Returns to Scale in U.S. Production: Estimates and Implications

Susanto Basu

University of Michigan and National Bureau of Economic Research

John G. Fernald

Board of Governors of the Federal Reserve System

A typical (roughly) two-digit industry in the United States appears to have constant or slightly decreasing returns to scale. Three puzzles emerge, however. First, estimates often rise at higher levels of aggregation. Second, apparent decreasing returns contradicts evidence of only small economic profits. Third, estimates with value added differ substantially from those with gross output. A representative-firm paradigm cannot explain these puzzles, but a simple story of aggregation over heterogeneous units can. Theory and evidence on aggregation invalidate the common use of demand-side instruments. Finally, we discuss implications of heterogeneity for macroeconomic modeling: A one-sector macroeconomic model that ignores heterogeneity may sometimes require firm-level parameters, but at other times the model may require the ''biased'' aggregate parameters.

This is a substantially revised version of a paper previously circulated as ''Constant Returns and Small Markups in U.S. Manufacturing.'' We thank Russ Cooper, Dale Jorgenson, Pete Klenow, Sam Kortum, Greg Mankiw, Stephanie Schmitt-Grohé, John Shea, and an anonymous referee for helpful comments and Barbara Fraumeni for help with the data. We particularly thank Mike Woodford for extensive written comments on an earlier version of this paper. Basu is grateful to the National Science Foundation for financial support and to the Hoover Institution for its hospitality. This paper represents the views of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or other members of its staff.

[*Journal of Political Economy,* 1997, vol. 105, no. 2]
 0022-3808/97/0502-0001$01.50

Why is productivity procyclical? That is, why do measures of labor productivity and total factor productivity rise in booms? The answer to this question sheds light on the relative merits of different models of business cycles. One recent class of explanations emphasizes the potential role of imperfect competition and increasing returns to scale. Measured total factor productivity growth then reflects not just technology shocks but also variations in input use. Hall (1988, 1990), especially, has argued that relaxing the traditional assumptions of perfect competition and constant returns helps explain procyclical productivity.

In addition, recent papers show that increasing returns and imperfect competition can modify and magnify the effects of various shocks in an otherwise standard dynamic general equilibrium model. In response to government demand shocks, for example, models with countercyclical markups can explain a rise in real wages whereas models with increasing returns can explain a rise in measured productivity. Perhaps most strikingly, if increasing returns are large enough, they can lead to multiple equilibria, in which sunspots or purely nominal shocks drive business cycles.[1]

In assessing the merits of these models, one thus needs to know the empirical importance of increasing returns and imperfect competition. In particular, some models offer predictions that differ significantly from those of more standard models only with substantial increasing returns and a high degree of imperfect competition. Indeed, in some cases new results exist only with sufficiently large increasing returns (e.g., Farmer and Guo 1994). Are these models merely intellectual curiosities, or do they provide genuine insight into the macroeconomy?

This paper provides new empirical evidence on the importance of deviations from constant returns and perfect competition. Using data on 34 industries that together constitute the U.S. private business economy, including 21 roughly two-digit manufacturing industries, we conclude that a typical industry has roughly constant returns to scale, implying at most small markups of price over marginal cost. This finding contrasts with those of Hall (1990) and Domowitz, Hubbard, and Petersen (1988), for example, whose estimates suggest large increasing returns. However, we find substantial heterogeneity across sectors, which turns out to have cyclical implications: neither industries nor aggregates behave as though they were single firms.

[1] See, e.g., Rotemberg and Woodford (1992), Beaudry and Devereux (1994), and Farmer and Guo (1994). For a survey of dynamic general equilibrium models with imperfect competition, see Rotemberg and Woodford (1995).

In particular, aggregation potentially explains three puzzles in the data. First, using an extended version of Hall's (1990) procedure, we find that a typical industry appears to have significantly *decreasing* returns to scale. In the absence of large pure profits, decreasing returns at a firm level implies that firms consistently price output below marginal cost, which obviously makes no economic sense. Second, point estimates vary with the level of aggregation, with a typical industry showing apparent decreasing returns but total manufacturing and the total private economy showing apparent increasing returns.[2] Third, value-added estimates differ substantially from gross-output estimates and appear less robust. Value added is not a natural measure of output and can in general be interpreted as such only with perfect competition. With imperfect competition, the use of value added, even at a firm level, suffers from an omitted-variable bias; in addition, value added suffers different aggregation biases than gross output. A priori, we cannot say whether aggregation biases are larger for value added or for gross output, but as an empirical matter it appears that aggregation biases can explain the lack of robustness in empirical estimates when value-added data are used.

Much of the difficulty in providing simple estimates of returns to scale arises from the substantial heterogeneity in the data. Macroeconomic theory provides relatively little guidance for dealing with heterogeneity since most models assume identical firms. How, then, does heterogeneity affect macroeconomic models? Consider three uses of fully specified models. First, they serve as parables: simple but precisely told stories illustrating a particular economic mechanism. Second, they help in understanding the sources of economic fluctuations, particularly by matching key moments in economic data. Third, they are laboratories used to study the effects of particular policy interventions.

The importance of heterogeneity in production differs in each of these uses. For the model qua parable, with little cost, one can probably abstract from heterogeneity if the point is to explore other economic mechanisms. For example, that production takes place at heterogeneous plants is probably irrelevant for making the point that intertemporal substitution in labor supply is an important propagation mechanism in business cycle models.

But the abstraction may be more costly when one seeks to understand business cycles or predict the effects of policy changes. We

[2] Caballero and Lyons (1992) observed the difference in estimates of returns to scale at different levels of aggregation but interpreted this as evidence of productive spillovers across industries. We argue that aggregation bias provides a better explanation.

make this point with a simple, highly stylized model. In one case, despite heterogeneity, the aggregate degree of returns to scale (which exceeds the average firm's returns to scale) is a sufficient statistic for understanding and calibrating the model. In the second case, this summary statistic does not suffice. The example economies differ only in nonproduction institutions. Hence, general results appear unlikely: The importance of heterogeneity depends on particular features of the model.

Section I extends existing firm-level theory estimating returns to scale. We review Hall's simple nonparametric method and relate two concepts of returns to scale used in macroeconomic studies: returns to scale in gross output and value added. Section II makes the common macroeconomic assumption that industries and aggregates can be modeled as firms. We identify three puzzles, which lead to our proposed solution in Section III: Because of aggregation effects, neither industries nor aggregates can be costlessly modeled as individual firms. Section IV controls empirically for aggregation effects when we move from our industries to various aggregates, demonstrating the importance of these effects. In Section V, we reflect on the implications of our results for calibrating macroeconomic models. Section VI presents conclusions.

I. Estimating Firm-Level Returns to Scale

In this section we quickly review and generalize Hall's (1990) method for estimating returns to scale. We then relate estimates of returns to scale from gross-output data to those from value-added data and evaluate the potential biases from using value added as a measure of production.

A. *Methods for Estimating Returns to Scale*

At a firm level, the correct model of production relates gross output, Y, to primary inputs of capital, K, and labor, L, as well as purchased intermediate inputs of materials and energy, M. Letting T denote the state of technology, we write the firm's production function as

$$Y = F(K, L, M, T). \tag{1}$$

We define s_J as the share of costs for input J in total revenue and c_J as the share in total cost. We assume that firms are price takers in factor markets. Cost minimization then 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-output-augmenting productivity growth dt. That is, if dl, dk, and dm are the growth rates of L,

K, and M, then

$$\begin{aligned} dy &= \gamma \cdot [c_L dl + c_K dk + (1 - c_L - c_K)\, dm] + dt \\ &\equiv \gamma \cdot dx + dt. \end{aligned} \tag{2}$$

This equation generalizes equation 5.29 in Hall (1990). Hall's equation, in turn, generalizes the equation defining Solow's residual, allowing for both non–constant returns and the possibility of economic profits.

Cost minimization implies that returns to scale γ equals the ratio of average to marginal cost. Increasing returns can take different forms: for example, no fixed costs but diminishing marginal cost or fixed costs with flat or upward-sloping marginal cost. Once we estimate γ, we can also calculate a corresponding markup of price over marginal cost, μ. An identity links returns to scale and the markup:

$$\gamma = \frac{AC}{MC} = \left(\frac{P}{MC}\right) \cdot \left(\frac{AC}{P}\right) = \mu \cdot (1 - s_\pi). \tag{3}$$

We calculate a required return to capital in order to construct the cost shares in (2) and use this estimate to calculate the average share of economic profits, s_π. An estimate of γ then implies an estimate of μ. We estimate an average pure profit rate of at most 3 percent.[3] Hence, given relatively small profits, equation (3) shows that μ approximately equals γ; large markups, for example, require large increasing returns.

Given low estimated profits, equation (3) also shows that strongly diminishing returns imply that firms consistently price output below marginal cost. Since this implication makes no economic sense, we conclude that firm-level returns to scale must be either constant or increasing.

The natural model of production at the firm level uses gross production as the concept of output and explicitly models the use of materials input, as in equation (2). However, Hall and many subsequent authors instead relate value added to primary inputs of capital and labor alone. We discuss the relative merits of this choice later. For now we simply note the Divisia definition of real value added:[4]

[3] Rotemberg and Woodford (1995) also provide a variety of evidence suggesting that rates of pure economic profit are close to zero.

[4] Beginning in 1995, the U.S. National Income and Product Accounts measure real gross domestic product and sectoral real value added as chain-linked Fisher indices. Chain-linked Fisher indices are one discrete-time approximation of the continuous-time Divisia definition used here.

$$dv \equiv \frac{dy - s_M dm}{1 - s_M} = dy - \left(\frac{s_M}{1 - s_M}\right)(dm - dy). \tag{4}$$

Value added is like a partial Solow residual, subtracting materials growth from output growth, weighted by the share of intermediate inputs in revenue. The second equality shows that if the materials-to-output ratio is constant, then value added grows at the same rate as gross output and intermediates.

The cost-weighted measure of primary input growth, dx^V, is defined analogously to dx, so that

$$dx^V = \left(\frac{c_K}{c_K + c_L}\right)dk + \left(\frac{c_L}{c_K + c_L}\right)dl.$$

Thus the equation estimated by Hall (1990) and others is

$$dv = \rho \cdot dx^V + v. \tag{5}$$

As in equation (2), the error term is interpreted as a shock to production technology. We discuss the relationship between ρ and γ below.

B. Value Added as a Measure of Production

Why do macroeconomists often use value-added data? A compelling reason is that macroeconomists are typically interested in understanding value-added aggregates, especially GDP. Summed across firms or industries, real value added has the desirable property of equaling total national expenditure. Thus aggregate value added is clearly appropriate for focusing on the *uses* of output. But as we now discuss, value added is not generally appropriate for studying productivity growth—that is, attempting to understand the *sources* of output change—since it is not in general a valid production measure.

Gross-output returns to scale γ and markup μ are the primitives of technology and behavior. Thus they are conceptually the natural parameters for, say, calibrating multisector models with imperfect competition. On the other hand, as we now demonstrate, value-added data generally yield biased estimates of returns to scale in the presence of imperfect competition. To show this, rewrite equation (2) as

$$dy = \gamma(1 - c_M)\,dx^V + \gamma c_M + dt. \tag{6}$$

Then, from definition (4), the definition of dx^V, and the first-order condition for cost minimization that $\gamma c_M = \mu s_M$, we find

$$dv = \left[\frac{\gamma \cdot (1 - c_M)}{1 - \gamma \cdot c_M}\right] \cdot dx^V + (\mu - 1)\left[\frac{s_M}{(1 - \mu s_M)(1 - s_M)}\right] \cdot (dm - dy) + \frac{dt}{1 - \gamma c_M}. \quad (7)$$

Thus equation (5) is generally misspecified: There is an omitted variable in the estimating equation that uses value added as the output measure. This variable is identically zero in two cases: if there is perfect competition, so that price equals marginal cost; or if the elasticity of substitution between materials and other inputs is zero, so that $dm = dy$. Since $\mu \geq 1$, the coefficient multiplying $dm - dy$ is weakly positive. So the sign of the omitted-variable bias depends on the sign of the covariance between the projection of dx^V on the instruments and $dm - dy$. If materials intensity is procyclical, the bias is positive.

Suppose for the moment that the bias is zero. Then equation (7) shows that ρ in equation (5) estimates $\gamma(1 - c_M)/(1 - \gamma c_M)$. Making further assumptions about technology provides some economic intuition for this quantity. Suppose that the production function (1) takes the following separable form:

$$Y = G(V^P(K, L, T), H(M)). \quad (8)$$

Following the logic used to derive equation (2), we can write the growth of productive value added 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):

$$dv^P = \gamma^V dx^V + dt^V, \quad (9)$$

where γ^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 $1 - \gamma c_M$ and γc_M. Hence, the relationship between γ^V and γ is

$$\gamma^V = \gamma \cdot \left(\frac{1 - c_M}{1 - \gamma c_M}\right). \quad (10)$$

Thus ρ corresponds to γ^V, the parameter often of interest to macroeconomists. For example, if G is Leontief in V^P and H, then in a representative-firm model γ^V equals returns to scale of the econo-

my's aggregate production function for GDP (Rotemberg and Woodford 1995).[5] If the materials-to-output ratio is acyclical (and hence orthogonal to dx^V), then using value-added data provides an unbiased estimate of γ^V. Of course, without any assumptions about the materials-to-output ratio, one can estimate a gross output γ and then calculate the implied γ^V from equation (10).

C. How Large Is "Large"?

To provide context for our results, we briefly discuss how large deviations from constant returns and perfect competition need to be to generate new results. We focus our discussion on γ^V and its corresponding value-added markup.[6]

Countercyclical markups generally magnify the effects of imperfect competition and increasing returns. In a model with implicit collusion and countercyclical markups, for example, Rotemberg and Woodford (1992) find that a steady-state markup of 1.2 suffices to establish a qualitatively new result: Real wages rise when government purchases rise, since falling markups shift out the labor demand schedule. Thus relatively small deviations from constant returns and perfect competition can significantly change model predictions.

For many recent models of indeterminacy, however, striking results exist only with sufficiently large markups and returns to scale. Schmitt-Grohé (1995) explores the calibration of four prominent models of indeterminacy and finds the minimum values of γ^V that transform the steady state of the neoclassical growth model into a sink rather than a saddle point, which allows "animal spirits" to matter in a rational expectations equilibrium. Models with a constant markup, such as that of Farmer and Guo (1994), typically require markups in excess of 1.75. Even a model with countercyclical markups requires markups in excess of 1.4. (Schmitt-Grohé emphasizes, however, that these minimums unrealistically assume an infinite labor supply elasticity and a labor cost share of 0.7.)

This question of "how large is large" can be answered only in a specific model. Hence, while existing models provide a benchmark for interpreting our empirical results, our results also provide an

[5] Basu (1995*b*) explores some of the implications of dropping the assumption that G is Leontief. Note that the data do not support the assumption of a zero elasticity of substitution. Bruno (1984) reviews a number of studies and concludes that the elasticity is between 0.3 and 0.4. Rotemberg and Woodford (1993, app. 3) conclude that a reasonable value for this elasticity is 0.7.

[6] Most existing models assume the existence of a representative producer whose production function satisfies the conditions in eq. (8), with zero elasticity of substitution between V and M. Since profits are small, calibrations in terms of returns to scale or markups are roughly equivalent, so we use the terms interchangeably.

input into the development of ''reasonable'' models. Some recent models, for example, incorporate features to get indeterminacy in ways that do not require sizable increasing returns. For example, multisector models can generate indeterminacy with relatively small increasing returns, essentially because each sector faces downward-sloping supply curves for capital and labor even if aggregate factor supplies are increasing functions of factor prices and are quite inelastic (Benhabib and Farmer 1995; Perli 1995).

II. Data and Puzzling Results

We now apply the theory from Section I to estimate returns to scale in U.S. production at various levels of aggregation. We follow the standard macroeconomic practice of applying firm-level theory to relatively aggregated data; within manufacturing, for example, we use data on industries defined at approximately the two-digit level of Standard Industrial Classification (SIC) codes. We use aggregates, rather than firms, since comprehensive firm-level data exist for only narrow sectors of the U.S. economy. Even in manufacturing, where relatively detailed data are available, time-series data on firms are incomplete, are of short duration, and ignore entry and exit.

But applying firm-level theory to aggregate data produces three puzzling results: Returns-to-scale estimates differ at different levels of aggregation, estimated returns to scale are sometimes strongly diminishing, and results are too sensitive to the use of gross output versus value added. We document these puzzles here and attempt to resolve them in the following sections of the paper.

A. Data

We use data provided by Dale Jorgenson for 34 industries that together constitute the U.S. private business economy for 1959–89. These data seek to provide complete sectoral production data and represent a massive effort to ensure consistency with production theory. Hence, the data set contains observations on primary and intermediate inputs as well as gross output. Jorgenson, Gollop, and Fraumeni (1987) document the data thoroughly.

Appendix A describes in greater detail how we use the data. In brief, we construct several variables from the data. First, for each industry we calculate output and input growth rates as log changes. Second, we calculate payments to capital by estimating capital's user cost, as in Hall (1990). Third, we calculate Divisia aggregates of output and input growth at the level of nondurable, durable, and total manufacturing, as well as the total private business economy. Fourth,

we construct value-added data for each industry and aggregate using definition (4).[7]

We present regressions both instrumented and uninstrumented. Hall (1990) argues for demand-side instruments; he uses changes in the world price of oil and government defense spending, and a dummy variable indicating the political party of the president. For comparability, we use these instruments along with one lag of each.

In later sections, however, we emphasize the uninstrumented results, for two reasons. First, the instruments may not be completely exogenous, uncorrelated with the disturbance term. For example, for value added, equation (7) shows that anything that changes the intensity of intermediate input use (as oil prices may) is not a valid instrument. For gross output as well, if technological change is energy biased, then lower oil prices are associated with faster technology growth. More important, we argue in Section IV that aggregation effects appear important in the data and that these effects are correlated with aggregate demand shocks. Second, our instruments are relatively weakly correlated with inputs for some industries. Even if the instruments are not so weak that they lead to small-sample problems (such as those pointed out by Nelson and Startz [1990] and Staiger and Stock [1994], among others), instruments that are both relatively weak and potentially correlated with the disturbance term suggest that instrumental variables may be more biased than ordinary least squares (OLS).[8]

B. Results

Tables 1 and 2 report our estimates from equations (2) and (5). Table 1 reports single-equation estimates using data at various levels of aggregation; table 2 reports weighted averages of the results of similar regressions for the industries included in those aggregates. In both tables, the first row reports estimates of γ from gross-output data. The second row uses equation (10) to convert these gross-output estimates to estimates of γ^V. The third row reports estimates of ρ (which may or may not be good estimates of γ^V) using our constructed measures of value added. Panel A shows two-stage least-squares results and panel B shows OLS results.

[7] A previous version of this paper also used Hall's (1988, 1990) data set. The qualitative features of our results do not change much with that data set; the puzzles we document remain.

[8] The first-stage *F*-statistic averages about three. Whether or not we instrument does not affect our qualitative results and usually has only a small effect on our quantitative results. We cannot tell whether this invariance reflects small OLS bias or weak or invalid instruments.

TABLE 1

AGGREGATE ESTIMATES

Parameter	Private Economy (1)	Manufacturing (2)	Manufacturing Durables (3)	Manufacturing Nondurables (4)
	A. Two-Stage Least Squares			
Gross output γ	1.27	1.09	1.14	.86
	(.10)	(.07)	(.05)	(.15)
Implied value added γ^V	1.72	1.29	1.46	.66
	(.36)	(.29)	(.21)	(.28)
Direct value-added estimate	1.46	1.10	1.40	.04
	(.38)	(.33)	(.27)	(.63)
	B. Ordinary Least Squares			
Gross output γ	1.23	1.12	1.13	.99
	(.06)	(.04)	(.03)	(.10)
Implied value added γ^V	1.57	1.41	1.40	.97
	(.20)	(.20)	(.12)	(.31)
Direct value-added estimate	1.29	1.21	1.32	.52
	(.23)	(.19)	(.14)	(.44)

NOTE.—Sample period is 1959–89. In both panels, the first row presents single-equation estimates of eq. (2), which are converted in the second row using eq. (10). The third row estimates eq. (5). Instruments are the price of oil, government defense spending, and the political party of the president, with one lag of each.

TABLE 2

WEIGHTED AVERAGE OF SECTORAL ESTIMATES

Parameter	Private Economy (1)	Manufacturing (2)	Manufacturing Durables (3)	Manufacturing Nondurables (4)
	A. Two-Stage Least Squares			
Gross output γ	.97	.92	1.08	.73
	(.12)	(.05)	(.03)	(.11)
Implied value added γ^V	1.16	1.06	1.32	.67
	(.23)	(.08)	(.12)	(.12)
Direct value-added estimate	.94	.87	1.26	.26
	(.22)	(.15)	(.16)	(.29)
	B. Ordinary Least Squares			
Gross output γ	.83	.93	1.07	.77
	(.04)	(.03)	(.02)	(.05)
Implied value added γ^V	.89	1.03	1.28	.66
	(.07)	(.05)	(.06)	(.06)
Direct value-added estimate	.54	.66	1.07	.06
	(.09)	(.08)	(.07)	(.15)

NOTE.—Sample period is 1959–89. In both panels, the first row presents gross-output weighted averages of single-equation industry estimates of eq. (2). The second row converts each industry estimate using eq. (10), and then averages with value-added weights. The third row presents value-added estimates of eq. (5). Instruments are the price of oil, government defense spending, and the political party of the president, with one lag of each.

In table 2, for example, column 1 is the weighted average of results for all 34 individual industries. The weights are the shares of each industry in the aggregate. In the first row, the weights are in total nominal gross output; in the second and third rows, the weights are in total nominal value added. Note that the second row of table 2 is not a simple transformation of the estimate in row 1 since γ^V is calculated for each industry before averaging.[9]

We first discuss the instrumented results in table 1. The first row shows that the production of gross output in the total private economy as well as in nondurables manufacturing shows statistically significant evidence of aggregate increasing returns. Within manufacturing, durables shows the strongest evidence of increasing returns. As shown in row 2, the magnitude of increasing returns is sizable: For the entire private economy, the estimate of γ^V is 1.72; taken at face value, this estimate is large enough to justify some existing models of multiple equilibria. Even the estimate for durables manufacturing, 1.46, is capable of justifying multiple equilibria in some models.

Given questions about the instruments, panel B shows the regressions uninstrumented. The point estimates are, for the most part, little changed; the estimates are (surprisingly) lower uninstrumented for the entire economy as well as for durables. Point estimates generally rise for nondurables. The major difference is the change in standard errors. For the uninstrumented regressions, we can now reject constant returns at the 1 percent level for total manufacturing gross output.

Comparing tables 1 and 2 documents our first puzzle: Estimated returns to scale are strikingly larger in aggregate than in industry-level data. In table 2, only durable-goods manufacturing industries show any evidence of increasing returns to scale. Even there, the point estimate for γ^V is only 1.28; if this estimate applied to the entire economy, it would be large enough to give some interesting macroeconomic results but would not generate multiple equilibria in most models. For overall manufacturing, the typical industry has diminishing returns to gross output, whereas in value added (row 2), the typical industry has roughly constant returns. For the entire econ-

[9] We convert each estimate of γ to γ^V using the sample average c_M for the corresponding industry or aggregate. Standard errors are calculated by linearizing eq. (10) and using the standard variance formulas for linear equations. For the weighted averages, standard errors are calculated assuming the estimates are independent and using the standard formula for the variance of a sum of random variables. Estimating the system of equations for manufacturing using seemingly unrelated regressions (allowing for different parameters for each industry) would allow for nonzero covariances in the estimates; this makes only a small difference to the standard errors of the weighted average, however, on the order of 0.01.

omy, returns to scale appear strongly diminishing, particularly when value-added data are used. In nondurables, returns to scale are strikingly low by any measure.

This finding constitutes our second puzzle: Returns to scale estimates are often much smaller than one. Returns to scale are larger in the instrumented regressions (opposite what we expect if the problem is positive feedback from technology shocks to input use), but even there, where the average value of γ^V is 1.16 for the entire private economy, the median estimate (not shown) is 1.01. Hence, half the estimates are less than one. (Statistically, five individual estimates are significantly less than one at the 5 percent level; three are significantly greater than one.) As noted in our discussion following equation (3), strongly diminishing returns imply sizable economic profits, which are not observed in U.S. data. Furthermore, the possibility of replication suggests that in the long run, marginal cost should not much exceed average cost. We thus believe a priori that returns to scale should be no lower than constant, but the data seem to contradict us.

Our third puzzle comes from comparing our gross-output and value-added results. In table 1, the point estimates using value-added data are consistently smaller than the implied γ^V from the gross-output estimates, and standard errors are large enough that we can never reject constant returns in the instrumented regressions. The most striking estimate pertains to manufacturing nondurables, where the point estimate suggests returns to scale of about zero. (The standard error is so large, however, that the regression is largely uninformative.) Table 2 shows a similar pattern.

Equation (7) showed that value-added estimates suffer a potential omitted-variable bias: Can this bias explain the differences between gross-output and value-added estimates of γ^V? The bias should depend on the coefficient from regressing the omitted variable, $dm - dy$, onto dx^V, suitably instrumented. For 22 of the 34 industries, the coefficient is positive, averaging about 0.1. For all our aggregates as well, the coefficient is positive. Since a priori the markup is greater than or equal to one, equation (7) implies that value-added estimates should be biased upward.

In tables 1 and 2, however, direct value-added estimates tend to be smaller than the indirect gross-output estimates. Thus the firm-level theory that predicts differences between these two estimates does not explain the differences we actually find.[10]

[10] With heterogeneity, Jensen's inequality also implies that row 2 should yield smaller estimates than row 3 since eq. (10) is convex. Thus heterogeneity heightens the puzzle of why gross-output and value-added estimates of γ^V differ.

The results in tables 1 and 2 thus leave us with three puzzles. First, the aggregate results differ significantly from the sectoral results, tending to show increasing returns. Second, the weighted average of sectoral estimates in table 2 often shows a statistically significant degree of diminishing returns to scale. Third, results are too sensitive to whether they are estimated from gross output or value added. The rest of this paper attempts to resolve these puzzles.

C. Comparison with Existing Literature

Consistent with our findings, previous literature reports a wide range of returns-to-scale estimates, depending on type of data, level of aggregation, and estimating method. First, several widely cited papers find sizable industry increasing returns or markups. Hall (1990), for example, reports large increasing returns using data for two-digit manufacturing value added, whereas in table 2 we report decreasing returns with similar data. This difference arises primarily because Hall estimates equation (6) in reverse, regressing input growth on output growth and then inverting the resulting coefficient. In our data, the reverse regressions almost always show strong and statistically significant increasing returns as well.

The reverse regressions are not trustworthy, however. The OLS bias is large since output (the right-hand-side regressor) necessarily covaries with disturbances to output. Instrumental variables do not necessarily eliminate the resulting bias, for two reasons. First, the literature on the small-sample properties of instrumental variables suggests that if OLS bias is large, the instruments must be relatively strongly correlated with the endogenous explanatory variable; otherwise, instrumental variables estimates are biased in the direction of the OLS bias. Bartelsman (1995) presents Monte Carlo evidence suggesting that the resulting small-sample bias of the reverse regressions is severe. Second, if the instruments are not only weak but also bad (correlated with the error term), then the resulting bias is likely to be larger in the reverse regression. This bias depends on the covariance between the disturbance term and the fitted value of the endogenous explanatory variable; with bad instruments, the fitted value for output necessarily covaries with the disturbance term. In Section IV, we argue that the instruments are bad. Putting output on the right-hand side of the regression thus puts very strong and probably unwarranted faith in the explanatory power and exogeneity of the instruments; the forward regression follows more naturally from theory and suffers fewer econometric problems.

Domowitz et al. (1988) use four-digit gross output data and report

gross-output markups of around 1.6. We find their results implausible, largely because their data are incomplete in ways that make their productivity residuals spuriously cyclical. Perhaps most important, their labor compensation data are incomplete, omitting payments for social security, health insurance, and pensions, for example. As a result, labor's share of value added is only about one-third (see Norrbin 1993); since labor hours are much more cyclical than the capital stock, measured productivity becomes more cyclical than true productivity.

Second, some other recent work at the industry level also finds the puzzle of decreasing returns. For example, Burnside (1995) explores the robustness of our industry results across data sets, instrument lists, and sample periods. Using Hall's manufacturing value-added data for 1953–84, he reports a weighted-average returns to scale of about 0.9. Burnside, Eichenbaum, and Rebelo (1995) use several data sets; in quarterly three-digit manufacturing data, for example, they report estimates of γ^V between 0.8 and 0.9. (In some, though not all, specifications, the results are statistically less than one. See their table 5.)

Third, several authors argue that procyclical productivity and apparent increasing returns result from cyclical variations in the intensity of input use, resulting, for example, from labor hoarding (see, e.g., Shapiro 1993; Bils and Cho 1994; Burnside et al. 1995; Basu 1996). In this paper we take no account of cyclical variations in capacity utilization, even though we do not doubt that such variations exist. Such variations cannot, however, explain our three puzzles. First, they explain apparent increasing returns, not the apparent decreasing returns found in the industry data. Second, they cannot explain the difference between industry and aggregate estimates since capacity utilization explains apparent aggregate increasing returns by explaining apparent industry (and firm) increasing returns. Third, variations in intensity of capital and labor use affect both value-added and gross-output regressions in the same direction and cannot account for different results with different data. (In practice, attempts to control for capacity utilization usually have surprisingly little effect on returns-to-scale estimates. For example, Burnside [1995] tries to use energy use to control for variations in capital utilization; his estimate of the weighted-average gross-output returns to scale rises from 0.87 to 0.91.)

Fourth, most plant and engineering studies find essentially constant returns to scale. For example, Baily, Hulten, and Campbell (1992) use plant-level data and find about constant returns, with more estimates slightly below one than above. Griliches and

Ringstad (1971) argue that essentially constant returns are needed to rationalize the observed large dispersion of establishment sizes within a given industry.

Finally, Caballero and Lyons (1992) first noted the puzzle that returns to scale rise at higher levels of aggregation. They interpreted this as evidence of enormous productive spillovers across industries. However, we do not find their explanation compelling. First, it is hard to identify specific examples of such spillovers affecting manufacturing plants. Second, Basu and Fernald (1995*b*) find that the Caballero-Lyons specification, which involves including aggregate inputs in industry regressions, shows apparent externalities in value-added data but not in gross-output data. If these are true productive spillovers, the effect should be present in both sources of data. Thus we confirm their stylized fact in our data but are not convinced by their explanation.

III. Aggregating over Firms

Macroeconomics is microeconomics plus aggregation (Fisher 1993). The previous two sections focused on microeconomics, asking how far one can take the representative-firm paradigm. We concluded that the data seem inconsistent with firm-level theory. This section focuses on aggregation, asking when it matters that each of our industries comprises thousands of firms.

We aggregate gross output and value added as Divisia indices. Thus aggregate gross-output growth dy and value-added growth dv are

$$dy = \sum_i w_i \cdot dy_i,$$
$$dv = \sum_i wv_i \cdot dv_i, \tag{11}$$

where w_i is the share of the firm's revenue in total industry revenue and wv_i is the share of the firm's nominal value added (revenue minus intermediate-input costs) in total industry value added.

We assume that there are no economic profits (so total cost equals total revenue) and that factor markets are perfectly competitive.[11] These assumptions imply that the growth of industry inputs, dx,

[11] Aggregation can also fail because of factor rents or profit rates that differ across sectors. These effects are difficult to estimate, and existing empirical evidence on their importance is controversial. For simplicity, we thus abstract from these failures of aggregation here. However, we do consider factor rents theoretically in Sec. V. See Basu and Fernald (1995*a*) for a complete derivation.

equals the weighted sum of firm-level inputs dx_i:

$$dx = \sum_i w_i \cdot dx_i. \tag{12}$$

Substituting into equation (11) for the growth rates of dy_i, we find

$$dy = \sum_i w_i \cdot \gamma_i \cdot dx_i + \sum_i w_i dt_i. \tag{13}$$

Defining $\bar{\gamma}$ as the weighted-average returns to scale in the industry, $\sum w_i\gamma_i$, we can rewrite this as

$$dy = \bar{\gamma}dx + R + dt, \tag{14}$$

where dt equals the weighted average of firm technology shocks, and

$$R = \sum_i w_i \cdot (\gamma_i - \bar{\gamma}) \cdot dx_i. \tag{15}$$

Aggregate gross output growth in (14) depends on technology shocks, dt, plus returns to scale in the ''typical'' firm, $\bar{\gamma}$, multiplied by aggregate input growth. In addition, output growth depends on the reallocation R, which contributes positively to growth if firms with high returns to scale have higher than average growth in inputs dx_i. Define the sectoral ''cyclicality'' parameter β_i as the regression coefficient of dx on dx_i. The bias caused by the reallocation term (in the OLS regression) is then $\sum w_i(\gamma_i - \bar{\gamma})\beta_i$. Hence, the bias is positive if γ_i is positively correlated with β_i.

For value added, substitute into equation (11) from equation (8). With some rearranging, we can write aggregate value-added growth in a similar way, reflecting the direct contribution of inputs multiplied by ''average'' value-added returns to scale $\bar{\gamma}^V$, technology, and aggregation terms:

$$dv = \bar{\gamma}^V \cdot dx^V + R^V + I + dt^V, \tag{16}$$

where dt^V is the value-added weighted average of technology shocks dt_i^V, and

$$\begin{aligned} R^V &= \sum_i wv_i \cdot (\gamma_i^V - \bar{\gamma}^V) \cdot dx_i^V, \\ I &= \sum_i wv_i \cdot (\gamma_i^V - 1) \cdot \left(\frac{s_{Mi}}{1 - s_{Mi}}\right) \cdot (dm_i - dy_i). \end{aligned} \tag{17}$$

As with gross output, the value-added reallocation effect R^V reflects the ''covariance'' between returns to scale and the cyclicality of input growth. The term I reflects variations in the intensity of intermediate-input use within firms that are imperfect competitors.

The aggregation equations (14) and (16) have several implications. First, aggregate demand instruments need not be uncorrelated with reallocation effects and hence may be invalid. The problem is that demand shocks need not lead to equiproportionate changes in input use for all firms within the aggregate. For example, government defense spending may fall unequally on firms within the industry; some firms may be more sensitive than others to changes in oil prices.

Second, aggregation can plausibly explain the puzzles discussed in Section II. If heterogeneity is cyclical, then estimates at different levels of aggregation are likely to differ. Aggregation effects can in principle cause estimates to either rise or fall at different levels of aggregation. Hence, aggregation can explain the first puzzle, of higher estimates in aggregates than in industries, as well as the second puzzle, of apparent diminishing returns in some industries. Finally, aggregation effects are likely to differ in gross output and value added, potentially explaining why the two types of data yield different results.[12]

Third, the parameters we estimate from aggregate data are not structural: they are complex combinations of structural parameters and behavioral responses to reallocation-inducing shocks. There need not be a stable relationship between inputs and outputs over different time periods or in response to different economic shocks. Hence, estimates may be sensitive to data and sample period.

Fourth, it now becomes unclear what parameter one wants for calibrating returns to scale. Authors calibrating one-sector models tend to focus on estimates of the average value-added returns to scale, $\bar{\gamma}^V$; if estimates differ at different levels of aggregation, this strategy may not be appropriate. We return to this point in Section V.

We conclude by noting that it might seem strange that aggregation fails even under conditions in which firm-level ''productive'' value added V^P can be written as functions of two physically homogeneous factors of production, capital and labor, and in which these factors can move costlessly between firms. Under these assumptions, the theorems of Fisher (1993) would seem to assure the existence of an aggregate production function. Fisher's theorems do not apply in our setup, however, since factors are not necessarily allocated efficiently to maximize output. Since different firms may have different degrees of market power (corresponding to differences in de-

[12] We can generate examples in which gross-output data give better estimates of firm-level parameters than value-added data, and vice versa. Thus there is no general proposition on this issue; it becomes an empirical matter which is preferred.

grees of returns to scale), the same factor has a different value of marginal product in different uses. Thus reallocating a factor from one firm to another can change aggregate output.

IV. Aggregation Results

This section explores the empirical importance of the aggregation effects identified in the previous section. At a minimum, the results suggest that these effects matter. Moreover, these results plausibly yield more accurate estimates of firm-level parameters than those in table 2.

We implement the decomposition in equations (14) and (16), using our sectoral estimates of returns to scale to calculate estimates of R, R^V, and I. We estimate ''aggregation-corrected'' gross-output and value-added growth by subtracting these aggregation terms from actual growth in aggregate gross output and value added. We then regress the result on aggregate inputs.

If aggregation effects operate only across industries but not within industries, then aggregation-adjusted estimates should provide unbiased estimates of the true average parameters. In this case, the estimates reported in this section should be close to those reported in table 2 for the industry averages, differing only because of statistical estimation error. On the other hand, suppose that industries themselves contain substantial unobserved aggregation effects. We argue below that these effects may largely cancel out in the aggregate regressions. In this case, we may obtain better estimates of firm-level parameters from (corrected) aggregate data than from partially disaggregated data.

We return to the issue of interpretation after presenting results. We present only uninstrumented results because in general the presence of aggregation effects implies that demand-side instruments are not valid. To check the empirical importance of this point, we regressed the estimated aggregation terms on each of our instruments.[13] (Since the null hypothesis is that the instruments are valid, we focus here on estimates of the gross-output and value-added terms using the instrumented industry estimates underlying table 2. As reported in App. B, the uninstrumented industry estimates give similar results.) For the change in the world price of oil, we can almost always reject the null. For the private economy as a whole, for example, the coefficient on the current oil price is significantly negative at the .001 level for both value added and gross output; the coefficient on the lagged change is significantly negative at the .01

[13] Appendix B presents these regression results in detail.

TABLE 3

AGGREGATE ESTIMATES CORRECTED FOR REALLOCATIONS

	Private Economy (1)	Manufacturing (2)	Manufacturing Durables (3)	Manufacturing Nondurables (4)
Gross output γ	1.01	1.08	1.11	.96
	(.05)	(.04)	(.03)	(.08)
Implied value added γ^V	1.02	1.26	1.33	.87
	(.11)	(.16)	(.11)	(.23)
Direct value-added estimate	1.03	1.19	1.36	.81
	(.18)	(.18)	(.15)	(.27)

NOTE.—Sample period is 1959–89. Estimated aggregation terms are subtracted from gross-output growth and value-added growth before OLS regressions on input growth.

level for value added and the .05 level for gross output. Government defense spending positively affects reallocations for manufacturing durables but is otherwise insignificant, as is the political party of the president. These results thus verify the empirical importance of the theoretical point that the instruments we (and many others) use are not valid as instruments at an aggregate level. We have no reason to expect that the instruments are any more valid at an industry level.

In table 3, we present aggregation-corrected aggregate results, using the uninstrumented industry estimates. We begin with two general observations. First, overall the estimates for the total private economy in column 1 suggest constant returns. Second, in all cases the implied values of γ^V in row 2 are very close to the direct value-added estimates in row 3. This was not true in the results reported in Section II and suggests that reallocation effects are an important explanation for the differences in results between gross output and value added.

Although the typical industry appears to have constant returns, manufacturing shows some evidence of increasing returns. This is particularly true in durables, where all the estimates show statistically significant increasing returns. In nondurables, all the estimates show slight decreasing returns, though they are not statistically different from one.

It is worth emphasizing that nowhere in table 3 do we have the puzzle of statistically significant decreasing returns. Only in nondurables is there evidence of decreasing returns, and even there the results are not statistically different from one. We cannot say that the results are statistically different from those shown in table 1 since the standard error in the uncorrected regression is very high; but it is worth noting that the standard error falls dramatically since the fit

of the regression improves substantially. These results are, however, statistically significantly different from those in table 2.

Overall, aggregation effects appear to be procyclical, leading to an economically substantial increase in estimates of returns to scale. Using value-added data is mainly a problem with nondurables; those results change substantially in an economically sensible way when we control for aggregation. For durables, corrections to value added make less of a difference. This suggests that the value-added reallocation terms are particularly important for nondurables and relatively unimportant for durables.

The results in table 3 show that correcting for reallocation in aggregate data makes a substantial difference to the estimates but does not recover industry-level averages. Can we nevertheless interpret the results in table 3 as reliable estimates of the average firm-level parameters in the economy? Without detailed firm-level data we cannot be sure but can suggest the factors that matter. Consider aggregating in two stages. First, in equation (14) for gross output, aggregate from the level of firms to the level of individual industries i. Then aggregate over these industries to derive gross output dy at the level of the entire economy. Omitting the technology term gives

$$dy = \bar{\gamma} dx + \sum_i w_i R_i + R. \tag{18}$$

The R_i are the unobserved within-industry reallocation terms, and R reflects the across-industry reallocations that we estimated in this section. The R_i are plausibly correlated with both sectoral and aggregate input growth. For example, idiosyncratic shocks to dx_i may reflect procyclical entry by low-productivity firms and hence lower estimates of returns to scale; systematic shocks to dx may change factor prices and lead to procyclical exit by low-productivity firms and hence raise estimates. In addition, aggregate shocks likely reflect aggregate income changes, leading producers of more durable goods (even within a two-digit category) to increase their inputs relatively more. Thus industry-level reallocations associated with idiosyncratic and systematic shocks need not have the same sign. Suppose that the R_i are related in a (relatively) structural way to dx_i and dx:[14]

$$R_i = \delta_i dx_i + \eta_i dx. \tag{19}$$

Let σ_i equal $\mathrm{cov}(dx_i, dx)/\mathrm{var}(dx_i)$, the regression coefficient of dx

[14] In practice, the R_i need not be structural. For the heuristic argument that follows, this point is unimportant. For the same heuristic reason, we omit disturbance terms, which would add statistical error to the equations that follow without changing any of the substance.

on dx_i. Then the average estimate over industries, $\hat{\bar{\gamma}}$, equals $\bar{\gamma} + \bar{\delta} + \overline{\sigma\eta}$. If sectoral input growth is a poor proxy for aggregate input growth, so that the σ_i are small, then the bias in the industry estimates in table 2 primarily reflects $\bar{\delta}$. (The average $\bar{\sigma}$ is necessarily less than one; in our data it is about 0.3.) By contrast, even if the σ_i are small, the aggregate regressions in general reflect the full effect of $\bar{\eta}$ as well as $\bar{\delta}$.

Substituting $\hat{\gamma}_i$ and $\hat{\bar{\gamma}}$ into equation (15) for R gives $\hat{R}$. With some rearranging of equation (14), aggregation-corrected growth in aggregate output equals

$$dy - \hat{R} = (\bar{\gamma} + \bar{\delta} + \bar{\eta})\, dx - \sum_i w_i(\sigma_i\eta_i - \overline{\sigma\eta}) \cdot (dx_i - dx). \quad (20)$$

When we estimate equation (20), various results are possible, depending on the signs and magnitudes of the parameters. As one case, suppose that the η_i equal zero. Then equation (20) estimates $\bar{\gamma} + \bar{\delta}$. These are the same quantities estimated in table 2, with the same biases. In this case, our aggregation-corrected estimates should recover table 2.

As a second case, though, suppose that $\bar{\delta}$ and $\bar{\eta}$ are equal in magnitude and opposite in sign and that the σ_i are close to zero. In this case, the aggregate estimate from equation (20) correctly estimates $\bar{\gamma}$ without bias. Then the estimated $\hat{R}$ correctly controls not just for aggregation from industries to aggregates, but for aggregation from firms to industries. Thus the estimates in table 3 provide the correct weighted average of firm-level returns to scale, whereas the estimates in table 2 (though relying only on disaggregated data) do not.

In our data, we can only speculate about which case is closer to the truth since we do not observe the R_i. Nevertheless, our priors are that case 2 is closer to the truth and hence that reallocation-corrected estimates provide better estimates of the average firm parameters. First, table 2 suggests that $\bar{\delta}$ is negative, with a magnitude of at least -0.17 (given a point estimate of 0.83 and a minimum on economic grounds of 1.00). Second, results in this section suggest that the aggregate reallocation term R is procyclical; if the reallocation response to aggregate shocks is similar within as well as across industries, then $\bar{\eta}$ is positive. Third, while the estimated σ_i generally exceed zero, the estimates have little correlation with the sectoral cyclicality of inputs β_i, so that the bias from the omitted final term in equation (20) is probably small.[15]

[15] It is, of course, surprising that σ_i and β_i have such a small correlation, given that $\sigma_i = \beta_i \operatorname{var}(dx)/\operatorname{var}(dx_i)$. In the data, high-$\beta_i$ industries, such as autos, tend to be high-$\operatorname{var}(dx_i)$ industries. To get a sense for the magnitude of the potential bias,

The results in this section thus suggest that the average firm appears to have approximately constant returns to scale and that aggregation effects can explain the three puzzles presented in Section III. Nevertheless, reallocation produces the same changes in average productivity as though there were a representative firm with increasing returns. So the economic question remains, Which estimate more accurately captures the economic forces at work in business cycles? The next section explores this question.

V. Implications for Calibrating Macro Models

When can we ignore heterogeneity in production and act as though a representative firm produces all output? Doing so means modeling the production side of the economy using an aggregate production function for GDP. In this section we first ask, Is this procedure ever sensible if the world actually has significant heterogeneity? We then ask, What parameters should one use to calibrate the assumed aggregate production function? In particular, can one use estimates from aggregate data and ignore heterogeneity?

We address these questions using a very simple, stylized example. In one case, it is safe to ignore heterogeneity: we can model a multisector world as though there were only one sector and calibrate it using only aggregate data, as one would calibrate a one-sector model. But in another case, this procedure is misleading. The appropriate procedure depends on nonproduction institutions in the economy. Thus our example shows that there are no general answers to the questions we posed: the answers depend on the precise model.

Our example modifies Basu (1995*a*). The world is static. There is a continuum of consumers, indexed by $j \in [0, 1]$. Each maximizes expected utility from consumption c and leisure $\bar{l} - l$:

$$\max E(U) = E[\alpha c + v(\bar{l} - l)] \tag{21}$$

subject to a standard budget constraint $c \leq \pi + wl$, where $\bar{l}$ is an individual's endowment of time, l is labor supply, π is profit, and w is the real wage.

A. *An Economy with a Single Firm*

Initially, suppose that the production side of the economy consists of a single firm that employs all workers. Aggregate labor L is defined

assume that the unobserved η_i are identical across sectors. The bias in estimating returns to scale caused by the final term is then $-\bar{\eta} \sum w_i(\sigma_i - \bar{\sigma})\beta_i$, which in the data equals $-\bar{\eta}^*(0.05)$. Even if η is large (0.5, say), the bias is small.

as $\int_0^1 l_j dj$, and the firm's production function is

$$Y = AL^{\gamma}, \tag{22}$$

where $\gamma \geq 1$. If $\gamma > 1$, then the firm must price with a markup above marginal cost to cover its costs. Although only one firm produces at any given time, we assume that there are a large number of potential entrants that can enter costlessly and capture any economic profits. Thus the incumbent sets its price to make exactly zero profits, so from equation (3), the firm's markup μ equals returns to scale γ. The real wage in this economy is thus

$$w = \frac{\gamma}{\mu} AL^{\gamma-1} = AL^{\gamma-1}.$$

Equilibria of this economy are symmetric, so $L = l$. So far there is no uncertainty in the economy, and in equilibrium the first-order condition for utility maximization is

$$v'(\bar{l} - L) = \alpha AL^{\gamma-1}. \tag{23}$$

We assume that the marginal utility of leisure rises as the consumer supplies more labor; thus $L_1 > L_2$ implies $v'(\bar{l} - L_1) > v'(\bar{l} - L_2)$. To ensure the existence of an interior equilibrium, we assume conditions on v such that $l > 0$ and $l < \bar{l}$ for all L. As in Cooper and John (1988), this economy necessarily displays multiple, symmetric, Pareto-ranked equilibria if, for some L^* that solves equation (23), we have

$$\frac{d \ln v'}{d \ln l} < \left. \frac{d \ln w'}{d \ln l} \right|_{dL=dl} = \gamma - 1. \tag{24}$$

Clearly, for (24) to hold, γ must be sufficiently larger than one: Increasing returns of a strong form (diminishing marginal cost) lead to an upward-sloping labor demand curve and hence make the real wage rise with aggregate labor supply. If this effect is sufficiently strong, the increase in the wage more than offsets the disutility of supplying additional labor. Given the assumptions about v above, we shall then have an odd number of locally unique equilibria, with higher-L equilibria Pareto-preferred. In this economy, therefore, the critical parameter that determines whether there are multiple equilibria is the degree of returns to scale of the representative firm, which can be estimated using aggregate data.[16]

[16] As we noted above, the increasing returns need to take the form of diminishing marginal cost. This is a robust feature of models in which there are increasing returns to agglomeration with a constant markup (e.g., Farmer and Guo 1994). Of course, as eq. (3) shows, increasing returns to scale just means that average cost exceeds marginal cost and is compatible with any slope of the marginal cost curve. One can calibrate the slope of the marginal cost curve from estimates of the degree of returns to scale only by assuming that there are no fixed costs.

B. An Economy with Heterogeneity

We now consider two examples in which the production side of the economy consists of two firms, each producing with constant returns to scale. However, one of the firms has a higher level of productivity than the other. Thus $Y_i = A_i L_i$ $(i = 1, 2)$, and we assume $A_1 > A_2$.

Both firms are competitive and price at marginal cost. However, unions in firm 1 succeed in extracting all the benefits of higher productivity in the form of higher wages, so both firms have equal unit costs (and equal prices). Hours in the high-productivity firm are rationed. Thus w_1, the wage in firm 1, equals A_1 and w_2 equals A_2. The rationing rule takes the form of having a share, s, of aggregate labor hours devoted to work in firm 1, with the remainder going to firm 2: $L_1 = sL$ and $L_2 = (1 - s)L$. Thus the average wage, $\bar{w}$, is $sA_1 + (1 - s)A_2$. Finally, we let s depend on aggregate output: $s = s(Y)$. For reasons we do not model, we assume $s'(Y) > 0$.

Now suppose that we estimate ''returns to scale'' by estimating equation (2) or (6) using aggregate data from this economy. As in Basu (1995*a*), we find

$$\hat{\gamma} \equiv \frac{d \ln Y}{d \ln L} = \frac{1}{1 - s\epsilon_s \dfrac{A_1 - A_2}{sA_1 + (1 - s)A_2}}, \tag{25}$$

where $\epsilon_s \equiv (ds/dY)(Y/s)$. Since ϵ_s is positive, $\hat{\gamma} > 1$. Let us suppose that ϵ_s, A_1, and A_2 are chosen so that $\hat{\gamma}$ equals the representative firm's γ in equation (22). Does this parameter play the same role here as it does in the single-firm economy, where (for given v) it determines whether the economy has multiple equilibria?[17]

1. Egalitarian Rationing

Suppose that we first consider one institution of labor supply, which we might term ''egalitarian rationing.'' Under this scheme, each worker is allowed to work a fixed number of hours, H, in firm 1 and then is free to work as many hours as he or she wishes in firm 2. In equilibrium, $H = s(Y)l = s(Y)L$. So an individual worker's budget constraint becomes $c \leq \pi + w_1 H + w_2(\bar{l} - H - x)$.

[17] In this example the failure of an aggregate production function comes from imperfect competition in the labor market. We assumed away this source in Sec. III for simplicity, as well as because we have no way to assess its importance in our data. Katz and Summers (1989) argue that there is empirical evidence in favor of significant labor rents of the sort assumed in this example, but Topel (1989) criticizes their assumptions. This argument does not concern us, however, since our simple, static example is not meant to describe the world, but simply provides a convenient example to establish that certain results are possible.

Note that at the margin (as long as H is not too large) the wage is always the low wage, w_2. Small changes in Y (and hence $s(Y)$ and H) do not change this property. From the point of view of the worker, the change in the average wage is inframarginal and is not reflected in a change in the marginal compensation for labor. Then we can see that equation (24) is not satisfied, and this economy has a unique equilibrium. Thus, although this economy has procyclical labor productivity (and procyclical labor compensation per hour) to exactly the same degree as our one-firm economy, it does not have multiple equilibria. In this case, assuming a representative firm and calibrating it with the parameter estimated from aggregate data does not accurately describe the economics at work.

2. Rationing by Lottery

Now suppose that instead of high-paid hours being rationed evenly among workers, the opportunity to work at the high-wage firm is allocated by lottery. All workers enter the lottery and specify in advance the number of hours they are willing to work, regardless of the firm at which they are employed. Thus workers make their labor supply decisions without knowing their individual ex post real wages. They do know, however, the probability distribution of wages.

Since the problem is now one of individual (though not aggregate) uncertainty, the first-order condition becomes

$$v'(\bar{l} - l) = \alpha E(w) = \alpha\{s(Y)A_1 + [1 - s(Y)]A_2\}. \tag{26}$$

The condition for multiple equilibria thus is

$$\frac{d \ln v'}{d \ln l} < \frac{d \ln w'}{d \ln l} = \frac{A_1 - A_2}{sA_1 + (1 - s)A_2} s\epsilon_s \frac{d \ln Y}{d \ln L} = \hat{\gamma} - 1. \tag{27}$$

Equation (27) shows that in this case the two-sector economy with heterogeneity has multiple equilibria under just the same condition as the one-sector economy: if the $\hat{\gamma}$ estimated from aggregate data is above the same critical value in both cases.

C. Implications

This example shows that one cannot draw general lessons about calibrating one-sector models from aggregate data in the presence of significant heterogeneity. Nevertheless, two features of the example seem to have general implications.

First, we note that if any single summary statistic is useful, it is

likely to be the degree of returns to scale estimated from aggregate data *without* composition corrections. In all the cases we examined, this parameter correctly captures the procyclical behavior of average labor productivity and average labor compensation (under the assumption that the $s(Y)$ function is structural). A one-sector model calibrated with the average γ (always one in the multifirm economy) would be unable to replicate this behavior.

Second, this example shows that the aggregate parameter also can fail to capture the relevant economics in the multifirm economy, although it always does so in the single-firm economy. The reason is that in the multifirm economy, the procyclicality of aggregate productivity always reflects the procyclical ''average marginal product'' of labor, but only under some circumstances does that translate to a procyclical ''marginal marginal product'' of labor. However, it is the latter that is relevant for economic decisions (in this case, labor supply). Thus the lesson may be that it is safer to ignore reallocation if the institutions being modeled ensure that the average and marginal factor prices move more or less together. However, it is difficult to say when this would occur since there is no completely accepted theory of imperfect competition that explains how the same input can have different marginal products in different uses.

VI. Conclusions

On both empirical and theoretical grounds, heterogeneity in production appears important for macroeconomics. Although estimates of returns to scale vary widely across relatively disaggregated industries, the average industry appears to produce with constant or even decreasing returns. Taken literally, apparent widespread decreasing returns contradicts empirical evidence of small economic profits. These industry results also contrast sharply with aggregate results at the level of manufacturing and the private business economy, which show large increasing returns. A representative-firm framework, used in many recent macro models, cannot easily explain these findings. Another popular explanation for procyclical productivity—unobserved changes in factor utilization—also cannot explain these findings since variable utilization explains apparent industry increasing returns rather than apparent decreasing returns.

An explanation for these findings is that neither industries nor aggregates behave like firms. In particular, aggregating across imperfectly competitive producers can explain the puzzles we identify since similar factors employed in different industries can then have different marginal products. For example, durable-goods industries appear to have larger returns to scale than the average industry. The

output of durable-goods industries is also more procyclical than the average. Thus the additional factors employed in a boom have marginal products that are higher than the average products of factors in use, leading aggregate productivity to be procyclical.

If long-run returns to scale cannot be lower than one, this story in its pure form requires that some industries have increasing returns.[18] However, aggregation can exaggerate a modest degree of increasing returns, and an econometrician might misinterpret this as evidence for large increasing returns at a representative firm. On the other hand, reallocation effects can in principle work in the opposite direction, making productivity countercyclical in some industries and thus explaining the puzzle of apparent industry diminishing returns. Hence, aggregation issues can explain differences in results at different levels of aggregation, in different types of data, and over different sample periods.

We correct for economywide reallocation to the extent feasible with the available data. These results suggest that aggregation effects are important in the data. They also suggest that a typical firm produces with approximately constant returns to scale, although we cannot reject a modest degree of increasing returns. These results are consistent with those from plant-level studies, such as Baily et al. (1992). In addition to finding constant returns at a plant level, Baily et al. find significant differences in productivity levels among plants. These are of course the features of our example in Section V.

Many macroeconomic issues depend on the production parameters within the economy. However, the relevant parameter may not be the weighted average of returns to scale over all firms in the economy. Cyclical reallocations are not simply a bias in estimating returns to scale; they cause real output to vary and may themselves serve as important macroeconomic mechanisms. We illustrate this theoretical possibility in a simple model with multiple heterogeneous firms, each producing with constant returns to scale. This economy can display multiple Pareto-ranked equilibria, a result requiring large increasing returns if production took place at a single firm. Furthermore, the critical parameter is the degree of returns to scale that an econometrician would estimate using aggregate data without correcting for reallocations.

This multifirm economy does not always exhibit multiple equilibria, however, and whether it does so depends on nonproduction features of the economy. Thus, to investigate many recent hypotheses in macroeconomics, one may need to construct multisector dynamic

[18] As our example of Sec. V shows, this need not be true if there is imperfect competition in factor markets.

general equilibrium models with imperfect competition and heterogeneity and confront the aggregation issues from first principles. Until then, the fact that the world seems best described by approximately constant returns at the firm level does not necessarily allow us to reject macroeconomic parables in which increasing returns at a representative firm play a central role in explaining economic fluctuations. Ascertaining which paradigm provides better macroeconomic insights is an important task for future research.

Appendix A

Data Sources and Methods

We use unpublished data provided by Dale Jorgenson and Barbara Fraumeni on industry-level gross output and inputs of capital, labor, energy, and materials. The data set covers 21 manufacturing industries and 13 nonmanufacturing industries.[19] These sectoral accounts seek to provide accounts that are, to the extent possible, consistent with the economic theory of production. These data are available both with and without an adjustment for input quality. The quality adjustment essentially involves taking account of changes over time in input composition. Computers, for example, give a higher service flow per dollar than factories since they depreciate faster. Similarly, engineers and janitors make different marginal contributions to output, and one can use information on relative factor payments to adjust for the differences.

We calculate the variables we need from the Jorgenson data. The equations in the text are all derived in continuous time; in all cases, we approximate differentials with log differences and instantaneous shares with averages in periods t and $t - 1$. This Tornquist approximation is exact if the underlying production function is translog; otherwise, it provides a flexible second-order approximation to any function.

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, $P_K K$, is then $r\pi^K K$, where $\pi^K K$ 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 50 types of capital, plus land and inventories, distinguished by the Bureau of Economic Analysis in constructing the national product accounts. Hence, for each of these 52 assets, we compute the user cost of capital as

[19] The manufacturing industries match the two-digit classification, except that transport equipment (SIC 37) is divided into ''motor vehicles'' (SIC 371) and ''other transport equipment.'' The nonmanufacturing industries comprise agriculture; metal mining; coal mining; oil and gas extraction; nonmetallic mining; construction; transportation; communications; electric utilities; gas utilities; trade; finance, insurance, and real estate; and services.

$$r_s = (\rho + \delta_s)\frac{1 - \mathrm{ITC}_s - \tau \cdot d_s}{1 - \tau}, \quad s = 1, \ldots, 52, \tag{A1}$$

where ρ is the required rate of return on capital. For each asset, δ_s is the depreciation rate, ITC_s is the investment tax credit, and d_s is the present value of depreciation allowances; τ is the corporate tax rate. We assume that the required return ρ equals the dividend yield on the Standard & Poors 500. We obtained unpublished data on ITC_s, d_s, and τ from Dale Jorgenson.

We estimate that the typical industry has an average profit rate of about 3 percent. Given uncertainty about whether the dividend yield appropriately captures the cost of funds, we have experimented with several alternative measures of the capital cost, as discussed in Basu and Fernald (1995*b*). Even assuming zero profits, so that revenue and cost shares are equal, has little effect on our results. This is unsurprising since economic profits do not appear large in any of our measures.

We create aggregates as Divisia indices over the underlying industries. Note that aggregate gross output suffers substantial double counting from the expenditure side of the national accounts. If our interest is on the production side—for estimating returns to scale, for example—then aggregation effects are a significant issue but double counting is not.

Appendix B

Correlation of the Hall-Ramey Instruments with Aggregation Errors

This Appendix suggests that the Hall-Ramey instruments are not valid because they are correlated with aggregation effects. Hence, while demand-side instruments are likely to be valid at a firm level, they do not appear to be valid at higher levels of aggregation. This Appendix presents the results cited in Section IV.

For each level of aggregation, we calculate the gross-output aggregation term R and the value-added terms $R^V + I$, defined in Section III. In aggregate data, these terms constitute part of the regression disturbance term. We regress these aggregation terms on each of the Hall-Ramey instruments; the tables below present marginal significance levels, as in Hall (1990, tables 5.1, 5.2). If the instruments are valid instruments, uncorrelated with the disturbance term, then it is necessary that they be uncorrelated with the aggregation terms. (It is, of course, not sufficient. As discussed in the text, the instruments may be invalid for reasons other than aggregation biases. Moreover, even if the instruments affect reallocations across industries, we may not detect this effect statistically if the standard errors are large.)

To calculate the aggregation terms, we need industry-level estimates of returns to scale. If the instruments are valid, one wants to use them to estimate the industry returns to scale. If the instruments are not valid, of course, then these industry estimates may also be invalid. (Even if there is feedback from technology shocks to input growth, so that OLS is biased, there is

no reason to expect this to cause the calculated reallocation terms to be correlated with the instruments.) We calculate these terms both ways: table B1 shows results using the *uninstrumented* estimates of industry parameters, and table B2 shows results using the *instrumented* estimates.

The instruments are the current and once-lagged values of (i) the growth rate of the world price of oil (oil), (ii) the growth rate of government defense spending deflated by the GDP deflator (gdef), and (iii) the political party of the president/party in power (pip).

For the change in the world price of oil, we can usually reject the null that the instruments are not associated with aggregation effects. In table B2 for the private economy as a whole, for example, the coefficient on the current oil price is significantly negative at the .001 level; the coefficient on the lagged change is significantly negative at the .01 level for value added and the .05 level for gross output. Government defense spending positively affects reallocations for manufacturing durables but is otherwise insignificant, as is the political party of the president. These results thus verify the empirical importance of the theoretical point that the instruments we (and many others) use are not valid as instruments at an aggregate level.

TABLE B1

UNINSTRUMENTED ESTIMATES

	Private Economy	Nonmanufacturing	Manufacturing	Durables Manufacturing	Nondurables Manufacturing
	A. Dependent Variable: Estimated Gross-Output Aggregation Terms				
oil	.034*	.012*	.113	.455	.019*
oil(−1)	.015*	.021*	.181	.600	.056
gdef	.382	.587	.189	.001**	.945
gdef(−1)	.760	.495	.847	.579	.185
pip	.766	.311	.887	.107	.711
pip(−1)	.904	.549	.730	.309	.689
	B. Dependent Variable: Estimated Value-Added Aggregation Terms				
oil	.064	.104	.049	.430	.507
oil(−1)	.021*	.013*	.293	.860	.014*
gdef	.367	.685	.063	.001*	.663
gdef(−1)	.829	.404	.634	.730	.970
pip	.689	.341	.739	.065	.560
pip(−1)	.955	.689	.911	.276	.351

NOTE.—Entries are marginal significance levels for regression on each of the instruments.
* Significant at the 5 percent level.
** Significant at the 1 percent level.

TABLE B2

INSTRUMENTED ESTIMATES

	Private Economy	Nonmanufacturing	Manufacturing	Durables Manufacturing	Nondurables Manufacturing
	A. Dependent Variable: Estimated Gross-Output Aggregation Terms				
oil	.000**	.000**	.105	.267	.060
oil(−1)	.036*	.007**	.233	.603	.265
gdef	.763	.924	.523	.006**	.400
gdef(−1)	.191	.262	.577	.779	.070
pip	.334	.300	.809	.439	.370
pip(−1)	.570	.372	.538	.794	.767
	B. Dependent Variable: Estimated Value-Added Aggregation Terms				
oil	.001**	.014*	.015*	.247	.168
oil(−1)	.005**	.002**	.250	.777	.005**
gdef	.797	.780	.323	.021*	.547
gdef(−1)	.241	.193	.962	.826	.790
pip	.377	.525	.986	.250	.550
pip(−1)	.428	.586	.974	.734	.415

NOTE.—Entries are marginal significance levels for regression on each of the instruments.
* Significant at the 5 percent level.
** Significant at the 1 percent level.

References

Baily, Martin Neil; Hulten, Charles; and Campbell, David. ''Productivity Dynamics in Manufacturing Plants.'' *Brookings Papers Econ. Activity: Microeconomics* (1992), pp. 187–249.

Bartelsman, Eric J. ''Of Empty Boxes: Returns to Scale Revisited.'' *Econ. Letters* 49 (July 1995): 59–67.

Basu, Susanto. ''Comment.'' In *NBER Macroeconomics Annual 1995,* vol. 10, edited by Ben S. Bernanke and Julio J. Rotemberg. Cambridge, Mass.: MIT Press, 1995. (*a*)

———. ''Intermediate Goods and Business Cycles: Implications for Productivity and Welfare.'' *A.E.R.* 85 (June 1995): 512–31. (*b*)

———. ''Procyclical Productivity: Increasing Returns or Cyclical Utilization?'' *Q.J.E.* 111 (August 1996): 719–51.

Basu, Susanto, and Fernald, John G. ''Aggregate Productivity and the Productivity of Aggregates.'' Working paper no. 5382. Cambridge, Mass.: NBER, December 1995. (*a*)

———. ''Are Apparent Productive Spillovers a Figment of Specification Error?'' *J. Monetary Econ.* 36 (August 1995): 165–88. (*b*)

Beaudry, Paul, and Devereux, Michael. ''Monopolistic Competition, Price Setting and the Effects of Real and Monetary Shocks.'' Manuscript. Vancouver: Univ. British Columbia, 1994.

Benhabib, Jess, and Farmer, Roger E. A. ''Indeterminacy and Sector-Specific Externalities.'' Manuscript. New York: New York Univ., October 1995.

Bils, Mark, and Cho, Jang-Ok. ''Cyclical Factor Utilization.'' *J. Monetary Econ.* 33 (April 1994): 319–54.

Bruno, Michael. ''Raw Materials, Profits, and the Productivity Slowdown.'' *Q.J.E.* 99 (February 1984): 1–29.

Burnside, Craig. ''What Do Production Function Regressions Tell Us about Increasing Returns to Scale and Externalities?'' Manuscript. Washington: World Bank, November 1995.

Burnside, Craig; Eichenbaum, Martin; and Rebelo, Sergio. ''Capital Utilization and Returns to Scale.'' In *NBER Macroeconomics Annual 1995,* vol. 10, edited by Ben S. Bernanke and Julio J. Rotemberg. Cambridge, Mass.: MIT Press, 1995.

Caballero, Ricardo J., and Lyons, Richard K. ''External Effects in U.S. Procyclical Productivity.'' *J. Monetary Econ.* 29 (April 1992): 209–25.

Cooper, Russell, and John, Andrew. ''Coordinating Coordination Failures in Keynesian Models.'' *Q.J.E.* 103 (August 1988): 441–63.

Domowitz, Ian; Hubbard, R. Glenn; and Petersen, Bruce C. ''Market Structure and Cyclical Fluctuations in U.S. Manufacturing.'' *Rev. Econ. and Statis.* 70 (February 1988): 55–66.

Farmer, Roger E. A., and Guo, Jang-Ting. ''Real Business Cycles and the Animal Spirits Hypothesis.'' *J. Econ. Theory* 63 (June 1994): 42–72.

Fisher, Franklin M. *Aggregation: Aggregate Production Functions and Related Topics.* Collected Papers of Franklin M. Fisher, edited by John Monz. Cambridge, Mass.: MIT Press, 1993.

Griliches, Zvi, and Ringstad, V. *Economies of Scale and the Form of the Production Function: An Econometric Study of Norwegian Manufacturing Establishment Data.* Amsterdam: North-Holland, 1971.

Hall, Robert E. ''The Relation between Price and Marginal Cost in U.S. Industry.'' *J.P.E.* 96 (October 1988): 921–47.

———. ''Invariance Properties of Solow's Productivity Residual.'' In *Growth/ Productivity/ Unemployment: Essays to Celebrate Bob Solow's Birthday,* edited by Peter Diamond. Cambridge, Mass.: MIT Press, 1990.

Hall, Robert E., and Jorgenson, Dale W. ''Tax Policy and Investment Behavior.'' *A.E.R.* 57 (June 1967): 391–414.

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

Katz, Lawrence F., and Summers, Lawrence H. ''Industry Rents: Evidence and Implications.'' *Brookings Papers Econ. Activity: Microeconomics* (1989), pp. 209–75.

Nelson, Charles R., and Startz, Richard. ''Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator.'' *Econometrica* 58 (July 1990): 967–76.

Norrbin, Stefan C. ''The Relation between Price and Marginal Cost in U.S. Industry: A Contradiction.'' *J.P.E.* 101 (December 1993): 1149–64.

Perli, Roberto. ''Indeterminacy, Home Production, and the Business Cycle: A Calibrated Analysis.'' Manuscript. New York: New York Univ., April 1995.

Rotemberg, Julio, and Woodford, Michael. ''Oligopolistic Pricing and the Effects of Aggregate Demand on Economic Activity.'' *J.P.E.* 100 (December 1992): 1153–1207.

———. ''Imperfect Competition and the Effects of Energy Price Increases on Economic Activity.'' Manuscript. Cambridge: Massachusetts Inst. Tech., 1993.

———. ''Dynamic General Equilibrium Models with Imperfectly Competitive Product Markets.'' In *Frontiers of Business Cycle Research,* edited by Thomas F. Cooley. Princeton, N.J.: Princeton Univ. Press, 1995.

Schmitt-Grohé, Stephanie. ''Comparing Four Models of Fluctuations Due to Self-Fulfilling Expectations.'' Manuscript. Washington: Board Governors, Fed. Reserve System, 1995.

Shapiro, Matthew D. ''Cyclical Productivity and the Workweek of Capital.'' *A.E.R. Papers and Proc.* 83 (May 1993): 229–33.

Staiger, Douglas, and Stock, James H. ''Instrumental Variables Regression with Weak Instruments.'' Technical Working Paper no. 151. Cambridge, Mass.: NBER, January 1994.

Topel, Robert H. ''Comment.'' *Brookings Papers Econ. Activity: Microeconomics* (1989), pp. 283–88.