We also find that inputs flow predictably from low- to high-productivity uses as the business cycle goes from trough to peak. For example, the capital-intensive durable-goods industries have comparatively high returns to scale. Durable-goods output is also highly procyclical. Thus, much of the procyclicality of productivity (especially within manufacturing) comes from the reallocation of inputs to these industries over the business cycle.

Therefore, interpreting aggregate data within a representative-firm paradigm leads to highly misleading conclusions. In a real-business-cycle context, one can misinterpret the positive correlation between productivity and output as evidence for the importance of high-frequency technology shocks, even if demand shocks drive business cycles. In an increasing-returns setup one might observe the same correlation, and conclude that each firm produces with very large returns to scale. But this conclusion would also be wrong, since it confuses reallocation of inputs between firms with production at a single representative firm.

On the other hand, aggregation effects allow us to explain several empirical puzzles in the recent macro productivity literature. For example, we can explain why estimated returns to scale are larger at higher levels of aggregation without invoking high-frequency technological spillovers.[2] This literature also finds strongly diminishing returns to scale in several industries (notably non-durable manufacturing).[3] As a statement about firm-level parameters this finding is a puzzle, but in our framework it is easy to explain.

From the standpoint of calculating technical change or estimating structural parameters, composition effects are just a source of bias. But aggregate productivity change is a meaningful economic quantity, even though it cannot be given a production-function interpretation. It represents the change in an economy's ability to produce final consumption or investment goods from a given

[2] Caballero and Lyons (1992) offer the spillovers interpretation.
[3] For example, Burnside (1994).

quantity of primary inputs, and hence is a natural measure of welfare change. Under the standard conditions of perfect competition and constant returns, productivity changes only if technology changes. But in general, technical progress is only one of many ways in which an economy can produce more output from given primary inputs. In particular, with imperfect competition, increasing returns, or factor immobility, an economy can produce more output from a more efficient distribution of existing inputs. This improved distribution contributes to aggregate productivity growth.

Although largely ignored by the recent macroeconomic literature, some previous productivity literature makes related points. For example, Jorgenson, Gollop and Fraumeni (1987) decompose aggregate productivity change into sectoral technology changes plus reallocation effects. They focus on long-run growth, however, and do not present results at a cyclical frequency. They also assume constant returns and perfect competition. We, by contrast, find that in U.S. data, the effects of small sectoral deviations from constant returns and perfect competition are the most important causes of the differences between aggregate productivity and aggregate technology. Some industry-level work does study cyclical productivity, and also emphasizes issues of aggregation.[4] By their nature, however, these papers cannot examine the implications for aggregate productivity, and are typically constrained by data limitations to study labor productivity rather than total factor productivity. TFP is, however, the right concept for studying most macroeconomic issues. There is also a microeconomic literature that derives necessary conditions for an aggregate production function to exist when technology is embodied in vintage capital or when some factors are not mobile.[5] This literature is less relevant, since the conditions necessary to aggregate production functions are much more rigorous than those needed to aggregate productivity growth.

The paper is structured as follows. In Section I, we provide simple examples to motivate our discussion. In Section II, we present the determinants, technological and otherwise, of firm-level value-added productivity. Section III shows how aggregate productivity is related to firm-level productivity, demonstrating how imperfect competition and reallocation effects change aggregate

[4] See Olley and Pakes (1992), Aizcorbe and Kozicki (1995), and especially Bertin, Bresnahan and Raff (1995).
[5] For example, Fisher (1993) and Sato (1975).

productivity. In Section IV, we present examples suggesting that the standard measure of aggregate productivity, although not generally a measure of technology, does have economic interpretation as a measure of welfare. In Section V, we discuss the data and parameters we use to implement our productivity decomposition. Section VI contains our results, and Section VII concludes.

I. Composition Biases in Aggregate Data: Some Examples

As we argued, composition effects can cause one to draw highly misleading inferences from aggregate data. In this section, we illustrate this point with several examples. For simplicity, our examples use only one of the biases we identify in our general derivation of Section III.

Consider an industry that comprises two firms, A and B. Both have the production function

$$Y_i = L_i^{\gamma_i} \quad i = A, B.$$

Firm A has returns to scale of 1.5 ($\gamma_A = 1.5$) and firm B has returns to scale of 2 ($\gamma_B = 2$). Aggregate industry data on input and output are the sums of the inputs and outputs of the two firms.

For simplicity, we assume that only one firm produces at any given time. If both firms made exactly the same product, a social planner would always allocate all inputs to firm B. But in a decentralized equilibrium with imperfect competition the social optimum will generally not prevail. Also, it is possible that the two firms make slightly different products (e.g. high-quality and low-quality shoes) that are nevertheless lumped together into a single aggregate industry, much as Lincolns and Chevrolets are treated as one product, Motor Vehicles. In this case, even the social optimum could require that both firms produce.

Our first example gives a possible distribution of firm and industry inputs and outputs across booms and recessions:

	Labor in Busts	Labor in Booms	Output in Busts	Output in Booms
Firm A	0	3	0	5.2
Firm B	2	0	4	0
Industry	2	3	4	5.2

In this example, firm A produces only in booms and firm B only in recessions. Now suppose a researcher uses any standard method (e.g. the one proposed by Hall [1990]) to estimate the degree of returns to scale of the "average firm" in the industry. One can estimate firm-level returns to scale, $\hat{\gamma}$, using firm-level data, as $\frac{\Delta \ln Y}{\Delta \ln L}$. Applying this method to industry data, however, gives an estimated returns to scale of 0.64.

This example thus demonstrates two important points. First, the returns to scale estimated from aggregate data need not bear any resemblance to the firm-level parameters one wants to estimate. Second, the aggregate data can lead one to find strongly *diminishing* returns to scale in an industry where all firms have *increasing* returns. Hence, composition effects can explain one of the puzzles we noted in the introduction. This is important, since diminishing returns makes no economic sense as a firm-level statement: it implies that firms produce, on average, above efficient scale. Also, profit-maximization implies that price must equal or exceed marginal cost, while diminishing returns means marginal cost exceeds average cost. Hence, returns to scale of, say, 0.7 (which are sometimes estimated) imply that at least 30 percent of output is pure profit. There is no evidence of such huge profit rates in U.S. data.

Next, suppose we reverse the distribution of inputs while keeping aggregate inputs the same. Now inputs and outputs are given by:

	Labor in Busts	Labor in Booms	Output in Busts	Output in Booms
Firm A	2	0	2.83	0
Firm B	0	3	0	9
Industry	2	3	2.83	9

Suppose we again estimate returns to scale from industry data. We find $\hat{\gamma} = 2.88$ — now much larger than the returns to scale of either firm. Interestingly, this difference between sectoral and aggregate returns to scale is a stylized fact documented for U.S. manufacturing by Caballero and Lyons (1992). They interpret the higher degree of returns to scale in aggregate data as evidence for productive spillovers between sectors that are internalized at higher levels of aggregation. As this

example shows, their interpretation need not be correct: returns to scale using aggregate data may be overestimated because of composition biases. This result leads us to hope that we can explain another puzzle: why returns to scale estimates rise with the level of aggregation.

One might ask how these two effects can simultaneously be at work. We conjecture that the first effect is at work at a very disaggregated level — within narrowly-defined industry groups, recessions "cleanse" industries of inefficient firms, as in the model of Caballero and Hammour (1994). But we see the second effect at work at higher levels of aggregation. For example, there are good economic reasons for durable-goods output to be procyclical relative to non-durables output. We also find that durable-goods industries have larger average returns to scale. This combination of factors makes aggregate productivity seem procyclical for the reasons given in the second example.

As these examples show, composition effects can potentially explain many of the puzzles of cyclical productivity. We now derive the full relationship between aggregate productivity and firm-level technology and begin to apply it to the data in an effort to see whether composition effects are empirically important.

II. Firm-Level Productivity

This section analyzes the determinants of measured firm-level productivity growth. The next section then uses these microfoundations to analyze aggregate productivity growth. At both the firm and aggregate level, we measure productivity in terms of real value added output and primary inputs of capital and labor. At a disaggregated level, the use of value-added data requires some explanation, since real value added is an artificial construct. The natural measure of firm output is gross output, not value added. Real value added is bread without flour; books without paper or ink; shoes lacking leather. Technology shocks, as well, are most naturally thought of as affecting the ability of a firm or sector to produce gross output, not the ability to produce value added.

It is useful, however, to focus on firm-level value added because of our ultimate interest in aggregates. The natural measure of aggregate output is aggregate final expenditure, since this

measures the amount that society can consume today or save for tomorrow, i.e., the aggregate of private and public consumption, investment, and net exports. Because of the national accounts identity, we know that aggregate final expenditure equals aggregate firm-level value added: intermediate-input use cancels out, since the quantity of goods and services that are sold as intermediate inputs to other firms necessarily equals the amount of intermediate inputs that these firms purchase. Thus, aggregating over firm value added allows us to derive an economically sensible aggregate.

In nominal terms, it is clear how to define firm-level value added and hence the aggregate index. Nominal value added in a firm, $P_i^V V_i$, is defined as the difference between the value of gross output and the cost of the intermediate inputs used to produce it:

$$P_i^V V_i = Q_i Y_i - P_{Mi} M_i.$$

Y_i is sectoral gross output and Q_i is its price; M_i is the quantity of intermediate inputs used in a sector and P_{Mi} is its price. Aggregate value added is then the arithmetic sum of firm-level value added.

In real terms, however, there are several internally consistent index-number methods that can be used to calculate constant-dollar measures of firm value added and aggregate final expenditure. The national accounts identity holds in constant prices as long as we use the same index number method to calculate real value added as we use to calculate real final expenditure. Analytically, the Divisia index method is the most useful.[6] Divisia indices are defined in terms of growth rates, so define dy_i and dm_i as the growth rates of gross output and intermediate inputs. Then the Divisia definition of real value added growth, dv_i, is implicitly defined by writing output growth as a weighted sum of

[6] In 1995, the Bureau of Economic Analysis announced plans to report real GDP in the National Income and Product Accounts (NIPA) as a chain-linked Fisher index, so that sectoral value added will also be calculated as chain-linked indices. Chain-linked Fisher indices are one method of approximating in discrete time the continuous-time Divisia definitions used here. NIPA historically used a Laspeyres (or double-deflated) index of value added. Divisia indices (or their discrete time approximations) have better index number properties on both the expenditure and the product side, leading NIPA to change its accounting methods. Of particular relevance here, the Divisia method is also analytically simpler for productivity analysis. Nevertheless, none of the analytic conclusions in this section or the next would be affected by the use of other measures of value added; the algebraic expressions would be identical except that there would be additional additive terms. See, for example, Basu and Fernald (1995a, Appendix).

intermediate input growth and output growth, using as weights shares in the value of gross output. Hence, we can write this as:

$$dv_i = \frac{Q_i Y_i}{P_i^V V_i} dy_i - \frac{P_{Mi} M_i}{P_i^V V_i} dm_i = dy_i - \left(\frac{P_{Mi} M_i}{P_i^V V_i} \right) (dm_i - dy_i). \quad (1)$$

As one would expect, this expression tells us that if intermediate inputs grow at the same rate as gross output, then value added grows at this same rate. Similarly, if intermediate inputs grow faster than gross output, then value added grows slower than gross output.

As we emphasize below, this standard definition of real value added does not in general have an interpretation as a measure of production. It is useful as a national accounting device, however, since it properly accounts for the fact that the aggregate quantity of output used as intermediate input equals the aggregate quantity of intermediate inputs used by all firms. The appropriateness of this standard construction of real value added does not require any assumptions about optimizing behavior, let alone any assumptions about technology or market structure. Because national expenditure is closely related to welfare, the aggregate of this standard definition of firm real value added is also useful for studying welfare. (We pursue this point in Section IV.)

By making assumptions about cost-minimizing behavior, production technology, and market structure, we can relate changes in real value added to changes in inputs and technology. We begin by specifying the primals of the production technology. We assume that production by each firm is characterized by a gross-output production function:

$$Y_i = F^i(K_i, L_i, M_i, T_i), \quad (2)$$

where Y is gross output, K, L and M are inputs of capital, labor, and materials, and T is an index of technology. The firm's production function F may be homogeneous of arbitrary degree γ in K, L, and M. γ is not constrained to be one, so F may have non-constant returns to scale. Output by each firm may be sold with a markup, μ, above marginal cost: $\mu_i = P_i / MC_i$, where P is price and MC is marginal cost. There are N such firms.[7]

[7] We shall assume throughout that the number of firms is fixed. This assumption greatly simplifies the derivations in Section III, below, and does not constrain our empirical implementation.

We now define several measures of input shares that are useful in the derivations below. First, define s_{Ji} as the share of costs for input J $(J=K,L,M)$ in total *revenue* of firm i:

$$s_{Ji} \equiv \frac{P_{Ji}J_i}{Q_iY_i}. \tag{3}$$

Second, define c_{Ji} as the share of cost of input J in total *cost* of firm i:

$$c_{Ji} \equiv \frac{P_{Ji}J_i}{P_{Ki}K_i + P_{Li}L_i + P_{Mi}M_i}. \tag{4}$$

Third, define "value added" cost shares, c^V_{Ji}, as the shares of cost of inputs J $(J=K,L)$ in total primary-input cost:

$$c^V_{Ji} \equiv \frac{P_{Ji}J_i}{P_{Ki}K_i + P_{Li}L_i}. \tag{5}$$

In equations (3) through (5), the input prices are all defined as market prices or market rental rates. In particular, if the firm makes economic profits that are paid out to owners of capital, these profits are excluded from the rental price of capital. We assume that firms are price takers in factor markets, so the observed prices of labor and materials inputs equal the cost of these inputs. Thus the only difference between total revenue and total cost lies in the treatment of payments to capital. Rather than assuming that required payments to capital are identically equal to the residual after other factors are paid, we construct a required rental rate series for the capital stock of each sector. As well as assuming price-taking in factor markets, we assume that all factors are freely variable (i.e., there are no quasi-fixed factors).[8]

Following Hall (1990), we decompose output growth into the contribution of inputs plus the contribution of technology shocks. Cost minimization[9] implies that the growth rate of output, dy, equals returns to scale, γ, multiplied by the cost-share-weighted growth in inputs, dx, plus gross-

[8] Quasi-fixity of inputs, where factors are sunk in the short run, matters here only if we allow for time-variation of the cost shares in equation (5). In continuous time, the shares are of course constant for infinitesimal changes. We also assume they are constant in discrete time for finite changes. This can be viewed as a first-order log-linear approximation to equation (5). To a first approximation all production functions are Cobb-Douglas and the Cobb-Douglas production function implies that output elasticities are constant, making quasi-fixity irrelevant for productivity calculations.

[9] Contrary to some of the statements in the literature, the derivation does not require profit-maximization. Hence the relationship we derive below is robust to any form of price-setting behavior; for example, it allows for sticky output prices and for complex dynamic pricing strategies derived from supergames (e.g. Rotemberg and Saloner [1986]). See Basu and Fernald (1995b, Appendix).

output-augmenting productivity growth, $\frac{F_T T}{F} dt$. That is, if *dl*, *dk*, and *dm* are the growth rates of *L*, *K*, and *M*, then

$$\begin{aligned} dy_i &= \gamma_i \left[c_{Li} dl_i + c_{Ki} dk_i + (1 - c_{Li} - c_{Ki}) dm_i \right] + \frac{F_T^i T_i}{F^i} dt_i \\ &\equiv \gamma_i \, dx_i + \frac{F_T^i T_i}{F^i} dt_i . \end{aligned} \tag{6}$$

This is the standard Hall estimating equation for returns to scale. Since our ultimate interest is in value added, it is useful to rewrite this equation in terms of primary input growth and the growth in the materials-to-output ratio. First, write this as:

$$dy_i = \gamma_i (1 - c_{Mi}) dx_i^V + \gamma_i \, c_{Mi} dm_i + \frac{F_T^i T_i}{F^i} dt_i , \tag{7}$$

where dx_i^V equals the cost-share-weighted growth of primary inputs. Second, subtract $\gamma_i c_{Mi} dy_i$ from both sides, and divide through by $1 - \gamma_i c_{Mi}$. Output growth can then be written as:

$$dy_i = \left[\frac{\gamma_i (1 - c_{Mi})}{1 - \gamma_i c_{Mi}} \right] dx_i^V + \left[\frac{\gamma_i c_{Mi}}{1 - \gamma_i c_{Mi}} \right] (dm_i - dy_i) + \frac{F_T^i T_i}{F^i} \frac{dt_i}{1 - \gamma_i c_{Mi}} . \tag{8}$$

Hence, we can write the growth rate of real value added from equation (1) as follows:

$$dv_i = dy_i - \frac{s_{Mi}}{1 - s_{Mi}} (dm_i - dy_i) = \left[\frac{\gamma_i (1 - c_{Mi})}{1 - \gamma_i c_{Mi}} \right] dx_i^V + \left[\frac{\gamma_i c_{Mi}}{1 - \gamma_i c_{Mi}} - \frac{s_{Mi}}{1 - s_{Mi}} \right] (dm_i - dy_i) + \frac{F_T^i T_i}{F^i} \frac{dt_i}{1 - \gamma_i c_{Mi}} . \tag{9}$$

To aid interpretation, note that the first-order conditions for firm optimization imply that $\gamma_i c_{Mi} = \mu_i s_{Mi}$, so we can write this expression as

$$dv_i = \left[\frac{\gamma_i (1 - c_{Mi})}{1 - \gamma_i c_{Mi}} \right] dx_i^V + (\mu_i - 1) \left[\frac{s_{Mi}}{(1 - \mu_i s_{Mi})(1 - s_{Mi})} \right] (dm_i - dy_i) + \frac{F_T^i T_i}{F^i} \frac{dt_i}{1 - \gamma_i c_{Mi}} . \tag{10}$$

Although this expression looks complicated, we can make several qualitative statements about the terms on the right-hand-side. First, under constant returns and perfect competition, the first term equals dx_i^V, and the second term disappears. Hence, with constant returns and perfect competition, value-added growth equals primary input growth plus technological change, and it may be reasonable to interpret value added as a measure of "net output." Second, the term multiplying $(dm_i - dy_i)$ is necessarily non-negative, and is positive in the presence of imperfect competition, when μ_i exceeds

one. Third, if returns to scale are not constant, then the term multiplying primary input growth is mapped away from unity (for example, if γ_i equals 1.1 and c_{Mi} equals 0.6, then this term equals about 1.29.)

Although it is not necessary to make further assumptions about the firm's technology, doing so allows us to provide economic interpretation to the terms in equation (10). Suppose that the production function in (2) takes the following separable form:

$$Y_i = F^i(K_i, L_i, M_i, T_i) = G^i\left(V^{Pi}(K_i, L_i, T_i), H^i(M_i)\right), \tag{2'}$$

Following the logic we used to derive equation (6), we can write dv^P in terms of the cost-weighted growth in primary inputs dx^V, plus technology shocks (without loss of generality we normalize to one the elasticity of productive value added V^P with respect to technology):

$$dv_i^P = \gamma_i^P dx_i^V + dt_i. \tag{11}$$

γ_i^V equals the sum of elasticities of V^P with respect to capital and labor. We cannot, in general, make any statements about the magnitude of this parameter. To do so, we make the further substantive assumption that all returns to scale are in V^P, arising perhaps from overhead capital or labor. This requires that G be homogeneous of degree one in V^P and H, and that H be homogeneous of degree one in M. The sum of output elasticities with respect to all inputs is γ, which in turn is the sum of $(1-\gamma c_M)\gamma^V$ and γc_M. Hence, the relationship between γ and γ^V is

$$\gamma_i^V = \gamma_i \frac{1-c_{Mi}}{1-\gamma_i c_{Mi}}. \tag{12}$$

Returning to equation (10), we can now rewrite it as follows:

$$dv_i = \gamma_i^V dx_i^V + \left[\frac{\gamma_i^V c_{Mi}}{(1-c_{Mi})} - \frac{s_{Mi}}{(1-s_{Mi})}\right](dm_i - dy_i) + dt_i. \tag{13}$$

Real value-added growth depends on primary input growth, changes in the materials-to-output ratio, and technology. The first term shows that primary inputs are multiplied by value-added returns to scale. The second term reflects the extent to which the standard measure of value added differs from V^P, and hence does not properly measure the productive contribution of intermediate inputs. Intuitively, the standard measure of value added subtracts off intermediate input growth using

revenue shares, whereas with imperfect competition the productive contribution of these inputs exceeds the revenue share. The third term is the value-added-augmenting technology shock.

Our focus is productivity measurement, so we now define the firm's productivity residual. We follow Hall (1990), and define the cost-weighted value-added productivity residual, dp, as $dv - dx^V$. Hence,

$$dp_i = (\gamma^V - 1)dx_i^V + \left[\frac{\gamma^V c_{Mi}}{(1-c_{Mi})} - \frac{s_{Mi}}{(1-s_{Mi})}\right](dm_i - dy_i) + dt_i. \qquad (14)$$

Firm-level productivity growth measured in terms of value added depends in part on returns to scale, as emphasized by Hall. In the presence of imperfect competition, however, productivity growth also depends positively on changes in the relative intensity of intermediate-input use.

III. Aggregate Productivity Measurement with Increasing Returns and Imperfect Competition

We now aggregate from the firm level to the economy-wide level. We find that technology shocks are only one contributor to changes in measured aggregate productivity. Other non-technological effects on aggregate productivity in general depend on changes in aggregate primary inputs, changes in the average intensity of intermediate input use, and finally changes in the distribution of inputs and outputs across firms.

Hence, aggregate productivity in general measures many things other than technology. In particular, aggregate productivity growth may be procyclical even if aggregate technology is fixed and the average sector has constant returns to scale. The cyclicality of aggregate productivity depends especially on changes in the distribution of inputs and output across sectors, and indicates changes in the economy's ability to produce final consumption goods from given quantities of primary inputs. Thus, we conclude that the common practice of using changes in aggregate productivity to calculate technology change or estimate returns to scale is seriously flawed.

A. Definitions

We first define the relationships between firm-level and aggregate quantities.[10] Aggregate inputs are defined as simple sums of the firm-level quantities:

$$K \equiv \sum_{i=1}^{N} K_i,$$
$$L \equiv \sum_{i=1}^{N} L_i.$$

These definitions are consistent with the methods used by the BEA to construct its estimates of the capital stock and labor input.

We define the aggregate (rental) prices of capital and labor as the factor payments divided by aggregate quantities:

$$P_K \equiv \frac{\sum_{i=1}^{N} P_{Ki} K_i}{K},$$
$$P_L \equiv \frac{\sum_{i=1}^{N} P_{Li} L_i}{L}.$$

We define the growth rate of aggregate value added as a Divisia index of the underlying sectoral value-added growth rates:

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

where w_i is the sector's share of nominal value added:

$$w_i \equiv \frac{P_i^V V_i}{\sum_{i=1}^{N} P_i^V V_i}.$$

We now define the analogue of w_i for an industry's share in the total *cost* of producing aggregate value added:

$$w_i^c \equiv \frac{P_{Ki} K_i + P_{Li} L_i}{\sum_{i=1}^{N} \left(P_{Ki} K_i + P_{Li} L_i \right)}.$$

We allow each firm to face different input costs for capital and labor; this would occur, for example, if there were monopoly unions with different degrees of market power in different sectors. Allowing for sector-specific prices, we can now define the aggregate cost shares for capital and labor in the cost of producing value added:

---

[10] As noted above, we abstract from entry and exit and assume a fixed number of firms.

$$c_K^V \equiv \frac{\sum_{i=1}^N P_{Ki}K_i}{\sum_{i=1}^N \left(P_{Ki}K_i + P_{Li}L_i\right)},$$

$$c_L^V \equiv \frac{\sum_{i=1}^N P_{Li}L_i}{\sum_{i=1}^N \left(P_{Ki}K_i + P_{Li}L_i\right)}.$$

B. Derivation

Our objective is to measure the rate of technological progress at the aggregate level. We follow Hall (1990) and calculate a cost-based Solow (1957) residual. Hence aggregate productivity growth, dp, is defined as:

$$dp \equiv dv - c_K^V dk - c_L^V dl. \tag{15}$$

With some algebraic manipulation, we can write the product of the aggregate labor cost share c_L^V and the growth of aggregate labor input dl as

$$c_L^V dl = \sum_{i=1}^N w_i^c c_{Li}^V \left[\frac{P_L}{P_{Li}}\right] dl_i \ .$$

Deriving the analogue for capital, we can now sum sectoral productivity growth dp_i using weights w_i^c:

$$\sum_{i=1}^N w_i^c \, dp_i = \sum_{i=1}^N w_i^c \, dv_i - \sum_{i=1}^N w_i^c c_{Li}^V \left[\frac{P_L}{P_{Li}}\right] dl_i + \sum_{i=1}^N w_i^c c_{Ki}^V \left[\frac{P_K}{P_{Ki}}\right] dk_i .$$

By adding and subtracting aggregate productivity, dp, from the right-hand side, we can rewrite this expression as

$$\begin{aligned} \sum_{i=1}^N w_i^c \, dp_i = dp &+ \sum_{i=1}^N \left(w_i^c - w_i\right) dv_i \\ &- \sum_{i=1}^N w_i^c c_{Li}^V \left[\frac{P_{Li} - P_L}{P_{Li}}\right] dl_i - \sum_{i=1}^N w_i^c c_{Ki}^V \left[\frac{P_{Ki} - P_K}{P_{Ki}}\right] dk_i \end{aligned} \tag{16}$$

Hence aggregate productivity growth equals:

$$\begin{aligned} dp \ = \ &\sum_{i=1}^N w_i^c \, dp_i + \sum_{i=1}^N \left(w_i - w_i^c\right) dv_i \\ &+ \sum_{i=1}^N w_i^c c_{Li}^V \left[\frac{P_{Li} - P_L}{P_{Li}}\right] dl_i + \sum_{i=1}^N w_i^c c_{Ki}^V \left[\frac{P_{Ki} - P_K}{P_{Ki}}\right] dk_i \end{aligned} \tag{17}$$

The aggregate residual is a weighted sum of the sectoral residuals, plus input and output reallocation terms. Substituting our final expression for sectoral productivity, equation (14), into equation (17), we find:

$$dp = \sum_{i=1}^{N} w_i^c \left(\gamma_i^V - 1\right) dx_i^V + \sum_{i=1}^{N} w_i^c \left[\frac{\gamma^V c_M^i}{\left(1 - c_M^i\right)} - \frac{s_M^i}{\left(1 - s_M^i\right)}\right] \left(dm_i - dy_i\right)$$
$$+ \sum_{i=1}^{N} \left(w_i - w_i^c\right) dv_i + \sum_{i=1}^{N} w_i^c c_{Li}^V \left[\frac{P_{Li} - P_L}{P_{Li}}\right] dl_i + \sum_{i=1}^{N} w_i^c c_{Ki}^V \left[\frac{P_{Ki} - P_K}{P_{Ki}}\right] dk_i$$
$$+ \sum_{i=1}^{N} w_i^c \, dt_i \tag{18}$$

Equation (18) forms the basis for our discussion of aggregate productivity. It shows that in general aggregate productivity measures many things other than the weighted average of sectoral technology shocks. Nevertheless, (18) shows that aggregate productivity does measure only technological progress if there are constant returns and perfect competition *in every sector* as well as perfect competition and free mobility in factor markets. Under these assumptions total cost equals total revenue. Thus the first and second terms on the right-hand side of (18) are clearly zero: returns to scale are 1 and $c_M = s_M$. Since there are no profits $w_i^c = w_i$, so the third term disappears, The third and fourth terms are zero if labor and capital receive the same wages and rents in all sectors. With perfect competition and free mobility in factor markets, these last two terms are zero as well.[11]

We can decompose these terms further, separating out the mean effects of the first two terms from their reallocation effects. For example, note that two conceptually distinct effects are at work in the first term of (18). First, if every sector has identical returns to scale and if this degree of returns to scale is larger than 1, then productivity is procyclical because each sector is taking advantage of increasing returns. This is the effect stressed by Hall (1990). On the other hand, suppose that sectors have different degrees of returns to scale but the mean returns to scale is zero. Even so, the first term can contribute to aggregate productivity growth if the sectors with above-average returns to scale experience above-average input growth: this is just the intuition of our example in Section I.

[11] See also Jorgenson, Gollop and Fraumeni (1987), who derive an analogous equation under the assumption of constant returns to scale in production and perfect competition in output markets. Thus the first three terms in (18) are identically zero in their setup; the fourth and fifth terms may be non-zero.

Define the following averages:

$$\bar{\gamma}^V \equiv \sum_{i=1}^N w_i^c \gamma_i^V,$$

$$\bar{\rho} \equiv \sum_{i=1}^N w_i^c \left[\frac{\gamma^V c_M^i}{(1-c_M^i)} - \frac{s_M^i}{(1-s_M^i)} \right],$$

and

$$\overline{(dm-dy)} \equiv \sum_{i=1}^N w_i^c (dm_i - dy_i).$$

Substituting these definitions into (18), we find

$$\begin{aligned} dp = {} & (\bar{\gamma}^V - 1)dx^V + \sum_{i=1}^N w_i^c (\gamma_i^V - \bar{\gamma}^V)(dx_i^V - dx^V) \\ & + \bar{\rho}\overline{(dm-dy)} + \sum_{i=1}^N w_i^c \left[\left(\frac{\gamma^V c_M^i}{(1-c_M^i)} - \frac{s_M^i}{(1-s_M^i)} \right) - \bar{\rho} \right] \left[(dm_i - dy_i) - \overline{(dm-dy)} \right] \\ & + \sum_{i=1}^N (w_i - w_i^c) dv_i + \bar{\gamma}^V \sum_{i=1}^N w_i^c c_{Li}^V \left[\frac{P_{Li} - P_L}{P_{Li}} \right] dl_i + \bar{\gamma}^V \sum_{i=1}^N w_i^c c_{Ki}^V \left[\frac{P_{Ki} - P_K}{P_{Ki}} \right] dk_i \\ & + \sum_{i=1}^N w_i^c dt_i \end{aligned} \tag{19}$$

C. Intuition

Equation (19) is rather complex, and we discuss its economic interpretation in two parts. Here we discuss the intuition for why each term contributes to measured productivity. In Section IV, we present examples of economies in which different terms from equation (19) are economically significant.

The first two terms on the right-hand side of (19) reflect the contribution of increasing returns. As we discussed earlier, there are two such effects: a "mean effect" and a "redistribution (covariance) effect." We find below that the redistribution effect is significant. For example, durable-goods industries are much more cyclical than industries producing non-durables. Basu and Fernald (1995b) also show that durable-goods industries have higher returns to scale. So even though they find that manufacturing industries have constant returns to scale overall, the fact that in booms a greater share of the marginal output is produced by industries with increasing returns can help us explain why aggregate manufacturing productivity is procyclical.

The next two terms represent the extent to which measured real value added depends on the intensity of intermediate-input use. As we discussed in Section II, firm-level value added is useful for national accounting, regardless of technology or market structure. With imperfect competition, however, changes in the materials-to-output ratio in general affect a firm's value added since the marginal product of these intermediate inputs exceeds their cost. Since this otherwise-uncounted marginal product represents real goods, the change in value added in turn affects aggregate output and aggregate productivity. Thus, the degree of vertical integration can have real economic consequences in an economy with imperfect competition. We again separate this effect into a mean and a redistribution effect.

The next three terms, in the third line of equation (19), come from differences in productivity levels (or marginal products) across firms. Suppose one firm is more efficient than another. This difference in efficiency must be reflected in differences in returns to at least one factor: to owners of the firm, in the form of profits; to owners of labor, in the form of wages; or to owners of capital, in the form of rents.

The first of the three terms says that aggregate productivity grows if a firm making above-average profits expands. The high profit rate implies that this firm is unusually efficient in turning inputs into output, so aggregate productivity rises if these efficient firms account for a higher share of output growth. In this context, "efficiency" just means that the firm's output commands a high relative price; the first example in the next section shows that a firm that succeeds in charging high markups is also efficient in this sense. However, we show that it *is* efficient (in the economic sense) for such firms to have above-average output growth.

The next two terms are relatively straightforward. They represent gains in productivity from redistributing factors of production from low-rent to high-rent firms. From the point of view of productivity analysis, this is quite sensible. Suppose that factors are homogeneous, but in one firm they are paid more than in other firms. Since we assume that employers act competitively in factor markets, the firm with high-priced capital and labor uses less of these inputs and thus *ceteris paribus* has higher marginal products of these factors. Then if that firm expands disproportionately, more of

the primary inputs are being used in a firm where they have high marginal product. Naturally aggregate productivity rises.

Why might labor, say, be paid a higher wage in one firm than another? First, efficiency wage considerations may lead to differences in wages across firms in different industries, as emphasized by Katz and Summers (1989). Second, labor may not be fully mobile across sectors. Third, a union with monopoly power might choose to charge different wages to different firms. Whatever the reason, shifting resources from firms where labor is relatively unproductive to firms where labor is relatively more productive increases aggregate output, even if total input does not change.

Finally, from the point of view of applying this derivation to data, note that we have assumed there is no unobserved utilization. That is, we assumed that the dt_i defined in equation (14) is a true firm-level technology shock. Of course, this assumption is almost certainly not true: variable capacity utilization and unobserved changes in labor effort have always been prominent among the proposed explanations for the puzzle of procyclical productivity. Burnside, Eichenbaum and Rebelo (1995) and Basu (1995b) show how to adjust these residuals for capacity utilization. Thus, we should not expect that correcting for aggregation biases alone will give us a true measure of technology change. However, we focus on aggregation effects because that is the novel contribution of our paper.

IV. Welfare Interpretation of the Aggregate Productivity Residual

In this section, we ask whether aggregate productivity is an economically meaningful concept. We illustrate the economic intuition behind equations (18) and (19) via three examples. We show that in fairly general cases the terms in (18) do have natural economic content, and the aggregate Solow residual has an interpretation as a measure of welfare.

Since we have already examined the effects of returns-to-scale differences in Section I, all of our examples feature constant-returns production. Thus, the terms in (18) that depend on $\left(\gamma^V - 1\right)$ are always zero in our examples. In the first and third examples we also abstract from materials use,

so the terms that depend on $(dm - dy)$ are also identically zero. For simplicity, we focus on one-period examples where capital is not a factor of production.[12]

A. Example 1: Markup Pricing and Factor Rents

Suppose the economy consists of two firms, each producing one type of good which is only for final consumption. There are two distortions — markups and labor rents — that each lead firm 1 to produce less than is socially optimal relative to firm 2. In this example, shifting resources towards firm 1 raises both productivity and welfare.

The representative consumer has a Cobb-Douglas utility function over the two types of goods:

$$U = C_1^{\alpha} C_2^{1-\alpha}. \tag{20}$$

The consumer inelastically supplies $\bar{L}$ units of labor. Although labor is homogeneous, workers in each firm may be paid a different wage. In particular, we assume that workers in firm 1 may be able to raise their wages above the Walrasian level. For the purposes of this example, it is not necessary to specify exactly how this is done. The labor market in firm 2 is competitive. Thus the consumer maximizes (19) subject to the budget constraint

$$P_1 C_1 + P_2 C_2 = W_1 L_1 + W_2(\bar{L} - L_1) + \Pi_1 + \Pi_2, \tag{21}$$

where the Π_i are the profits of the two firms. Following our convention, the aggregate wage is defined as:

$$W = \frac{L_1}{L_1 + L_2} W_1 + \frac{L_2}{L_1 + L_2} W_2.$$

Both firms have identical production functions:

$$Y_i = L_i \qquad i = 1,2. \tag{22}$$

However, both firms sell their output at a markup over marginal cost, μ_i. Assume $\mu_1 \geq \mu_2$. Again, it is not necessary to specify the details of the game that allows firms to have markups.

Substituting into equation (18), we can find the growth rate of aggregate productivity, dp, in this economy:

---

[12] In future work, we shall show that under quite general conditions welfare is equivalent to a particular measure of productivity, which reduces to the cost-based Solow residual if there are zero profits. This result holds even when capital is a factor of production and leisure enters the utility function.

$$dp = \left\{ \left[\frac{\mu_1 W_1 L_1}{\mu_1 W_1 L_1 + \mu_2 W_2 L_2} - \frac{W_1 L_1}{W_1 L_1 + W_2 L_2} \right] - \left[\frac{\mu_2 W_2 L_1}{\mu_1 W_1 L_1 + \mu_2 W_2 L_2} - \frac{W_2 L_1}{W_1 L_1 + W_2 L_2} \right] \right\} dl_1$$
$$+ \left\{ \frac{W_1 L_1}{W_1 L_1 + W_2 L_2} \left[\frac{W_1 - W}{W_1} \right] - \frac{W_2 L_1}{W_1 L_1 + W_2 L_2} \left[\frac{W_2 - W}{W_2} \right] \right\} dl_1. \quad (23)$$

Note that productivity increases with labor input in sector 1 if the markup is strictly higher in firm 1 ($\mu_1 > \mu_2$) or if the wage is strictly higher in firm 1 ($W_1 > W_2$) or both. Equation (23) says that a pure reallocation — increasing the size of firm 1 which, since aggregate labor input and technology are constant, must imply a decrease in the size of firm 2 — increases productivity.

Why should we wish to increase aggregate productivity, as defined by the cost-based residual? It is easy to show that productivity thus defined is an exact measure of welfare in this economy. Taking log differences of the utility function (20) and noting that $dL_1 + dL_2 = d\bar{L} = 0$, we find:

$$\begin{aligned} du &= \alpha dy_1 + (1-\alpha) dy_2 \\ &= \frac{P_1 Y_1}{P_1 Y_1 + P_2 Y_2} dy_1 + \frac{P_2 Y_2}{P_1 Y_1 + P_2 Y_2} dy_2 \\ &= dy \\ &= dp \end{aligned} \quad (24)$$

This result is intuitive: since distortions are larger in firm 1 its output is even lower relative to the social optimum than the output of firm 2. Thus welfare increases if we devote relatively more resources to producing good 1. The social production possibilities frontier is a straight line with slope -1. Given the utility function, the social optimum dictates that a fraction α of the total labor of the economy should be devoted to the production of good 1, and $(1-\alpha)$ to the production of good 2. But if $\mu_1 > \mu_2$ or $W_1 > W_2$, then $L_1 < \alpha \bar{L}$, so the representative consumer has lower utility than at the social optimum.[13]

Interestingly, in this example the increase in welfare from optimal policy also corresponds to an increase in productivity as measured by the cost-based Solow residual. Hulten (1978) shows that under perfect competition aggregate productivity represents a welfare improvement because it is a

[13] Our example is related to the literature on domestic distortions and "industrial policy": e.g. Bhagwati, Ramaswami and Srinivasan (1969). Bulow and Summers (1985) update this literature by using efficiency wages to justify intersectoral wage differentials, and Katz and Summers (1989) try to measure labor rents in different industries. We later use the Katz-Summers estimates for our decomposition.

shift of the social PPF coming from changes in production technology. Our example shows that when resource allocation is distorted, increases in productivity as well as welfare can come from shifts along the PPF — better allocation of existing resources — and need not come from shifts of the PPF.

One special feature of this example is that there is no distortion if $\mu_1 = \mu_2$ (and $W_1 = W_2$). This result may be surprising, given that markups of any size usually imply some distortion. That is not the case here because labor is supplied inelastically. If leisure were a third good in the utility function (and the labor market was competitive), then the existence of markups (of any size) would distort the consumption-leisure decision.

B. Example 2: Production of Commodities by Means of (Manufactured) Commodities

We now concentrate on the terms in equation (18) that depend on $(dm - dy)$. In this example, we show that markup pricing of intermediate inputs causes the economy to produce within its production possibilities frontier. Cyclical reductions in markups move the economy closer to the PPF, thereby increasing productivity as well as welfare.

We again consider a two-good example, but to aid intuition we now assume that the goods are perfect substitutes in consumption. So the consumer's utility function is

$$U = C_1 + C_2. \tag{25}$$

Labor is again supplied inelastically, but we now assume that the labor market is competitive and consequently the wage, W, is the same for all firms.

Again suppose there are two firms, but now assume that each firm needs to use materials to produce output. Firms can use either their own output or the other firm's output, but in either case they must purchase the output at the market price. Since the goods are perfect substitutes they have the same price, so we follow the convention that each firm uses the other firm's output as materials input. Both firms mark up their output above marginal cost. Firm 1's production function is

$$Y_1 = L_1^{\alpha} M_1^{1-\alpha}, \tag{26}$$

where M_1 is part of the output of firm 2. Firm 2's production function is

$$Y_2 = L_2^{\beta} M_2^{1-\beta}. \tag{27}$$

We assume $\beta > \alpha$. Since the goods are perfect substitutes $P_1 = P_2$, so $\beta > \alpha$ implies $\mu_2 > \mu_1$.

Since the goods are the same from the standpoint of final consumption, final output is just

$$Y_F = (Y_1 - M_2) + (Y_2 - M_1). \tag{28}$$

Repeated use of the conditions for cost-minimization and some tedious algebra show that

$$\begin{aligned} dp &= dy_F - dl \\ &= \frac{Y_2}{Y_F}\frac{Y_2 - M_2}{Y_2}\frac{(1-\beta)-\frac{M_2}{Y_2}}{\beta\left(1-\frac{M_2}{Y_2}\right)}\beta(dw - dp_1) + \frac{Y_1}{Y_F}\frac{Y_1 - M_1}{Y_1}\frac{(1-\alpha)-\frac{M_1}{Y_1}}{\alpha\left(1-\frac{M_1}{Y_1}\right)}\alpha(dw - dp_2) \\ &\quad + \left[\frac{Y_2 - M_2}{Y_F} - \frac{L_1}{L}\right]dl_2 + \left[\frac{Y_1 - M_1}{Y_F} - \frac{L_2}{L}\right]dl_1 \end{aligned} \tag{29}$$

Some substitution shows that the right-hand side of (29) equals the third and fourth terms on the right-hand side of (18).

We can now give the economic intuition behind these terms. Note that in this model economy there is only one distortion: materials are sold with markups.[14] This has two effects. First, materials are underused relative to primary inputs; thus, aggregate productivity could be increased by shifting to more materials-intensive production. Second, since markups are higher in firm 2, aggregate productivity could be increased by having firm 2 grow relative to firm 1. This would cause firm 1 to become more materials-intensive *relative to firm 2*, which would also increase efficiency.

The first two terms in (29) capture the first effect. Focusing on the first term, note that the difference between $(1-\beta)$ and $\frac{M_2}{Y_2}$ is the difference between the elasticity of output with respect to materials in firm 2 and the share of materials in production in that firm. Materials share is less than its elasticity because of markup pricing in the firm *using* materials: as usual, markup pricing leads to underuse of inputs. Given that materials are being underused, productivity would rise if production became more materials-intensive; the size of the effect is proportional to the markup *in firm 2*. Materials use will increase if the price of good 1, the materials input to firm 2, falls relative to the price of labor: that is, if the markup falls *in firm 1*. (Note that $\beta(dw - dp_1)$ equals $(dm_2 - dy_2)$.) The second term in (29) is symmetric, with the roles of the two firms reversed.

[14] The fact that final consumption goods are priced above marginal cost is not a distortion because labor supply is inelastic. Since the goods are perfect substitutes and are sold at the same price, their different markups also do not cause distortions in the allocation of aggregate consumption across goods.

In terms of understanding equation (18), we find that the contribution of the $(dm - dy)$ term to productivity growth depends on the *level* of the markup in the materials-using firm, and the *cyclicality* of the markup(s) in the materials-supplying firm(s).[15] So while the size of this effect depends on the extent of the inefficiency in the initial equilibrium, we also need the degree of that inefficiency to change over the business cycle. Obviously, the size of the effect depends on the importance of intermediate goods in production, i.e. on the sizes of $(1-\beta)$ and $(1-\alpha)$. It also depends on another technological feature of the production function, the elasticity of substitution between materials and other inputs in production. If the production function is Leontief in materials, as Rotemberg and Woodford (1995) assume, then $dm = dy$ always, and this effect disappears.

In this example, the production possibilities frontier now represents quantities of goods produced for *final* consumption. The distortion from markups implies that the economy generally is *within* the PPF (not just at an inefficient point on it, as in Example 1). Reductions in the markup move the economy closer to the PPF, again increasing both welfare and productivity.

Now we discuss the second effect at work in (29), which is embodied in the third and fourth terms of that equation. We argued that for the first two terms to be positive, we had to have positive markups in the materials-using firm. But what if the production structure were linear, rather than circular as in equations (26) and (27)? That is, instead of all firms using the outputs of other firms in production, suppose there were an upstream firm producing materials using only labor, and a downstream firm using those materials and labor to produce final consumption goods. Suppose furthermore that the upstream firm sells its output with a markup, but the downstream firm does not. As long as the downstream firm can substitute between materials and labor there is clearly a distortion in this economy, but it is not captured by the $(dm - dy)$ terms, which are always zero.[16]

Note, however, that in order to generate the upstream-downstream example we had to invoke asymmetric behavior. The third and fourth terms in (29) (and in general the third term in (18)) capture asymmetries in production and pricing. In the upstream-downstream example, since the

[15] Thus we confirm the intuition of Basu (1995a), who presents a symmetric example where this effect is at work.
[16] These terms are zero because the upstream sector does not use materials, while the downstream sector does not price with a markup.

upstream firm has a value-added share larger than its cost share, relative growth in that firm increases productivity. In our original example of equations (25)-(27), since good 2 is sold with a higher markup than good 1, productivity also rises if firm 2 grows relative to firm 1. The intuition is the same as in Example 1 above.

C. Example 3: Level Effects

Finally, we consider the effects of differences in the levels of productivity across establishments. It might seem that the simplest composition effect is the possible reallocation of inputs between high- and low-productivity plants, but it is not immediately apparent from (18) how this would change our measure of aggregate productivity. This example may be quite significant empirically. Establishment-level studies often find large differences in productivity levels between plants in the same four-digit industry.[17] Baily, Hulten and Campbell (1992) conduct an establishment-level study and show that a large fraction of long-run productivity growth comes from increases in the share of output produced by the most efficient firms. If this effect is also significant at higher frequencies, it may be responsible for a large fraction of the procyclicality of aggregate productivity.

We again use the specification of preferences in (25), in order to focus on an example where both firms produce the same type of good. Now assume

$$Y_i = A_i L_i \qquad i = 1,2. \tag{30}$$

Furthermore, assume $A_1 > A_2$. Again, since the goods are perfect substitutes, $P_1 = P_2$. Then the equilibrium must involve a higher markup in industry 1, or higher wages, profits, or capital rents in industry 1, or both. We have already seen in Example 1 how these differences contribute to both productivity and welfare changes. So differences in productivity levels do affect our measure of productivity growth, because they are captured by one of the three reallocation terms in equation (17).

Note that in order for level effects to matter in this way they must lead to differences in marginal factor products. For example, if the production functions in (30) had diminishing returns

[17] Baily, Hulten and Campbell (1992), Caves and Barton (1990).

to scale, differences in productivity levels would be consistent with a competitive equilibrium in which firm 1 produces relatively more output and marginal products equalize across firms. In that case, despite the differences in productivity levels, there would be no reallocation effects in aggregate productivity. However, Baily, Hulten and Campbell (1992) find that large differences in productivity levels coexist with essentially constant returns to scale. Thus, level effects are probably an important component of composition bias.

V. Data and Method

A. Data

Having discussed the theory of aggregation, we now investigate its empirical significance. We construct a "true" aggregate technology series, the last term in equation (18), and compare it to the conventional measure of aggregate productivity, dp. We do so by constructing estimates of the other terms on the right-hand side of (18) using data at two levels of aggregation. We first define our "aggregate" as the private U.S. economy, and our "firms" as 34 industries at roughly the two-digit SIC level. To facilitate comparison with previous work and to take advantage of the superior data quality in manufacturing, we also close our model within U.S. manufacturing. This makes the "aggregate" total manufacturing, and our "firms" 21 mostly two-digit manufacturing industries.[18]

The average two-digit manufacturing industry comprises about 18,000 firms, so it may seem odd to consider an industry as a firm. We do this for reasons of data availability: there are no firm-level data sets that span the economy. In principle, one could focus on a subset of the economy, and use data from the Longitudinal Research Database, say; doing so, however, requires sacrificing not only breadth, but also panel length and data quality. Our work, by focusing on aggregates, complements existing work (such as Bertin, Bresnahan, and Raff (1995)) that looks at highly disaggregated data for a small subset of the economy. Nevertheless, it is worth emphasizing that

[18] There are only 20 two-digit manufacturing industries. However, we divide S.I.C. 37 into Motor Vehicles (S.I.C. 371) and Other Transport Equipment (S.I.C. 372-79).

aggregation effects of the type we identify are likely to be important even within the disaggregated sectors that we use. We return to this point in the conclusion.

We use unpublished data provided by Dale Jorgenson and Barbara Fraumeni on industry-level inputs and outputs. These data consist of a panel of 34 industries (including 21 manufacturing industries) that constitute the U.S. private business economy for the years 1949-1989. These sectoral accounts seek to provide accounts that are, to the extent possible, consistent with the economic theory of production. Output is measured as gross output, and inputs are separated into capital, labor, energy, and materials. For our purposes, an essential aspect of the data is their inclusion of intermediate inputs. Basu and Fernald (1995a) contains a brief description of the data set; for a complete description, see Jorgenson, Gollop, and Fraumeni (1987).

We need to construct the cost shares defined above. To estimate the required payments to capital, we follow Hall and Jorgenson (1967), Hall (1990), and Caballero and Lyons (1992), and compute a series for the user cost of capital r. The required payment for any type of capital is then $rP_K K$, where $P_K K$ is the current-dollar value of the stock of this type of capital. In each sector, we use data on the current value of the 51 types of capital, plus land and inventories, distinguished by the BEA in constructing the national product accounts. Hence, for each of these 53 assets, the user cost of capital is

$$r_s = (\rho + \delta_s)\frac{(1 - ITC_s - \tau d_s)}{(1-\tau)}, \qquad \text{s = 1 to 53.}$$

ρ is the required rate of return on capital, and δ_s is the depreciation rate for this asset. ITC_s is the asset-specific investment tax credit, τ is the corporate tax rate, and d_s is the asset-specific present value of depreciation allowances. We follow Hall (1990) in assuming that the required return ρ equals the dividend yield on the S&P 500. Jorgenson and Yun (1991) provide data on ITC_s and d_s that is specific to each type of capital good. Given required payments to capital, computing the cost shares is straightforward.

To perform some of the estimation and to assess our series for the aggregate technology shock, we need demand-side instruments. We use versions of the Hall-Ramey instruments: the growth rate of

the price of oil deflated by the GDP deflator; the growth rate of real government defense spending; and the political party of the President.

B. Estimating Returns to Scale

In order to construct the first two terms on the right-hand side of (18), we must know γ^V. Since γ^V is not data, we estimate it for each sector using the method of Basu and Fernald (1995b), which is a correction of the ingenious procedure of Hall (1990).

We actually estimate the gross-output returns to scale, γ for each sector by running the regression in equation (5). In order to avoid the "transmission problem" of correlation between technology shocks and input use, we use the Hall-Ramey instruments noted above. We use the current value and one lag of each instrument.

The estimated values of γ are listed in Table 1. We see that median returns to scale are 0.98, very close to constant returns. Having estimated γ, we construct γ^V using equation (12). Although the γs are assumed to be constant over the sample period, the γ^Vs change over time because we treat the materials share as time-series data. Basu and Fernald (1995b, Appendix) discuss the pros and cons of estimating γ as opposed to estimating γ^V directly.

C. Capital and Labor Rents

In order to estimate the contributions of capital and labor reallocation to productivity growth, we need measures of above-market rents earned by capital and labor.

Our method of constructing required payments to capital implicitly assumes that the user cost of capital does not vary across sectors. Thus, the "reallocation of capital" term is identically zero. This point bears explanation. We certainly allow capital to earn rents: capital is the residual claimant in each sector, and may earn economic profits. This possibility is what forces us to construct the rental rate series above. But for the reallocation term to be non-zero, it is necessary that the rents to capital be allocative: that is, there must be differences in the *required return* to capital across sectors. We assume that there are no such differences.

However, we do allow for allocative labor rents, of the kind stressed by Katz and Summers (1989). They examine payments to categories of labor across industries, controlling for observable differences, and find that some industries exhibit substantial wage premia. They list their estimates of these industry wage premia in their Table 6. We use their estimates to construct the contribution of labor reallocations across industries to aggregate productivity.[19]

VI. Results

We now adjust the cost-based Solow residual dp, as required by equation (18), to get a series dp^* that should correspond to true technology growth. We find that our corrected series is less volatile, and less correlated with both input and output growth. It better fits our priors about technology; for example, it is less correlated with oil price changes or monetary contractions. Composition effects also explain why we find approximately constant returns at the micro level but large increasing returns in macro data.

True aggregate technology growth is the weighted average of the sectoral technology shocks, the last term on the right-hand side of (19). Creating this true technology series requires us to construct the first four terms on the right-hand side of (18) (recall that our method makes the fifth term identically zero), using the methods we discussed in the previous section.[20] We call the sum of these adjustment terms ξ. Thus

$$\begin{aligned}\xi &= \sum_{i=1}^{N} w_i^c \left(\gamma_i^V - 1\right) dx_i^V + \sum_{i=1}^{N} w_i^c \left[\frac{\gamma^V c_M^i - s_M^i}{\left(1-c_M^i\right)\left(1-s_M^i\right)}\right]\left(dm_i - dy_i\right) \\ &\quad + \sum_{i=1}^{N}\left(w_i - w_i^c\right) dv_i + \sum_{i=1}^{N} w_i^c \frac{c_L^i}{1-c_M^i}\left[1 - \frac{\sum_{j=1}^{N} P_L^j L_j}{P_L^i L}\right] dl_i \qquad (31) \\ &\equiv \xi_1 + \xi_2 + \xi_3 + \xi_4.\end{aligned}$$

---

[19] Outside the manufacturing sector, our industry definitions are often different than those used by Katz and Summers. Where they do not cover a particular industry (e.g. agriculture), we set the labor rents to zero. In cases where their industry definitions are broader than ours we assume that all the sub-industries within a particular industry have the same labor rents.

[20] An alternative method would be to aggregate the residuals from estimating equation (6), since by hypothesis they are true technology shocks. However this procedure would not allow us to investigate the different components of our adjustment, nor would it allow us to examine the issue of returns to scale.

We then subtract this sum from the standard measure of productivity, dp, to create the true measure of technology change, dp^*:

$$dp^* = dp - \xi \tag{32}$$

We also examine the properties of a series that includes all the corrections except for the average returns-to-scale effect:

$$\xi' = \xi - \bar{\gamma}^V dx^V. \tag{33}$$

ξ' provides a measure of our new correction to the usual productivity series, since the returns-to-scale adjustment is now almost standard.

As noted above, we close the model at two different levels of aggregation: at the level of the entire private economy, and at the level of one-digit manufacturing.

Table 2a presents summary statistics for aggregate private output growth, dv, weighted average of primary input growth, dx^V, the standard cost-based Solow residual for measuring productivity growth, dp, and our corrected series dp^*. Table 2b is a correlation matrix for these variables.

Table 2a shows that at the aggregate level our corrected technology series, dp^*, has a significantly lower variance than the cost-based Solow residual: about 75 percent of the variance of dp. It also has a considerably lower mean: about 80 percent of aggregate productivity growth actually comes from sectoral miscalculation and aggregation effects, not technology change.

The correlations in Table 2b show some of the striking differences between dp^* and dp. Note first that the two series are fairly highly correlated, with a correlation coefficient of 0.85. But while dp has a correlation with aggregate output growth of 0.84, the corrected measure dp^* has an output correlation of only 0.59. And while the standard measure has a positive correlation with aggregate inputs of 0.25, the corrected measure shows basically no correlation: -0.02.

In Tables 3a and 3b we present the analysis in Tables 2a and 2b, but now closing the model in manufacturing only. The results here are similar: the variance of technology shocks falls by about 25 percent after our corrections, as does mean technology growth (though only about 30 percent). The correlations also tell the same story: the output correlation falls even more than in the economy-wide data (by almost one-half), and the input correlation is again almost exactly zero.

In conjunction, these findings show that the standard technology-driven real-business-cycle model is even farther from fitting the data than one normally thinks. These models anyway tend to generate a higher correlation between output and the Solow residual than the data show: if the true correlation falls by 30 percent or more, then they are that much further from the truth. More importantly, business-cycle models must display the standard characteristic of business cycles: a comovement between output and inputs. If inputs and output move independently in response to a technology shock then technology shocks cannot be the dominant impulse driving business cycles.

However, there is one possible caveat to this conclusion. We have shown that accounting for composition bias reduces the variance of true technology shocks, but it is possible that composition effects themselves are a major new propagation mechanism for technology shocks.[21] For example, a positive technology shock in a multi-sector dynamic general-equilibrium model may boost output by more in durable-goods sectors than in non-durables. So although we show that output is less correlated with *contemporaneous* technology shocks after our composition corrections, it is possible that many of those composition changes are themselves driven by *lagged* technology shocks. Thus, our results might actually support technology-driven models by reducing the size of technology shocks without reducing their ability to explain output fluctuations.

To investigate this possibility, we regress the "reallocation only" component of ξ, ξ', on our derived series of lagged technology shocks. The results are in Table 4. The signs are generally consistent with the hypothesis that composition changes are propagation mechanisms for lagged technology shocks, but the coefficients are small and insignificant.[22]

We proceed to investigate whether the technology shocks we compute better fit our *a priori* notions about technology change. In Tables 5 and 6 we regress the two measures of technology shocks — dp and dp^* — on sources of fluctuations that are plausibly identified as not being

[21] We are indebted to Marty Eichenbaum and Michael Horvath for suggesting this possibility.
[22] We do find that our lagged technology shocks significantly predict aggregate input growth. However, recall that our "technology shocks" also include variations in the utilization of inputs. Most optimizing models of input variation imply that lagged utilization changes should predict future input growth: in the short run, firms react to a demand shock mostly by increasing utilization, but in the long run utilization falls back to its steady-state level and firms accommodate the shock along the extensive margin only.

technological in nature. First, we use current and lagged growth rates of real oil prices.[23] Second, we use two lags of the Romer and Romer (1989) dates that identify monetary contractions.[24] The two series generally respond quite differently to both sets of variables.

At the level of the private economy, dp^* falls by 0.027 percent in response to a one percent rise in oil prices, but the coefficient is not significant and neither is the F-statistic for the regression. The standard series, however, falls about twice times as much — 0.063 percent — and the coefficient and the regression are both significant. The results are even stronger within manufacturing: the elasticity of dp^* with respect to oil prices is again -0.027, while the elasticity for dp is -0.11.

We next look at the responses of these series to monetary contractions. At the economy-wide level, dp^* does fall by 1.7 percentage points in response to a lagged Romer date. This decline is significant, but the F-statistic cannot reject the hypothesis that all coefficients are zero. (The same is true when we use the contemporaneous date as well as two lags.) On the other hand, the standard residual falls by 2.4 percentage points in response to a lagged Romer date, and the coefficient and F-statistic are both significant. This pattern repeats more strongly in manufacturing: dp^* falls by 1.9 percentage points in response to a lagged Romer date, but dp falls by 4.3 percentage points.

Thus, by both the measures we use, dp^* is a better index of technological progress, though both sets of regressions show that it is by no means perfect. Our computed series has some unavoidable errors that come from using estimated rather than population parameters. But we conjecture that the major reason why our corrected "technology" series behaves badly is that it still includes changes in utilization as well as composition changes that occur below the two-digit level of aggregation.

We have argued that aggregate productivity is a bad measure of technology change. We now show that it provides misleading estimates of returns to scale as well. This is a significant issue, since the degree of returns to scale has become a crucial parameter in recent business-cycle models. A number of recent influential papers have modeled business cycles as products of sunspots or

---

[23] Oil prices may or may not represent "supply shocks" — the concept is vague — but they are not technology shocks, since factor price changes do not shift production functions.

[24] We also tried a specification using the contemporaneous date, but the contemporaneous variable was never significant and there are good reasons to think that monetary policy affects the economy only with a lag.

indeterminacy,[25] which can even result in monetary non-neutrality.[26] A high degree of returns to scale is critical for the success of most of these models — their ability to explain business cycle fluctuations is not just reduced but eliminated if returns to scale are below a (high) minimum. Most of these papers assume a production sector where identical firms produce a single commodity with the same technology. Of course, a one-sector model of production is an abstraction from a world with many sectors and heterogeneous technology. Thus, it is unclear how it should be calibrated. One might take the model literally, and calibrate it using estimates from aggregate data alone. Alternatively, one might think that the right way to calibrate such a model is to use the weighted sum of sectoral returns to scale, thereby constructing the "representative producer." Only theory can decide this issue, but we show that the results depend critically on which of these two approaches one takes.

Trying the first approach, we regress aggregate productivity growth in the private economy on aggregate input growth to estimate the degree of returns to scale (actually $\gamma - 1$). The results are in the first line of Table 7. Using the Hall-Ramey instruments, the standard series gives a point estimate for γ of 1.27, which is significantly different from 1 both statistically and economically. Note, however, that this is the degree of returns to scale in the production of gross output. The right parameter for one-sector models of indeterminacy is the degree of returns to scale in the production of real value added — γ^V — which we calculate from γ using equation (8).[27] At the level of the private economy, we calculate γ^V to be 1.98. This is extremely high, though representative of numbers found elsewhere in the literature.[28] Importantly, it is also more than large enough for the sunspot models to display indeterminacy. Schmitt-Gröhé (1994) compares four such models, and concludes that they require returns to scale of at least 1.50 if markups are acyclical, and 1.37 if markups are countercyclical. By contrast, using the second approach of calculating the cost-share-weighted sum of sectoral returns to scale, we find $\sum w_i^c \hat{\gamma}_i = 1.03$. Converting this number to value-

[25] For example, Benhabib and Farmer (1994), Farmer and Guo (1994), and Gali (1994).
[26] Beaudry and Devereux (1994).
[27] We cannot estimate γ^V directly, using data just on real value added and primary inputs of capital and labor, for the reasons discussed in Section II.
[28] E.g. Hall (1988, 1990); Domowitz, Hubbard and Petersen (1988).

added returns, we get $\left(\sum w_i^c \hat{\gamma}_i\right)^V = 1.07$ — too small even to be a major propagation mechanism, let alone enough to produce indeterminacy. Thus, it appears that controlling for aggregation produces a major change in this important parameter.[29]

However, just comparing the estimated returns to scale at these two levels of aggregation does not prove that aggregation effects alone are responsible for the different results. We therefore adjust aggregate output growth as required by our derivation to eliminate aggregation biases. We use the series ξ', which adjusts the standard productivity series for all but the average returns-to-scale effect. We estimate the regression

$$\xi' = const + \gamma^V dx^V,$$

and obtain a direct estimate of $\hat{\gamma}^V$ corrected for composition effects. We find $\hat{\gamma}^V = 1.20$, larger than the 1.07 we predicted by averaging the sectoral figures, but insignificantly different from either 1.07 or 1. The results for manufacturing are similar to those for the aggregate economy, but the changes are less dramatic.

Note that by any of the measures we examine, the estimates of two-digit returns to scale are much smaller than the conventionally-estimated aggregate returns to scale. This is the fundamental stylized fact stressed by Caballero and Lyons (1992), who argue that it is evidence in favor of externalities to production between 2-digit industries that become internalized at higher levels of aggregation. This finding is the basis for their statistical model, which purports to find large spillovers between two-digit manufacturing industries. We have shown that their explanation is not the only one. The greater cyclicality of productivity at the 1-digit level results from aggregation effects rather than positive spillovers.[30] Several recent papers have used external rather than internal

[29] Whatever the right way to calibrate a one-sector model might be, taking literally the assumption of returns to scale as high as 1.98 produces unreasonable results. For example, if returns to scale are that large then true technology growth at the aggregate level is not positive, but rather -1.3 percent per year. (Rotemberg and Woodford [1995] show how one might reconcile short-run increasing returns to scale with long-run constant returns.)

[30] Caballero and Lyons (1992) test their hypothesis directly by regressing sectoral value added on sectoral primary input growth and aggregate output growth. Basu and Fernald (1995a) show that their test is flawed by their use of value-added data, for the reasons discussed in Section II. Using the appropriate gross-output data, the evidence for externalities disappears. However, Bartelsman, Caballero and Lyons (1994) find significant externalities in four-digit gross-output data.

increasing returns to model economic fluctuations.[31] Our results demonstrate that, as written, these models are no more plausible than those based on internal increasing returns.

One might ask whether the finding of constant returns and no externalities makes a large line of recent theoretical work empirically irrelevant. Such a conclusion is warranted by most of the models now extant, but we think it too hasty. We touch on this issue below.

Having noted that our corrections do in fact matter for computing productivity growth, we look more closely at the components of our adjustment term, ξ. Tables 8a and 8b provides summary statistics for the four series that comprise ξ. We see that the two terms with significant standard deviation are the returns to scale term and the correction for the difference between the two concepts of value added.

Comparing equations (18) and (19) we note that these two terms each comprise two effects. One effect comes from the average error in calculating sectoral technology shocks, the other comes directly from composition biases. We can ask which effect is more significant by one of the criteria we used above. Suppose we correct the aggregate productivity series for all but the two "average effects" defined in equation (19). Then, at the level of the private economy, the correlation of this new series with aggregate output would be 0.74. This compares to the correlation of 0.84 for the unadjusted series and 0.59 for the fully-adjusted series. So the aggregation corrections are somewhat less important than the "average" ones at the economy-wide level. But the situation is reversed within manufacturing, where the aggregation effects account for about 60 percent of the correlation with output. Thus both sets of corrections contribute significantly our results.

VII. Conclusion

We have explored the theory, and a little of the practice, of calculating aggregate productivity growth under non-constant returns and imperfect competition in output and factor markets. We come to a number of important conclusions.

[31] See, e.g., Baxter and King (1991) and Benhabib and Farmer (1994).

The first is the crux of our paper: without factor price equalization, perfect competition, and constant returns to scale in all sectors and markets, even the aggregate cost-based Solow residual is not the right measure of an economy's technological change. Moreover, the aggregate residual does not allow us to estimate the extent of the average sectoral departure from perfect competition and constant returns.

Second, we show that the standard definition of productivity — which, as we noted, is not a measure of aggregate technology change — is nevertheless an economically interesting quantity. It measures the change in final output produced from a given set of primary inputs, and thus is a natural measure of welfare. In fact, we present examples of economies with imperfect competition, where productivity as well as welfare change without any change in production technology.

Third, we show how to create a proper measure of aggregate technology from information on sectoral quantities. Unfortunately, simple summary measures of the sectoral quantities do not suffice: we need to know the full distribution of inputs and outputs, as well as the extent of sectoral returns to scale.

Fourth, we begin to apply our method to data. We show that aggregation effects explain much of the cyclicality of aggregate productivity. We obtain an estimate of true technology shocks by controlling for these aggregation effects. Compared with the cost-based Solow residual, true technology shocks are less volatile, have a significantly smaller covariance with output, and have almost no covariance with inputs. These changes matter for a number of issues at the heart of current business-cycle theory, including the plausibility of real-business-cycle models, the extent of increasing returns to scale, and the existence of productive externalities across sectors.

This paper stresses a negative view of composition effects by emphasizing the errors caused by failing to account for aggregation bias. Plausibly, however, composition effects are important economic mechanisms in their own right, and may contribute to the propagation of business cycles. Consider the second example of Section I, where aggregation led to higher returns to scale in the industry than in either firm. In a sense there really are "aggregate increasing returns" : the industry

(or the economy) in fact produced almost 3 percent more output for each percent increase in labor input.

For many macroeconomic purposes, the level of "aggregate returns" — including reallocation effects — may be the relevant parameter. For example, existing models of indeterminacy often require implausibly large firm-level increasing returns. However, a sufficiently large level of aggregate returns may suffice to generate the indeterminacy result. Aggregate increasing returns can, for example, arise from small differences in returns to scale across firms, or even from constant returns but differences in productivity levels. As noted in Section IV, several industry studies find roughly constant returns to scale within productive units but different levels of productivity across units.

We plan to extend our work by applying it to economic models where factors of production are quasi-fixed. Note that our "reallocation term" for labor used steady-state industry wage differentials. But steady-state wage differences are likely to be small relative to cyclical differences induced by changes in demand across sectors. With quasi-fixed primary inputs, payments to labor and capital can differ across sectors even without imperfections in factor markets. And the sectors with the highest wage premia would naturally experience maximum employment growth, reinforcing the productivity effects of inter-industry wage differences. Thus, we hope to show that the mechanisms we identify here can explain substantial procyclicality of productivity even in a purely neoclassical multi-sector model without technology shocks.

We are most excited, however, at the prospect of extending our empirical work. Our measured technology shocks, though less cyclical than the Solow residual, still contain two sources of cyclical bias. First, we aggregated from roughly two-digit industries upwards, assuming that each of these industries had an invariant production functions. Since each of these industries in fact comprises thousands of firms, our "sectors" are themselves subject to potentially huge aggregation effects. Second, although not emphasized in this paper, each productive unit likely experiences substantial variations in capacity utilization that are usually--incorrectly--measured as technology change. However, several authors show how one might correct for unobserved capacity utilization at the

establishment level.[32] The ultimate implication of our work is that if one wants a true technology series, then in principle one should control for unobserved capacity utilization at the establishment level, and aggregate upwards.

Such an undertaking, or as close to it as is feasible, is likely to explain even more of the puzzles of cyclical productivity. These puzzles include the question of why some two-digit industries have apparent decreasing returns to scale. We provided an example where all firms have increasing returns to scale, but the industry as a whole has diminishing returns. This is a case where aggregation effects make productivity appear less procyclical than is in fact the case.

Nevertheless, we believe that on net aggregation effects act to exaggerate, not damp, the cyclicality of productivity. We conjecture that the ultimate explanation of cyclical productivity will consist of at most a modest degree of increasing returns, some cyclical utilization, and a large correction for aggregation. The residual that is left — the true measure of technology shocks — may well be too small and acyclical to play much of a role in explaining business cycles.

[32] For example, Basu (1995b), and Burnside, Eichenbaum, and Rebelo (1995).

References

Aizcorbe, Ana and Kozicki, Sharon (1995). "The Comovement of Output and Labor Productivity in Aggregate Data for Auto Assembly Plants." Finance and Economics Discussion Series 95-33, Federal Reserve Board.

Basu, Susanto (1995a). "Intermediate Goods and Business Cycles: Implications for Productivity and Welfare." *American Economic Review 85* (June) 512-531.

_____ (1995b). "Procyclical Productivity: Increasing Returns or Cyclical Utilization?" *Quarterly Journal of Economics*, forthcoming.

Basu, Susanto and Fernald, John G. (1995a). "Are Apparent Productive Spillovers a Figment of Specification Error?" *Journal of Monetary Economics* (August).

_____ (1995b). "Constant Returns and Small Markups in U.S. Manufacturing." Mimeo, Federal Reserve Board.

Baily, Martin N.; Hulten, Charles and Campbell, David. "Productivity Dynamics in Manufacturing Plants." Brookings Papers on Economic Activity (Microeconomics), 1992, (1), pp. 187-267.

Bartelsman, Eric; Caballero, Ricardo and Lyons, Richard K. (1994). "Customer- and Supplier-Driven Externalities." *American Economic Review 84* (September) 1075-84.

Baxter, Marianne and King, Robert (1991). "Productive Externalities and Business Cycles." Federal Reserve Bank of Minneapolis Discussion Paper #53.

Beaudry, Paul and Devereaux, Michael (1994). "Monopolistic Competition, Price Setting, and the Effects of Real and Nominal Shocks." Mimeo, Boston University.

Benhabib, Jess and Farmer, Roger E. A. (1994). "Indeterminacy and Increasing Returns." *Journal of Economic Theory*, June 1994, 63, 19-41.

Bertin, Amy L.; Bresnahan, Timothy F. and Raff, Daniel M. G. (1995). "Localized Competition and the Aggregation of Plant-Level Increasing Returns: Blast Furnaces 1929-1935." Working Paper 95-11, Reginal H. Jones Center, Wharton School, University of Pennsylvania.

Bhagwati, Jagdish and Ramaswami, V. K. and Srinivasan, T. N. (1969). "Domestic Distortions, Tariffs, and the Theory of Optimum Subsidy: Some Further Results." *Journal of Political Economy*, November/December, 77(6), pp. 1005-1010.

Bulow, Jeremy I. and Summers, Lawrence H. (1986). "A Theory of Dual Labor Markets with Application to Industrial Policy, Discrimination, and Keynesian Unemployment." *Journal of Labor Economics*, July, 4(3), Part 1, pp. 376-414.

Burnside, Craig (1994). "What do Production Function Regressions Tell Us about Increasing Returns to Scale and Externalities?" Mimeo, University of Pittsburgh.

Burnside, Craig; Eichenbaum, Martin and Rebelo, Sergio (1995). "Capital Utilization and Returns to Scale." In Ben S. Bernanke and Julio J. Rotemberg, eds., *NBER Macroeconomics Annual.*

Caballero, Ricardo J. and Hammour, Mohammed (1994). "The Cleansing Effect of Recessions." *American Economic Review 84* (December) 1350-1368.

Caballero, Ricardo J. and Lyons, Richard K. (1992). "External Effects in U.S. Procyclical

Productivity." *Journal of Monetery Economics 29*, 209-226.

Caves, Richard E. and Barton, David R. (1990). *Efficiency in U.S. Manufacturing Industries.* Cambridge, MA: MIT Press.

Domowitz, Ian; Hubbard, R. Glenn and Petersen, Bruce C. (1988). "Market Structure and Cyclical Fluctuations in U.S. Manufacturing." *Review of Economics and Statistics*, February 1988, 70, 55-66.

Farmer, Robert and Guo, Jang-Ting (1994). "Real Business Cycles and the Animal Spirits Hypothesis." *Journal of Economic Theory 63*, 42-72.

Fisher, Franklin M. (1993). *Aggregation.* Cambridge: MIT Press.

Galí, Jordi (1994). "Monopolistic Competition, Business Cycles, and the Composition of Aggregate Demand." *Journal of Economic Theory*, June 1994, 63, 73-96.

Hall, Robert E. (1988). "The Relation Between Price and Marginal Cost in U.S. Industry." *Journal of Political Economy* 96 (Oct.) 921-947.

_____ (1990). "Invariance Properties of Solow's Productivity Residual." In Peter Diamond (ed.) *Growth, Productivity, Employment* (Cambridge: MIT Press).

Hall, Robert E. and Jorgenson, Dale W. (1967). "Tax Policy and Investment Behavior." *American Economic Review* 57 (June) 391-414.

Hulten, Charles (1978). "Growth Accounting with Intermediate Inputs." *Review of Economic Studies 45*, 511-518.

Jorgenson, Dale W.; Gollop, Frank and Fraumeni, Barbara (1987). *Productivity and U.S. Economic Growth.* Cambridge: Harvard University Press.

Jorgenson, Dale W. and Yun, Kun-Young (1991). *Tax Reform and the Cost of Capital.* Oxford: Oxford University Press.

Katz, Lawrence F. and Summers, Lawrence H. (1989) "Industry Rents: Evidence and Implications." *Brookings Papers on Economic Activity (Microeconomics)*, (1), pp. 209-290.

Olley, G. Steven and Ariel Pakes (1992). "The Dynamics of Productivity in the Telecommunications Equipment Industry." NBER Working Paper 3977.

Romer, Christina and Romer, David (1989). "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz." In Olivier J. Blanchard and Stanley Fischer, eds., *NBER Macroeconomics Annual*: 63-129.

Rotemberg, Julio J. and Saloner, Garth (1986). "A Supergame-Theoretic Model of Price Wars During Booms." *American Economic Review 76* (June)390-407.

_____ and Woodford, Michael (1995). "Dynamic General Equilibrium Models with Imperfectly Competitive Product Markets." In Thomas Cooley, ed., *Frontiers of Business Cycle Research.*

Sato, Kazuo (1975). *Production Functions and Aggregation.* Amsterdam: North-Holland.

Schmitt-Gröhé, Stephanie (1994). "Comparing Four Models of Fluctuations Due to Self-Fulfilling Expectations." Mimeo, Federal Reserve Board.

Solow, Robert M. (1957). "Technological Change and the Aggregate Production Function." *Review of Economics and Statistics*, 39(): 312-320.

Table 1. Estimates of Returns to Scale by Industry

$$dy_i = \gamma_i dx_i + dt_i$$

Industry (Approx. SIC)	$\hat{\gamma}$
Agriculture (01-09)	0.51
Metal Mining (10)	1.07
Coal Mining (11-12)	0.76
Oil & Gas Extraction (13)	0.02
Non-metallic mining (14)	0.39
Construction (15-17)	1.06
Food (20)	0.98
Tobacco (21)	0.64
Textiles (22)	1.00
Apparel (23)	0.90
Lumber (24)	0.79
Furniture (25)	1.08
Paper (26)	1.17
Printing & Publishing (27)	0.76
Chemicals (28)	0.30
Petroleum Products (29)	0.25
Rubber & Plastics (30)	1.00
Leather (31)	0.64
Stone, Clay & Glass (32)	0.98
Primary Metal (33)	1.19
Fabricated Metal (34)	1.16

Table 1 (cont'd)

Industry (Approx. SIC)	$\hat{\gamma}$
Non-Elect. Machinery (35)	0.89
Electrical Machinery (36)	1.04
Motor Vehicles (371)	1.20
Other Transport (372-79)	1.00
Instruments (38)	0.97
Miscellaneous Manuf. (39)	0.91
Transportation (40-47)	1.28
Communication (48)	1.06
Electric Utilities (491)	1.65
Gas Utilities (492)	0.93
Wholesale and Retail (50-59)	1.43
Finance, Real Estate (60-67)	1.00
Services (various)	1.22
Weighted Average $\left(\sum w_i^c \hat{\gamma}_i\right)$	**1.03**

Table 2a. Standard and Corrected Measures of Technology Growth for the Private U.S. Economy (1950-1989)

	Output Growth (dv)	Input Growth (dx^v)	Standard Productivity Growth (dp)	Corrected Productivity Growth (dp^*)
Mean	**0.036**	**0.024**	**0.011**	**0.002**
Std. Dev.	**0.031**	**0.017**	**0.022**	**0.019**
Maximum	**0.10**	**0.052**	**0.065**	**0.055**
Minimum	**-0.02**	**-0.01**	**-0.04**	**-0.03**

Table 2b. Correlations of Series in Table 2a.
(Private U.S. Economy, 1950-89)

	Output Growth (dv)	Input Growth (dx')	Standard Productivity Growth (dp)	Corrected Productivity Growth (dp')
Output Growth (dv)	1			
Input Growth (dx')	0.73	1		
Standard Productivity Growth (dp)	0.84	0.25	1	
Corrected Productivity Growth (dp')	0.59	-0.02	0.85	1

Table 3a. Standard and Corrected Measures of Technology Growth for U.S. Manufacturing (1950-1989)

	Output Growth (dv)	Input Growth (dx^*)	Standard Productivity Growth (dp)	Corrected Productivity Growth (dp^*)
Mean	**0.038**	**0.017**	**0.021**	**0.016**
Std. Dev.	**0.055**	**0.036**	**0.034**	**0.028**
Maximum	**0.119**	**0.075**	**0.076**	**0.064**
Minimum	**-0.074**	**-0.059**	**-0.081**	**-0.060**

Table 3b. Correlations of Series in Table 3a
(U.S. Manufacturing, 1950-89)

	Output Growth (dv)	Input Growth (dx^*)	Standard Productivity Growth (dp)	Corrected Productivity Growth (dp^*)
Output Growth (dv)	1			
Input Growth (dx^*)	0.80	1		
Standard Productivity Growth (dp)	0.77	0.24	1	
Corrected Productivity Growth (dp^*)	0.43	-0.01	0.71	1

Table 4. Effects of Technology Shocks on Composition

$$\xi' = c + \beta_1 dp^*_{-1} + \beta_2 dp^*_{-2}$$

Independent Variables	Private Economy	Manufacturing
dp^*_{-1}	**0.13 (0.10)**	**0.06 (0.14)**
dp^*_{-2}	**-0.04 (0.09)**	**0.06 (0.14)**

Sample period 1952-89.
Standard errors in parentheses.

Table 5. Effects of Oil Price Changes on Standard and Adjusted Technology Shocks

$$dp^{(*)} = c + \beta_1 d\ln(P_{Oil}/P_{GDP}) + \beta_2 d\ln(P_{Oil}/P_{GDP})_{-1}$$

	Private Economy		Manufacturing	
Independent Variables	***dp***	***dp****	***dp***	***dp****
$d\ln(P_{Oil}/P_{GDP})$	-0.06 (0.02)	-0.03 (0.02)	-0.11 (0.03)	-0.027 (0.03)
$d\ln(P_{Oil}/P_{GDP})_{-1}$	-0.02 (0.02)	0.00 (0.02)	-0.03 (0.03)	-0.02 (0.03)
F-statistic $\left(F^2_{37} = 3.26\right)$	5.28	0.95	7.22	0.73

Sample period 1950-89.
Standard errors in parentheses.

Table 6. Effects of Romer Dates on Standard and Adjusted Technology Shocks

$$dp^{(*)} = c + \beta_1 Romerdummy_{-1} + \beta_2 Romerdummy_{-2}$$

	Private Economy		Manufacturing	
Independent Variables	***dp***	***dp****	***dp***	***dp****
$Romerdum_{-1}$	-0.024 (0.001)	-0.017 (0.008)	-0.043 (0.014)	-0.019 (0.012)
$Romerdum_{-2}$	0.001 (0.001)	0.000 (0.009)	0.000 (0.015)	0.010 (0.013)
F-statistic $\left(F^2_{37} = 3.26\right)$	4.08	2.17	4.99	1.39

Sample period 1950-89.
Standard errors in parentheses.

Table 7. Returns to Scale Estimates with and without Composition Corrections
(IV, using Hall-Ramey instruments)

	Aggregate Data (no Comp. Corr.) ($\hat{\gamma}$)	Implied Value-Added ($\hat{\gamma}^V$)	Wtd. Avg. of Sectors $\left(\sum w_i^c \hat{\gamma}_i\right)$	Implied Value-Added $\left(\sum w_i^c \hat{\gamma}_i\right)^V$	Aggregate Data with Comp. Corr. ($\hat{\gamma}^{V*}$)
Private Economy	1.27 (0.08)	*1.98*	1.03	*1.07*	*1.20* (0.28)
Manufac-turing	1.13 (0.05)	*1.53*	1.01	*1.03*	*1.13* (0.22)

Sample period 1950-89.
Standard errors in parentheses.

Note: Value-added returns to scale are italicized.

Table 8a. Components of Adjustment to Standard Technology Growth
(Private U.S. Economy, 1950-89)

	ξ	ξ_1	ξ_2	ξ_3	ξ_4
Mean	0.0092	0.0070	0.0021	0.00041	0.00027
Std. Dev.	0.012	0.0076	0.0071	0.0029	0.0014
Maximum	0.029	0.018	0.021	0.0065	0.0031
Minimum	-0.021	-0.013	-0.014	-0.0077	-0.0032

Table 8b. Components of Adjustment to Standard Technology Growth
(U.S. Manufacturing, 1950-89)

	ξ	ξ_1	ξ_2	ξ_3	ξ_4
Mean	0.0044	0.0011	0.0020	0.00086	0.00030
Std. Dev.	0.024	0.013	0.017	0.0031	0.0020
Maximum	0.050	0.030	0.052	0.0092	0.0066
Minimum	-0.061	-0.032	-0.032	-0.0043	-0.0048