NBER WORKING PAPER SERIES

PRODUCTION FUNCTIONS: THE SEARCH FOR IDENTIFICATION

Zvi Griliches
Jacques Mairesse

Working Paper No. 5067

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

We are indebted to Aviv Nevo for able research assistance, to G. Chamberlain, B. Crepon, B. H. Hall. T.J. Klette, Y. Mundlak, and A. Pakes for helpful comments, only a fraction of which could be incorporated in this draft, and to the National Science Foundation and the Sloan Foundation for financial support. This paper is part of NBER's research program in Productivity. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Working Paper #5067
March 1995

PRODUCTION FUNCTIONS: THE SEARCH FOR IDENTIFICATION

ABSTRACT

Some aspects of the econometric estimation of production functions are discussed, focussing primarily on the issue of simultaneity and reviewing the stream of criticisms of Douglas' work and the response to it. We look in particular at the work that uses panel data on micro data for plants or firms and at some more recent multi-equation extensions of it. We find that researchers, in trying to evade the simultaneity problem, have shifted to the use of thinner and thinner slices of data, exacerbating thereby other problems and misspecifications. We describe the need for better data, especially on product prices at the individual observation level and on relevant cost and demand shifters, and for better behavioral theories which would encompass the large amount of heterogeneity observed at the micro level.

Zvi Griliches
Department of Economics
125 Littauer Center
Harvard University
Cambridge, MA 02138
and NBER

Jacques Mairesse
Institut National de la Statistique et
des Etudes Economiques, (INSEE)
CREST
15, boulevard Gabriel PERI
92245 Malakoff CEDEX
FRANCE
and NBER

PRODUCTION FUNCTIONS: THE SEARCH FOR IDENTIFICATION

Zvi Griliches and Jacques Mairesse[1]

> "...we have here one of those cases - so frequent in economic practice - where it can be "proved" by abstract reasoning that a solution is not possible, but where life itself compels us nevertheless to find a way out." Frisch (1934), p. 274.[1]

1. Introduction

Econometric production functions, as we know them today, originated in the work of Douglas (Cobb and Douglas, 1928). They started out as a tool for testing hypotheses about the workings of marginal productivity theory and the competitiveness of labor markets, using macro-data. After a number of withering attacks by critics, this line of work shifted from macro to micro data, especially in agriculture (see Tintner (1944), Mundlak (1961), and Heady and Dillon (1961) for examples), where the necessary assumptions for estimating something sensible appeared to be more plausible. The growing interest in technical change brought production functions back as a measurement framework for productivity and led to a resurgence of interest in their estimation. A "dual" literature on cost functions, factor demand systems, and flexible functional forms developed in parallel, but we will only touch on it in passing here (see Jorgenson (1986) for a review). Both approaches have undergone significant econometric questioning, focused primarily on whether the relationship of interest is really "identified" or even identifiable with the data at hand.

In the empirical implementation of econometric production functions one runs into a number of difficult conceptual and data problems: (1) Is what we are looking for really out "there"? What is it that we are approximating when we write down Q=F(K, L,...) and apply it to a particular data set, across different firms and different time periods? Does the assumed

---

[1] Harvard University and NBER, and Crest INSEE, Paris and NBER. We are indebted to Aviv Nevo for able research assistance, to G. Chamberlain, B. Crepon, B. H. Hall, T. J. Klette, Y. Mundlak, and A. Pakes for helpful comments, only a fraction of which could be incorporated in this draft, and to the National Science and Sloan Foundations for financial support.

functional form make sense? (2) Do we have the right data set for this enterprise? Are all outputs and inputs accounted for and measured correctly? Have allowances been made for differences in the quality of labor and other resources across our units observation? Differences in the utilization of capital? (3) How was the sample generated? Is it representative of the population of interest? (4) Is the proposed estimation procedure appropriate? Can one take all or most of the input variables as "independent" or should one embed this problem in a larger simultaneous equations framework? Do the market structure and behavioral assumptions make sense? Can they be tested? And much more.

In this paper we shall limit ourselves only to the last question, even though some of the other questions, especially those connected with data and sample construction, may be more important substantively.[2] We shall look at this question from a historical perspective, tracing out the changing ideas and justifications offered in defense of the various suggested econometric procedures and how this process was affected by developments in data availability and computing ability. We shall reiterate an old observation of ours (Griliches (1977), p.12-13) that living in a second-best world one should be careful in focusing only on one problem. Fixing it may only aggravate other problems in our data and not lead us, necessarily, closer to the "truth," wherever that may be.

Before we turn to the more substantive discussion, a word should be said why researchers are actually interested in estimating production functions. In most cases it is a tool, a framework for answering other questions, only partially related to the "production function" itself. It is hard even to pose some questions without embedding them in such a framework. For example, one may be interested in the presence or absence of economies of scale in production (Griliches and Ringstad (1971), Mairesse (1975)), in whether productivity differences are related to differences in the quality of labor as measured by education and to differences in public and private investments in research (Griliches, (1964), Griliches and Mairesse (1984), Griliches 1986)), whether marginal products are equated to the apparent factor prices (Hall (1988), Hellerstein and Neumark (1993)), and what is the market structure in various industries and how high are the markups there. Such questions require independent estimates of cost or production functions and they are too interesting for us to give up trying to answer them even though the estimation framework used for these purposes may be itself problematic. One has to start someplace.

2. The Basic Criticism.

One of the first penetrating criticisms of Douglas's work came from Oslo. In his 1938 **Econometrica** article, Horst Mendershausen, after thanking Frisch, Haavelmo, and Reirsol for their comments and referring to the then still unpublished Frisch book on Production Theory, argued that data used by Douglas were too multi-collinear to allow for a "reliable" determination of the production function coefficients. While not saying so explicitly, he asserted that the underlying production function is not "identifiable" because the input variables are determined simultaneously by the same forces.[3] He demonstrated the last point by computing the implicit coefficients from the "other direction" regressions and showing that there was not enough independent variability in the data in the **right** directions.[4]

The first application of the simultaneous equations methodology to production function estimation, based explicitly on Haavelmo's earlier papers, is in Marschak and Andrews (1944). Their formulation of the problem cannot really be improved upon:

"...Can the economist measure the effect of changing amounts of labor and capital on the firm's output - the "production function" - in the same way in which the agricultural research worker measures the effect of changing amounts of fertilizers on the plot's yield? He cannot, because the manpower and capital used by each firm is determined by the firm, not by the economist. This determination is expressed by a system of functional relationships; the production function, in which the economists happens to be interested, is but one of them...," Marschak and Andrews p. 144.[5]

Their main point can be illustrated with a very simple two-equation model of producer behavior. The first is the production function

$$(1) \qquad y = \alpha z + \beta x + u$$

where y is the logarithm of output, z is the logarithm of capital (or all "fixed" inputs) and x is the logarithm of labor input (or all "variable" inputs). This is just a logarithmic transformation of the simple Cobb-Douglas production function, with u representing all other disturbances, left out factors, efficiency differences, functional form discrepancies, and errors of measurement.[6] The critical point made by Marschak and Andrews, and by others, is that one cannot really treat the right side variables as "independent" variables and proceed with estimation by "ordinary" least squares (OLS) as was done originally by Douglas and is still being done by most applied practitioners of this genre of research. Because the "inputs" are not under the control of the econometrician but are chosen in some optimal or behavioral fashion by the producers themselves, the usual exogeneity assumptions that are required for the consistency of OLS are unlikely to hold for the data at hand, at least not without further detailed analysis.

Even if one accepts the notion that "fixed" inputs (the z's) are predetermined for the duration of the relevant observation period, variable inputs (the x's) may be adjusted by the decision maker. A simple profit maximizing model of the joint determination of x and y, given z, product prices p (which, for now, will be assumed to be equal for different producers and hence normalized to 1), and factor prices w (*e.g.*, wages, or rather, their logarithms), implies the following marginal productivity condition (or variable input demand function)

$$y = x + w + v - \ln(\beta) \tag{2}$$

where v represents all discrepancies from the assumed conditions of perfect competition, perfect foresight, absence of risk aversion, and possible measurement errors in y, x, and w.[7] Ignoring constants and other unnecessary complications at this stage, one can solve these two structural equations to yield the associated reduced form equations:

$$x = D(\alpha z - (w + v) + u) \tag{3}$$

$$y = D(\alpha z - \beta(w+v) + u) \tag{4}$$

where $D=(1-\beta)^{-1}$. It is clear that if x is chosen even approximately optimally, then the production function disturbance u is "transmitted" to this decision equation, and x is a function of it. Simple OLS estimates of the production function will be biased and will not possess the desired structural interpretation. This is the main and plain message of the Marschak and Andrews paper and of a large number of subsequent expositions of this point (see, especially, Hoch (1962), Mundlak (1963), Mundlak and Hoch (1965), and Mundlak (1994)).

The responses to this message have taken a number of forms: The most common has been and still is denial, either outright by ignoring this message, or by claiming that the problem does not apply to the data at hand.[7] At first, in the mid-1940's and 1950's, most of the work on production functions had shifted to micro data, especially farm data, and many of these criticisms seemed to have less force. (There were almost no comparable micro data for industry available at that time.) In particular, it was possible to assert that much of what was contained in u, such as weather, and pests, was unanticipated by farmers, and that most of the resources used (land, machinery, family labor) were largely predetermined, and hence uncorrelated with it. This folklore position was formalized by Zellner, Kmenta and Dreze in their 1966 paper by putting the producer's decision within an uncertainty framework and having the producer maximize expected profits without knowing the realization of u. Under these circumstances, there is no transmission and hence no simultaneity bias. In spite of the eminence of the authors and the locus of publication (Econometrica), hardly anyone has been willing to offer this rationale for the use of OLS without feeling guilty about it. The "weather" story seemed especially inapplicable to industrial data. Moreover, even within agriculture it was perceived that u contained also expected permanent components such as land quality, and that the growing availability of panel data allowed a somewhat better solution to this problem.

At the macro level, the statistical estimation of production functions was largely abandoned for a while, with the emphasis shifting to the acceptance of marginal productivity theory (and the competition and constant returns to scale assumptions that went with it) and the use of factor shares as appropriate surrogates for the unknown parameters (Klein (1953), Solow, (1957)). Somewhat later came the "Duality" revolution in production function theory (Shepard, McFadden, Jorgenson and Lau, Diewert and others), the associated functional form generalizations and the move toward the estimation of cost functions and factor demand systems (Arrow et al. (1961), Nerlove (1963), Christenson, Jorgenson, Lau (1973)). In the rush to develop an exciting new field of research, the problem of simultaneity was largely ignored (with the notable exception of Nerlove, who used the regulatory environment within which his electrical utility firms operated to claim, quite reasonably, exogeneity for his output measures and factor prices), even though the problem is actually fully symmetric.[8]

In this paper, we focus primarily on the use of panel data and transformations, such as "within" and first-differences, which purport to eliminate this problem; on the use of lagged inputs as instrumental variables; and on the use of additional proxies and equations to substitute for the unobserved disturbance u (Olley and Pakes (1992).[9]

Many of these responses correspond to somewhat different interpretations about the sources of the correlation between x and u. Thus before we proceed to discuss them in greater detail, we need to be more explicit about the potential components of u, the error in the production function, and how it affects the producer's input decision.

3. The Anatomy of Error.

To discuss the identification problem in greater detail we need to specify the various potential sources of the "errors" in equations (1) and (2). Turning to the error in the first equation first, we can write it, tautologically, as

$$u_{it} = a_{it} + e_{it} + \varepsilon_{it} \tag{5}$$

where a and e are components of the disturbance that are known, ultimately, by the producer but not by the econometrician while ε is the net error that is introduced by measurement, data collection, and computational procedures. (It is net in the sense that it combines the effects of measurement errors in output minus the measurement errors in all the specified inputs: $\varepsilon = \varepsilon_y - \alpha\, \varepsilon_z - \beta \varepsilon_x$.) The distinguishing characteristic of ε is that it is the econometrician's problem, not the producers. It has no effect on producers behavior, either currently or in the future. In the useful terminology of Mundlak and Hoch ε is never "transmitted."[10]

The a and e components, on the other hand, are known to the producer but unobservable by the econometrician; a_{it} is known in time to affect the current x choice decisions while e reveals itself only later on and is not predictable ex ante. That is, ignoring errors of measurement, we are describing the case of "partial transmission" of u, where a may affect current behavior, while e will not, though it may have an effect on future behavior if its realization changes the producer's future expectations about u (delayed transmission).

Given this formulation, the e components should be serially uncorrelated, but a, and hence also x (and possibly also z), may be affected by their past values. The e's represent unforeseen environmental changes such as unusual weather conditions and unanticipated changes in unmeasured "effort" due to unexpected demand conditions and are bona fide "disturbances," independent of current x and z.

The a components represent all the miss-specifications made by the econometrician in describing the producer's situation, beyond the mismeasurement of the major observable variables already encapsulated in ε. They contain both important left out variables and functional form approximation errors. These would include unmeasured (by the econometrician) capital components such as R&D stocks and other intangibles, technology levels and managerial efficiency, unmeasured input quality (land, labor, and capital), unmeasured effort variables such as man and capital hours, and, possibly, higher order terms (squares, crossproducts, ...) of x and z.[11]

The main point about the a components is that they are known to the producer and hence they are "transmitted," to the extent that it is relevant, to the choice of x, and also, with some delay, to the longer run choice of the level of z.[12] This is the essence of the simultaneity problem for production function estimation.

Given this interpretation, equation (2) has to be rewritten as

$$(2') \qquad y - e - \varepsilon_y = x - \varepsilon_x + w + v - \ln(\beta)$$

and a should be substituted for u in equations (3) and (4) which implies that what is determined by the marginal productivity condition is the ***expected*** factor share

$$(2'') \qquad (x + w - \epsilon_x) - (y - e - \epsilon_y) = \ln(\beta) - v$$

net of measurement errors ϵ and unanticipated components e, and subject to the factor allocation disturbance v.

In turn, v can also be thought of as containing different components

$$(6) \qquad v = m + \omega$$

where ω represents individual differences in expected prices, both product and input, around their mean (official deflator) levels, and *m* reflects potential differences in markups and all other deviations from "simple" optimally, including risk aversion factors, functional form approximation errors, and actual errors in optimization.[13] In general, it is doubtful that one could assume independence between the *a* component of (1') and the *v* component of (2'), though this is an assumption that is often made.

After this long taxonomical digression we are ready to consider the various suggestions that have been made for dealing with the problem that components of *u* may also appear in the reduced form equation for *x*.

4. The Panel Data Response.

The major response to the Marschak and Andrews criticism came through the increasing availability of panel data. In that context, it was possible to assert that the major misspecifications, which are transmitted to the factor decisions, such as differences in land, entrepreneurial, and labor quality, are largely fixed over time (or at least over the length of the available panel) and hence can be eliminated by an appropriate "within" transformation.[14] Then (1) can be written as

$$(1') \qquad y_{it} = \alpha z_{it} + \beta x_{it} + a_i + \lambda_t + e_{it}$$

where the a_i's and λ_t's are "fixed" individual and time effects and can be eliminated by subtracting appropriate individual and time means (or by introducing a parallel set of dummy variables) and we are ignoring the errors of measurement ϵ for the time being. Ignoring also the time effects λ_t to simplify the exposition (they can be subsumed within the *z* set of variables), we have

$$(7) \qquad (y_{it} - y_{i\cdot}) = \alpha(z_{it} - z_{i\cdot}) + \beta(x_{it} - x_{i\cdot}) + (e_{it} - e_{i\cdot})$$

where the $x_{i\cdot}$ notation represents averaging over the time dimension for each individual *i*. To the extent that *e* is "not transmitted" (uncorrelated with *x* and *z*), the problem of "simultaneity" has been solved. The rest being "weather" or other unexpected events à la Zellner, Kmenta and

Dreze.

To the best of our knowledge, this approach is first stated briefly in Hoch (1955), based on a suggestion of C. Hildreth, with a fuller version appearing in Hoch (1962). It gained its popularity, however, from a series of remarkable papers by Mundlak (1961, 1963) and Mundlak and Hoch (1965). But as empirical work on production functions began to shift back to industrial micro-data (Krishna (1967), Griliches and Ringstad (1971), Ringstad (1971), Mairesse (1975), Griliches (1980), and Griliches and Mairesse (1984)), it encountered two problems, one empirical and the other theoretical. In empirical practice, the application of panel methods to micro data for manufacturing firms or plants produced rather unsatisfactory results - low and often insignificant capital coefficients and unreasonably low estimates of returns to scale. The second, not unrelated problem, is the realization that the within transformation is either not doing enough, in the sense that there are still potential simultaneity problems left, or too much, in the sense that this transformation may be aggravating other pre-existing problems, such as errors in variables.

The theoretical alarm bell was sounded by Chamberlain (1982) who pointed out that the within transformation which results in the elimination of the a_i, introduces $e_{i\cdot}$ into the equation and requires, therefore, strict exogeneity of the x's and not just their predeterminency. That is, there can be no delayed transmission, the e_{it}'s must be pure errors. Chamberlain also provided a framework for specification testing and for a more efficient estimation of equation (1') in the context of heteroscedastic micro data. His approach can be illustrated assuming the availability of panel data for three periods and writing the "augmented" conventional regressions (ignoring the z's and the constant terms for simplicity), where future and past values of x are added to the conventional specification, as

$$
\begin{aligned}
y_{1i} &= \pi_{11}x_{1i} + \pi_{12}x_{2i} + \pi_{13}x_{3i} + u_{1i} \\
y_{2i} &= \pi_{21}x_{1i} + \pi_{22}x_{2i} + \pi_{23}x_{3i} + u_{2i} \\
y_{3i} &= \pi_{31}x_{1i} + \pi_{32}x_{2i} + \pi_{33}x_{3i} + u_{3i}
\end{aligned}
\tag{8}
$$

Under the null hypothesis that there are no relevant left out variables, the expected "Π-matrix" equals

(9) $$\Pi_o = \begin{bmatrix} \beta & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \beta \end{bmatrix}$$

While if there are individual a_i's and they are correlated with the *x*'s, then one can write their expectation as $E(a_i | x_1, x_2, x_3) = \delta_1 x_{1i} + \delta_2 x_{2i} + \delta_3 x_{3i}$ and the expected Π-matrix is now

(10) $$\Pi_c = \begin{bmatrix} \beta+\delta_1 & \delta_2 & \delta_3 \\ \delta_1 & \beta+\delta_2 & \delta_3 \\ \delta_1 & \delta_2 & \beta+\delta_3 \end{bmatrix}$$

Chamberlain showed that each of these assumptions implies a set of restrictions on the unrestricted Π-matrix and can be tested in this context and, if the relevant null hypotheses are accepted, the parameters of interest can be estimated in a more efficient manner.

Chamberlain also anticipated the next set of responses along these lines: first differencing and GMM estimation. To get around the strict exogeneity requirements on *x*, one could proceed to first difference (or "longer" difference) the available panel data, rather than going "within." This also eliminates all individual "fixed effects," yielding

(11) $$y_t - y_{t-1} = \alpha(z_t - z_{t-1}) + \beta(x_t - x_{t-1}) + \zeta_t + e_t - e_{t-1}$$

where $\zeta_t = a_t - a_{t-1}$ is the *change* in *a*. If a is firmly fixed, $\zeta_t = 0$, and (11) can be estimated by Least Squares **if** e_t is unanticipated and untransmitted to x_t and x_{t+1}, even though it may affect subsequent decisions about *x* (and *z*).[15] If the "news" in *e* (or ζ_t) is transmitted to the current *x*, then the difference in *x* needs to be instrumented. Since the number of available instruments, lagged *x*'s and *z*'s, depends on the length of the panel and changes from cross-section to cross-section, the optimal estimation procedures become more complex, calling for the use of General Method of Moments estimators (see Arellano and Bond (1991), Griliches and Hausman (1986), and Keane and Runkle (1992), Mairesse and Hall (1993), among others). It can be shown (Crepon and Mairesse (1995)), that all of these approaches are already implicit in Chamberlain's Π-matrix approach and the modelling possibilities contained therein.

The difficulty with this line of attack is that the available instruments are likely to be quite poor and possess little resolving power. We shall return to this point below. It will suffice to note here that many economic variables evolve in a random walk like fashion at the micro level, *i.e.*, their log differences are only weakly serially correlated. If the x's and z's were strictly random walks, there would be no power at all in their past levels as instruments for current growth rates (first differences of their logarithms). If they do have power, one must ask what behavioral theory implies the validity of their use. Lags in adjustments to shocks would do it, but even if valid, the available power is likely to be rather low. To do better requires bringing in some additional information from somewhere else: "external" instruments, more theoretical restrictions on the structure, and/or more equations. This is where the current frontier lies. But before we describe it, we need to indicate the magnitude of the empirical problem faced by researchers in this area.

5. The Empirical Reality.

From the beginning the within (covariance analysis) results were not what one had hoped for. In the original 1955 note by Hoch, when farm effects are eliminated from the analysis, the estimated returns to scale fall from 1.0 to 0.6, forcing Hoch to interpret this "shortfall" as reflecting returns to unmeasured (fixed?) entrepreneurial capacity, but leaving him with unreasonably low estimated marginal returns to the relatively quasi-fixed labor and land inputs.[16] A more recent set of examples is given in Figure 1, taken from Mairesse (1990). It plots estimates of α and $\mu=\alpha+\beta$ from thirteen year panels of manufacturing firms in France, Japan, and the United States. For each country it graphs estimates based on the thirteen year firm averages for these firms ("between" in levels), on the total pooled data set in levels (TL), on the "within" firm deviations from their averages, in levels (WL), and a parallel set of estimates based on first differences in these same data sets (to be explained shortly). The important result visible in this graph, which is true for most of the data sets known to us, is that as one guards oneself against "left-out-variables" biases by purging various "components" from the available data and basing ones inferences on progressively smaller fractions of the original data, the estimated coefficients tend to move towards zero, with the capital coefficient falling faster than the labor coefficient and often actually reaching zero.

It is possible, of course, to respond that the original estimates were indeed biased upward and that the various "within" estimates are correct. But in most cases the changes and the implied biases are just too large to be credible. An alternative interpretation is that as we reduce the amount of variance in x and z that we use to identify the relevant coefficients other misspecifications overwhelm the remaining signal in the data. Before we discuss this further, we need to explain briefly the additional estimators introduced in Figure 1.

In going to first differences we eliminate whatever fixed individual effects there might have been in the data, but we also do more.[17] One can think of the resulting T-1 set of cross-sections of differences as another panel and do the whole covariance analysis partition into "between" and "within" once more. The resulting estimates are not really nested within the previous framework, they require an expansion and reinterpretation of equations (1) and (5). If we rewrite (5) as

(5')
$$u_{it} = a_{it} + \lambda_{it} + e_{it} + \epsilon_{it},$$

where now the "not so fixed" individual effect is the $a_{it} = \zeta_{it} + a_{it-1}$, and $\lambda_{it} = \lambda_t + g_i t$ is a period related component, allowing for both common period effects and differential individual trends. In this model the change in a (ζ) is independent of current e (and ϵ). Thus, a could be a random walk, but it could also depend on lagged e's (delayed transmission). The main difference between this formulation and (1) is, however, the introduction of individual trend terms which are not eliminated by the standard within transformation. Thus going to first differences (TD), we have

(12)
$$du_{it} = \zeta_{it} + c_t + g_i + de_{it} + d\epsilon_{it}$$

where d is the first difference operator and $c = d\lambda$ is now the period constant. The major argument for this transformation is that it allows us to consider (it protects us against) individual effects that change smoothly over time, but at different rates.[18] It is now more reasonable to interpret de as an unpredictable and hence serially uncorrelated component. If ζ were not currently transmitted to the measured independent variables and $g=0$, then one could estimated this equation by ordinary least squares. If ζ is transmitted, we will need valid instruments to estimate it, a point to which will come back to again below.

If productivity trends differ across individuals then the first differenced equation will also contain individual constants, the g_i's. These can be eliminated in turn by performing another within transformation on the panel of differences, which is what the WD estimates in Figure 1 refer to. Associated with this transformation is its complement: the "between differences" (BD), which in our data represent the average growth rates of the various variables for the period as a whole. The resulting numbers can also be interpreted as a "normalized" long difference, since $\Sigma(dy)/(T-1) = (y_T - y_1)/(T-1)$. The within differences "protect" us the most against various individual level and trend misspecifications at the cost of using the most restricted amount of variance in x and z for the identification of the various coefficients. In our data, which we believe are quite representative, the between differences transformation eliminates individual fixed effects with the least impact on the variances of the other variables (relative to WL and WD).

Returning to Figure 1, we can see that the estimates of α and μ decline as we buy more protection, moving from TL to WL, TD, and WD. In Table 1 we give the numbers for the French firms (N=441). Column 1 of this table gives the estimated α's for the various methods while column 3 lists the "net" variance of z (log of capital), net of its relationship to labor and time and industry dummies. The last columns provide the ingredients for an errors-in-variables interpretation of these results, to which we'll return below. First notice, however, that the net variance of z declines roughly in parallel with the estimated α's.[19] It is clear that a lower net variance will give us more imprecise estimates (holding degrees of freedom constant), but that does not imply a larger downward bias per se. The bias must come from some other source whose influence grows as the "signal" contained in the remaining variance of z keeps declining.[20]

An attempt to eliminate such biases using twice lagged values of y, x, and z as instrumental variables in a GMM frame work does not do any better (see the last row of Table 1). The resulting capital coefficient is not statistically significant in spite of the large number of degrees of freedom. It's net variance is minuscule. (The first stage R^2 of dz on all the instruments is only .1).

The simplest and oldest interpretation offered for such results is the magnification of the role random measurement errors in z as one reduces the identifying variance by various within or differencing transformations (and also by adding more variables to the estimating equation).[21] With uncorrelated measurement errors, the plim of $\hat{\alpha} = \alpha - \alpha\sigma^2_\epsilon m/\sigma^2_{zz}$, and σ^2_{zz} is the "net" variance of z, net of all the other variables included in the estimated equation, and where m is the multiplier for the error variance introduced by the various transformations. Thus, for example, m is 1/T for the BL and 2 for the TD estimators. Looking in Table 1 we can see that for, this sample, random measurement error bias in z would be 140 times larger in the WL estimates than in the BL ones and 17 times larger in the WD estimates than in the BD (or long differences) one. If one thought this was the only problem with the data, one could infer both a consistent estimate of α and an estimate of σ^2_ϵ from the numbers given in the first and last columns of Table 1. Using all six "observations," a regression of the estimated α on the last column yields .244 and .0035 as the respective estimates of α and σ^2_ϵ. Dropping the first two observations (TL and BL), since they may be subject correlated fixed effects bias, gives .161 and .0031 as the respective estimates. The latter regression has an adjusted R^2 of .93 and a standard error of estimate .02.[22] So this "model" fits rather well. Unfortunately, it does not make much sense in this particular context.

Errors of measurement in the capital stock are likely to persist. Hence a within or difference transformation should reduce their role, rather than magnify it! What might be then an alternative interpretation for these results? The simplest alternative would assert that we have the timing of the influence of capital wrong. Our measure of capital is based on gross stocks. That is, a given investment has an **assumed** constant weight over its estimated life. If there are gestation lags, difference transformations could produce entirely irrelevant numbers and the GMM estimates would not be of any help either.[23] That would also be true for other misspecifications in the assumed productivity life-pattern of investment.[24]

Other sources of misspecification could also contribute to the observed results. (They all could be guilty!) If some of the short run effects in the e's are transmitted to x, *i.e.*, we do have a simultaneity problem in the within dimensions, then β would be biased upward and some of this bias would also show up as a downward bias in α, since x and z are likely to be positively correlated.[25] Also, as we turn to the more short-run aspects of the data (in the within and differences dimensions), the maintained assumption of a perfectly competitive environment may become more and more dubious as firms face a variety of shorter run market imperfections. We'll come back to this last point below.

6. Other Approaches.

One early "solution" to the simultaneity problem was to assume profit maximization ex ante (and ex post) and use the observed factor shares as estimates of the relevant production function parameters. This was first suggested by Klein (1953) (see also the exposition in Chapter 4 of Nerlove (1965)). Solow (1957) used this assumption and varying factor shares to compute residual technical change without actually "estimating" (*i.e.*, econometrically) the possibly more general underlying production function. The difficulty with this suggestion is that it begs some of the questions one might want to ask of the data, especially about the extent of profit maximization, competition, and economies of scale. But it may prove useful when it is applied to only some of the parameters of interest. Thus, Griliches and Ringstad (1971) used it to "solve" the simultaneity problem in x, estimating effectively a "partial productivity" equation $y - s_v x = \alpha z + u$, where s_v, the share of variable inputs in total sales, was taken as a consistent estimator of β, x was a similarly weighted sum of all variable inputs (*e.g.*, labor and materials), and z was assumed to be predetermined and to provide information on the scale parameter.[26] Hall (1988) freed up the coefficient of s_v, assumed constant returns to scale, and estimated $y - z = m(s_v)(x - z) + u$, to test the competitive assumption and estimate the "mark-up" coefficient m using a number of demand side macro variables as instruments.[27] A similar approach is also followed by Klette (1994).

The obvious extension of the factor shares solution is to use factor prices w as instrumental variables to identify the parameters of interest. This is the essence of the "duality" approach to estimation. At the micro level, the only available factor prices are usually average earnings of workers and, possibly also, fuel prices. There are, however, a number of difficulties in using such price data as valid instruments. Why do wages differ across firms at a point of time and within firms over time? The first is likely to be related to unmeasured differences in the quality of labor, the second may fluctuate with unexpected shifts in the amount of overtime compensation. Even if there are valid regional wage differences and also aggregate time-movements in the real wage, the use of firm and time dummy variables would largely eliminate their contribution from the data, leaving us mostly with inappropriate or erroneous variation in these numbers. Moreover, the right prices in such an analysis are "expected" prices, rather than the realized and possibly endogenous ones.[28]

As suggested by Mundlak (1963, 1994), a related (but also not very workable) solution idea is to use the "whole" of equation (2) as an instrument, since it contains both that is the factor prices w and the errors in optimization v. If all of u is transmitted to the x choice (there are no e's or ϵ's in (1)), then (2) can be written (ignoring constants) as y-x=w+v and is now independent of u. If u and v were independent, which might be reasonable after a "within" transformation which eliminates the common firm effects, labor productivity (y-x) would be a valid instrument for x in (1). Also, if there were more than one x, the various input ratios $x_1 - x_j$ would also be valid instruments. But the moment we introduce only partial transmission (*i.e.*, non-negligible e's, and/or ϵ's), this is no longer legitimate. What may be possible is the use the lagged values of such terms (on the assumption that the currently untransmitted news e and errors of measurement ϵ are serially uncorrelated). Then we come back to the use of Chamberlain or GMM type estimators (*i.e.* the use of lagged values of x and z and also possibly y as instruments, which we have already discussed above).

An interesting new approach, which we shall present at some length, is advocated by Olley and Pakes (1992) in their paper on "Dynamics of Productivity in the Telecommunication Equipment Industry." This paper deals with two topics; selectivity and simultaneity, in an intertwined fashion. The sample selectivity problem may be quite serious for panel data. If observations (and data) are not missing at random, estimates that are based on "clean" and "balanced" sub-samples could be badly biased. For example, a bad draw of u may force a firm or plant to exit from the industry. Such a negative correlation between estimated productivity shocks and future probabilities of exit was observed by Griliches and Regev (1995) in their analysis of Israeli industrial firms. They called it "the shadow of death." If the impact of negative u's on exit is stronger for smaller firms (the larger ones having more resources to survive them), then this will induce a negative correlation between u and z (the stock of capital) among the surviving firms and bias the estimated capital coefficient downward in such samples. To simplify our exposition, we deal with simultaneity first, but we will come back to selectivity shortly.

The major innovation of Olley and Pakes is to bring in a new equation, the investment equation, as a proxy for a, the unobserved transmitted component of u.[29] Trying to proxy for the unobserved a (if it can be done right) has several advantages over the usual within estimators (or the more general Chamberlain and GMM type estimators): it does not assume that a reduces to a "fixed" (over time) firm effect; it leaves more identifying variance in x and z and hence is a less costly solution to the omitted variable and/or simultaneity problem; and it should also be substantively more informative.

In our notation, their argument goes roughly as follows: the investment demand of the firm at time t can be written as

$$(13) \qquad i_t = i_t(a_t, z_t)$$

where the function $i_t(\,)$ is allowed to change over time. Assuming (13) can be inverted, it can be rewritten as

$$(14) \qquad a_t = h_t(i_t, z_t)$$

where the potentially observable function h_t is now a perfect proxy for the unobserved a_t. Substituting h_t for a_t into (5) and (1), and ignoring the measurement errors ϵ, leads to

$$(15) \qquad y_t = \beta x_t + \phi_t(i_t, z_t) + e_t$$

where, the function $\phi_t = \alpha z_t + h_t(i_t, z_t)$, "contains" the unobserved a_t variable. Since both i and z are observed, if the form of h or ϕ were known, β could be estimated consistently by OLS. Note, however, that α is not identified at this point since z_t is also in h_t. In practice Olley and Pakes approximate ϕ_t by a third or fourth order polynomial in i and z to estimate β. Using the estimated β one can now proceed to estimate α. If z_t and a_t were uncorrelated, one would just use the partial productivity equation

$$(16) \qquad y_t - \beta x_t = \alpha z_t + a_t + e_t$$

More reasonably, if one assumes that z_t is uncorrelated only with the change in a_t, $\zeta_t = a_t - a_{t-1}$ (i.e., a_t is a random walk), one can estimate α from

$$(17) \qquad y_t - \beta x_t - \phi_{t-1} = \alpha(z_t - z_{t-1}) + \zeta_t + e_t$$

where ϕ is the estimated polynomial in i and z, and $(\phi_{t-1} - \alpha z_{t-1})$ is proxying for a_{t-1}.[30] Returning to the selective issue, Olley and Pakes add to this system a probability of survival function which also depends on a_{t-1} and $\bar{a}_{t-1}$, the unobserved critical level of a which would cause a firm to exit, and is again inverted to yield an estimate of $\bar{a}_t$, which is a function of both the estimated probability of survival and a_{t-1}. Equation (17) then becomes

$$(18) \qquad y_t - \beta x_t = \alpha z_t + g_t(\phi_{t-1} - \alpha z_{t-1}, \lambda_t) + \zeta_t + e_t$$

where the assumption is made that $E(a_t | a_{t-1}$ and survival$) = g_t(a_{t-1}, \bar{a}_t)$, where $\bar{a}_t$, is proxied by λ_t, the probability of survival from the selection equation which is estimated non-parametrically. The function g_t is approximated by a general polynomial (of order 3) in λ_t and $(\phi_{t-1} - \alpha z_{t-1})$ and α is estimated non-linearly across all the terms that contain it.[31]

The main Olley-Pakes results are summarized in Table 2 (taken from their Table 6), while Table 3 presents an application of a simplified version of their model to data on U.S. R&D performing firms listed on the major stock exchanges.[32] The focus in Table 3 is on estimating the coefficients of two capitals: "physical" and "R&D." The data and the approach differ somewhat from Olley and Pakes because we used data separated by five years, rather than adjacent cross-sections and used only quadratic and crossproduct terms in the polynomial approximations and parametric assumptions (normality) for the selection equations.

The two sets of results tell largely the same story, though the selectivity problem is significantly more severe in the Olley-Pakes plant data. The magnitude of this problem can be seen, in Table 2, in the decline in the estimated capital elasticity α as one moves from the unbalanced sample (column 3) to the balanced one (column 1). Possibly, because exit is often a success for our R&D firms (being taken over) rather than a failure, the selection problem is less severe, in Table 3. Once one shifts to the unbalanced samples, the remaining selectivity (mainly attrition) does not appear to be too important (compare columns 5 and 6 in both tables).

As far as the simultaneity problem is concerned, either it is of no great import in these data or the introduction of investment and the associated Olley-Pakes procedure does not fully adjust for it. Looking at Tables 2 and 3, investment is highly "significant" in the production function, but at the end of the procedure (having allowed for selectivity and unbalance), the coefficients change only little (compare columns 3 and 5 in both tables).[33] They do more so in the Olley-Pakes data, where the labor coefficient is reduced by 12 percent and the capital coefficient is increased by about 12 percent, which goes in the "right" direction. In our own data, the changes are smaller and go slightly in the "wrong" direction (labor "up" and the sum of the "capital" coefficients down), though probably not significantly so.[34]

The Olley Pakes solution to the simultaneity problem is a clever way to exploit the fact that the unobserved "productivity shocks" a_t are transmitted to more than just one equation and should be estimated within a system of behavioral equations. It does rest, however, on two very strong assumptions: 1. that there is only one single-component unobservable in the system, the a_{it}, which follows a first order Markov process and is fully transmitted to the investment equation, and 2. that no other variables or errors appear in it. Macro variables are taken into account implicitly by letting the g function change over time. Investment will, however, depend also on individual factors such as interest rate expectations, tax treatments, and changes in future demand prospects not yet captured in the state variable z_t. In principle, there may be additional instrumental variables and other indicators of a_t, such as R&D, which could help to solve the errors in the investment equation problem, except for the extreme non-linearities introduced by the Olley-Pakes semi-parametric approach.[35]

Even more problematic is the assumption that i_t depends on the whole of a. If a_{it} consists of both fixed and variable "omitted" components, i_t would not depend on the former if z_t had already fully adjusted to them. In this multifactor interpretation of a_{it}, net investment of i_t would depend primarily on the "news" in a_{it} and not on its level.[36] Hence, it may not be able to proxy well for the various other (more permanent) misspecifications contained in a_{it} and solve the bias problems created by their omission. While being right on its own terms and making a significant contribution to the treatment of sample selectivity and simultaneity problems, the Olley-Pakes model does not nest within it the standard fixed effects model. A single first order Markov unobservable cannot really capture a mixture of fixed and variable effects. This might also explain why the "within" results which "do more" are so different, leaving us wiser, but with much of the original puzzle (the effect of unaccounted for heterogeneity) still intact.

7. Other Problems.

Besides simultaneity and selectivity there are a number of other "garden variety" problems with such data: lack of information on quality dimensions of labor and capital inputs; lack of utilization variables; and so forth. A major difficulty has been, however, largely ignored in this literature: the fact that we usually do not have "quantity" measures of output but rather just nominal sales (or value added) which are deflated by a common price deflator.[37] The assumption is made that "the law of one price" holds: that all firms in an industry charge the same price and that all have prices move in unison over time. That is wrong if different firms produce different varieties of the product and operate within somewhat different, though possibly interconnected, markets. Examples would be cement plants which operate in different geographic markets, and plants in the "computer equipment" industry, some of which produce mainframes, while others produce PCs or disc drives. The use of deflated revenue as a measure of "output" may have serious consequences for the interpretation of the resulting estimates.

This can be illustrated using the simple model in Klette and Griliches (1994): To the production function (1) add a demand function for the products of the *i*-th firm:

$$y_i = \eta\,(p_i - p_I) + \upsilon p_I + d \qquad (19)$$

where p_i is the firm's own price (or price index), η is demand elasticity with respect to the relative price of its own products, p_I is the aggregate industry price index (relative to the overall economy price level), υ is its elasticity, *d* are all other demand shifters for the products of this industry. If the variable that we observe is not "real" output *y*, but deflated revenue

$$r_i = (y_i + p_i) - p_I \qquad (20)$$

then, the "revenue production" function is

$$(21) \qquad r = \alpha z + \beta x + u + (p_i - p_I)$$

There would be no problem here if the p_i's were random and exogenous. But if firms have a modicum of market power, at least in the short run, p_i will be set by them and will be correlated with u, x, and z. Using (19) and (20) we can solve out for p_i and write the pseudo production function as

$$(21') \qquad r = [\alpha z + \beta x + u]/m - (d + \upsilon p_I)/\eta$$

where $m = \eta/(1 + \eta)$ is the "mark-up" coefficient and is likely to be larger than one. Since d and p_I are aggregates, they can be "controlled" for by the introduction of period dummy variables. It is clear, from (21'), that the estimates of α and β will be biased downward on the order of 1/m, implying diminishing returns to scale in contexts where there actually may be increasing returns. The size of the bias (the size of the markup) may depend on the time dimension of the data, being lower in the longer run between dimension, since entry is more likely; and higher in the shorter run within dimension. This may provide another possible explanation for the facts reported in Figure 1.[38]

Klette and Griliches show that under certain, not unreasonable conditions, one can substitute a trend variable for the period dummies and use aggregate output y_A as a proxy for the demand shifters $(d + \upsilon P_I)$, which will identify the function as a whole. An illustration of their analysis is given in Table 4, based on data for plants manufacturing transportation equipment in Norway (N=403, T=6). In their study all the variable inputs are aggregated into one x (using factor shares as weights). When a production function is estimated by OLS on first differences, column 1, it yields a non-significant capital coefficient (which is a standard result for such estimates) and a statistically significant estimate of **decreasing** returns to scale. Using the more predetermined employment change as an instrument for x does not change matters much, but when the aggregate demand shifter y_a is added to this equation, it is "significant" and provides a reinterpretation of these results, implying a "reasonable" elasticity of demand, $\eta = -7.2$, and an associate estimate of mildly **increasing** returns, 1.06, rather than what appeared to be decreasing returns before.[39]

While the solution offered by Klette and Griliches may be problematic, they have raised an important warning flag. Most micro-studies do not have detailed price data at the relevant level. To the extent that these missing prices are endogenous, the resulting estimates are likely to be biased, often badly so, as estimates of the desired "pure" production function parameters.

This reminds us that at the micro level, the firms or plants we analyze differ a great deal, even within what one might think of as a relatively well defined "industry," *i.e.* cement or bakeries. They differ in their geographic location and in their potential for exercising market power; they differ in the particular assortment of products that they may produce; and because of the preceding considerations, they may also differ in the inputs and technologies that they use to produce them.

Using data similar to that used in Figure 1 and Table 1, Mairesse and Griliches (1990) estimate separate α coefficients for each firm based on 13 years of data, assuming constant returns to scale. They find a great deal of heterogeneity in their estimates. Much more than could be explained by sampling error and/or reasonable dispersion of α at the individual level. They conclude, not too surprisingly, that the simple production function must be seriously misspecified. The variegated world of plants and firms cannot really be summarized adequately by looking only at their capital-labor ratios and size. It is much more complex than that. Unfortunately, standard census type data do not provide enough additional information or relevant product-plant characteristics to allow one to pursue a substantive analysis of such differences much further.[40]

8. Coda as Prologue.

We have surveyed one strand of research in this area, the attempt to solve the "simultaneity" problem in estimating production functions from micro data, perhaps in excessive length (but still leaving large areas uncovered). The main message has been, that in trying to evade this problem, researchers have shifted their focus to thinner and thinner slices of data, leading to the exacerbation of other problems and misspecifications. Much of this line of work has been guided, unfortunately, by what "econometrics" as a technology might be able to do for this problem, rather than focusing on the more important but technically less tractable problems of data quality and model specification.

Instead of throwing away more of the data as "contaminated," the future, we believe, is in finding circumstances and data which will provide credible identification.[41] What makes firms invest differently, hire more or less labor, or expand their R&D program? Is it possible to look for periods of unexpected tax changes which might impact firms differentially, depending on their previous history and financial condition? For that, one needs to match detailed financial data to the standard census-type data sources. One could also use data on capital subsidies in various countries (*e.g.* Israel) to construct differential cost-of-capital variables. Data on stock market values may convey information on expectations about future demand conditions and R&D opportunities, independent of the current productivity disturbances. The challenge is to find (instrumental) variables that have genuine information about factors which affect firms differentially as they choose their input levels, their "location on the isoquant."

Besides better and more data, we need also a richer theoretical framework to help us understand why firms are different, not only in their capital-labor ratios, but also in the product mix that they produce, the quality of their workforce, the technologies they use, their organizational structures, the markets that they serve. We started our work in this area with the hope that micro-data may be the answer to the various difficulties encountered at the aggregate level, primarily because this is the level which our theories claim to comprehend, and because we believed that this will reduce multicollinearity and provide us with more identifying variance. We also thought that one could reduce aggregation biases by reducing the heterogeneity as one goes down from such general mixtures as "total manufacturing" to something more coherent, such as "petroleum refining" or the "manufacture of cement." But something like Mandelbrot's fractals phenomenon seems to be at work here also: the observed variability-heterogeneity does not really decline as we cut our data finer and finer. There is a sense in which different bakeries are just as much different from each other, as the steel industry is from the machinery industry.

This paradox arises, in part, from the fact that our theories, while denominated in micro language of the firm or plant, have really been designed with macro questions in mind. They deal with reasonable crude aggregates: output, labor, capital which turn out to be rather vague concepts when we go down to the micro level and have to face the large number of products, labor types, machines, and technologies. We have neither the data or a convenient language to describe all this variability effectively.[42]

In spite of all these reservations, the production function framework that we have been analyzing over the last two decades has its uses. It is a major tool for asking questions about rates of technological change, about economies of scale, rates of returns to R&D, the relevant productive qualities of labor, and more, whose answers may not be all that sensitive to the biases discussed here. Moreover, we are unlikely to give this framework up just because it is imperfect. But to make further progress we need to infuse it with new data and appropriate theoretical and econometric models for dealing with the real heterogeneity that is the hallmark of the world we live in.

Endnotes

1. Quoted in O. Bjerkholt (1995).

2. We shall ignore a large number of other interesting and related topics: alternative functional forms; frontier production functions; non-parametric measures of productivity; estimation of cost functions and factor demand structures and much more. See Walters (1963), Nerlove (1968), Jorgenson (1986), and Mundlak (1994) for reviews of some of these topics.

3. In the micro language of subsequent expositions, if all firms are on the same production function and face the same prices, they would have the same input ratios and there would be no relevant variability on which to base ones estimates of the production function.

4. His other major criticisms were: (1) The capital variable should be capital used rather than capital in place (this issue of "utilization" is still with us); (2) Constant returns should be tested for, not imposed; and (3), technical change should not be ignored. His was only one voice in a chorus of criticisms. The work of Douglas was also criticized on a number of other grounds; among them disbelief in the existence and/or relevance of an aggregate production function (see, *e.g.*, the discussion about intra and inter firm industry production functions in Reder (1943)); and the importance of left out variables: trends, technology levels, and capital utilization (see, e.g., Schultz (1929), Durand (1937), and Smith (1940) and (1945)). Douglas (1948) acknowledges most of these criticisms but does not mention simultaneity or the Marschak and Andrews paper.

5. This passage is preceded by another similar and equally eloquent passage: "...the economist cannot perform experiments. That is, he cannot choose one variable as "dependent," and, while keeping the other, "independent," ones under control, ... watch the values taken by the dependent, *i.e.*, uncontrolled variable. The economist has no independent variables at his disposal because he has to take the values of all the variables as they come, produced by a mechanism outside his control. This mechanism is expressed by a system of simultaneous equations, as many of them as there are variables. The experimenter can isolate one such equation, substituting his own actions for all the other equations. The economist cannot." *ibid*, p. 143.

6. Here, and in what follows, we ignore constant terms unless they become important for the analysis. In Marshak's and Andrews ' own words u (in their notation ϵ, the deviations in the constant term parameter A_0 of the production function) "will depend on technical knowledge, the will, effort and luck of a given entrepreneur in a given year, as can be summarized in the word 'technical efficiency.'"

7. (2) can be rewritten as $x = \log(\beta) + y - (w + v)$ which is the demand function for x in the

Cobb-Douglas case. If one allowed free coefficients on *y* and *w*, one would be in the non CRS CES case.

8. One can also do a sensitivity analysis and convince oneself that such biases are relatively unimportant in one's own data (*e.g.*, Griliches and Ringstad (1971)).

9. Factor demand functions, derived from the cost function framework, interchange the roles of *x* and *y*. To estimate them by least squares or related methods requires the assumption that *y* is exogenous, uncorrelated with *v*, which contradicts equation (4), and that *y* is not subject to measurement error ($\epsilon_y = 0$). Moreover, the expected real factor prices *w* must also be measured correctly by the econometrician. In short, life is not any easier on the "other side."

10. This distinction is similar to the distinction made by Frisch (1938) between *aberrations* and *simuli*.

11. Another possible response, which we will not discuss here explicitly, is to try to reduce the role and variance of *u*, and thus lower the remaining biases, by introducing additional variables into the equation such as R&D stocks and labor quality measures (e.g. Griliches (1964) and Griliches and Mairesse (1984)).

12. The last two components represent individual deviations around the point of "average" behavior approximated by the fitted function. To the extent that they result from changes in factor and product prices, both anticipated and unanticipated, this may preclude the use of prices and demand side variables as valid instrumental variables.

13. To simplify the exposition, 2 was treated as strictly exogenous in deriving equations (3) and (4). If *z* is also correlated with *a* than these equations hold only after some appropriate transformation and/or an additional equation should be appended for the now endogenous *z*.

14. To the extent that these deviations from the standard assumptions are permanent, they would imply a non-zero expectation for *v* and vitiate the attempt to estimate β from the factor share or the more extended factor demand versions of this equation arising from more general translog and other types of cost functions.

15. This is the case where parameter estimates based on first differences (FD) would be consistent though the "within" estimates are not.

16. Going "within" is not the only way of eliminating "fixed" effects. Differencing over a single period or more also accomplishes this purpose. We shall return to this point below.

17. Similar results and a similar interpretation appear also in the original Mundlak (1961) study. See also Griliches (1957).

18. The idea of individual trend is not new and keeps reappearing in various forms. An early statement is Mundlak (1970).

19. If the original fixed effects model was correct, first difference estimators are consistent but not BLUE. This is the basis for the specification tests suggested in Griliches and Hausmann (1986). It can be shown that "within" estimator is an optimal combination of all possible difference estimators.

20. Instrumental variables based estimators (GMM) would reduce the available net variance even further but might also reduce some of the other biases.

21. The story is similar for the estimates of β, as can be seen from the plotted *m* in Figure 1.

22. For repeated restatements of this point see Griliches (1974), Mairesse (1978), and Griliches and Hausman (1986).

23. Obviously, more efficient GMM based estimates could be computed if one believed this model.

24. To simplify matters, let us consider the case where investment is a random walk and has an assumed life of six years. But let the weights of past *I*'s in the true *K* be (0, 1, 2, 2, 1, 0), instead of all ones. The coefficient of the first differenced incorrect capital measure will be exactly zero then. If we try the difference lagged one period, its coefficient will be only one-quarter of the appropriate size, assuming that *I* is serially uncorrelated and has a constant variance. It is not that capital does not matter. It is that our timing is wrong and it gets worse as we do various transformations.

25. If, for example, one constructed the capital stock assuming a declining balance (geometric) depreciation pattern while the truth were closer to non-negligible gestation lags and an inverted-U shape, with rising weights for a while, à la Pakes and Griliches (1984).

26. See Appendix C in Griliches and Ringstad (1971) for an exposition of this point.

27. Equation (2) can be rewritten as $\ln(s_x) = \ln(\beta) - v$, where $s_x = (WX/PY)$ is the share of *X* in total revenue, and gives an estimate of β. The question of what average and transformation to use for this estimate depends on the assumed distribution of *v* (is $Ev=0$?) and whether maximization is expected to be correct on the median, the arithmetic, or the geometric average. See Goldberger (1968) and Griliches-Ringstad, p. 73, for additional discussion of this set of issues.

28. Abbot et al (1988) and Eden and Griliches (1993) raise questions about the validity of Hall's instruments. If u contains short run specification errors such as capital utilization and labor "effort" measures, which are related to expectational errors made by the decision makers when choosing their "desired" capital and employment levels, then unanticipated macro shifts may cause such errors and cannot serve as valid instruments in this context.

29. This criticism applies also to the literature on cost function and factor demand systems estimation which leans rather heavily on the correctness and exogeneity of the price data used in their analyses.

30. This simplification (that a_t is a random walk) is made by us for expositional purposes. If it were an AR(1) process, there would be a ρ coefficient in front of the term $\phi_{t-1} - \alpha z_{t-1}$. More generally, Olley and Pakes assume only that a_t is an exogenous process and approximate it by $\hat{\alpha}_t + \zeta_t$, where predicted is $\hat{\alpha}_t$ a polynomial in $(\hat{\alpha}_{t-1} - \alpha z_{t-1})$, and estimate α non-linearly across the various terms.

31. In their notation a is ω and they refer to it simply as "productivity."

32. The underlying data set is described in Hall (1990). For a discussion of the concept of R&D "capital" see Griliches (1995).

33. Olley and Pakes present also results for a non-parametric kernel estimator which are essentially the same.

34. The standard errors in Table 3 are too small because they do not allow for the fact that β and ϕ are estimated in the previous stage. For the correct way of deriving standard errors in such contexts, see Pakes and Olley (1995).

35. The current state of estimating non-linear errors-in-variables models is not completely hopeless, but it is not easy either.

36. In the old fashioned accelerator model language, i_t would be a function of the change in y_t rather than of its level.

37. Though Marschak and Andrews were well aware of that. See also Griliches and Mairesse (1984) and Hall and Mairesse (1995).

38. Also, there is a countervailing upward bias in β if u is transmitted to the x decision. It can be shown (see Klette and Griliches) that the estimated β is bounded by β/m and 1.

39. Klette and Griliches get similar results for the other three industries they examined. They also present parallel cost function computations, showing that these are similarly affected by the use of the wrong output measure.

40. Mundlak (1988) discusses a framework which has α's as functions of other variables. In his work these coefficient shifters are largely macro variables which does not really fit our micro data world very well.

41. The search for "natural experiments" has become important in labor economics but has not caught on yet in the field of productivity analysis. For examples of the former see Angrist (1990) and Angrist et al (1995).

42. For a brief moment in time, we looked at activity analysis and "engineering production functions" as a possible solution, but the computer resources of the time and the overwhelming data requirements drove this approach and its practitioners out of the field. Today, if they still exist, they can be found in business and engineering schools, not in economics departments. We have given up trying to know something substantive about specific technologies in particular fields.

References

Abbott, T. A., Griliches, Z., and Hausman, J. A., "Short Run Movements in Productivity: Market Power versus Capacity Utilization," paper presented at NBER Summer Institute, 1988.

Arellano, M., and Bond, S., "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58, 1991, 277-297.

Arrow, K. J., Chenery, H. B., Minhas, B. S., and Solow, R. M., "Capital-Labor Substitution and Economic Efficiency," *Review of Economics and Statistics*, 43(3), 1961, 225-250.

Angrist, J. "Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records," *American Economic Review*, 80, 1990, 313-35.

Angrist, J., Imbens, G. W., and Rubin, D. B., "Identification of causal effects using instrumented variables," *Journal of the American Statistical Association*, 1995, forthcoming.

Bjerkholt, O., "Ragnar Frisch: The originator of econometrics," in *Foundations of Econometrics: The Selected Essays of Ragnar Frisch*, Elgar Publishing Co., in press 1995.

Chamberlain, G., "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, 18(1), 1982, 5-46.

Christensen, L. R., Jorgenson, D.W., and Lau, L. J., "Transcendental Logarithmic Production Frontiers," *Review of Economics and Statistics*, 55(1), 1973, 28-45.

Cobb, C. W., and Douglas, P. H., "A Theory of Production," *American Economic Review*, 18 (1), 1928, Supplement, 139-72.

Crepon, B., and Mairesse, J., "Chamberlain and GMM estimates: an overview and some simulation experiments," INSEE, Paris, Unpublished manuscript, 1995.

Douglas, P. H., "Are there laws of production?" *American Economic Review*, 38(1), 1948, 1-41.

Durand, D., "Some thoughts on marginal productivity, with special reference to Professor Douglas' analysis," *Journal of Political Economy*, 45, 1937, 740-758.

Eden, B., and Griliches, Z., "Productivity, Market Power and Capacity Utilization when Spot Markets Are Complete," *American Economic Review, Papers and Proceedings*, 83, 1993, 219-223.

Frisch, R., "Circulation Planning," *Econometrica*, 2(3), 1934, 258-336.

Frisch, R., "Statistical versus theoretical relations in economic microdynamics," unpublished memorandum, 1938. Reprinted in the *Foundation of Econometrics: Selected Essays of Ragnar Frisch*, forthcoming.

Goldberger, A. S., "The interpretation and estimation of Cobb-Douglas Functions," *Econometrica*, 36, 1968, 464-472.

Griliches, Z., "Specification Bias in Estimates of Production Functions," *Journal of Farm Economics*, 1957, 39(1), 8-20.

Griliches, Z., "Research Expenditures, Education and the Aggregate Agricultural Production Function," *American Economic Review* , 1964, LIV(6), 961-974.

Griliches, Z., "Errors in Variables and Other Unobservables," *Econometrica*, 1974, 42(6), 971-998.

Griliches, Z., "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica*, 45(1), 1977, 1-22.

Griliches, Z., "Returns to Research and Development Expenditures in the Private Sector," in J. W. Kendrick and B. Vaccara, ed., *New Developments in Productivity Measurement*, vol. 44 of *Studies in Income and Wealth*, Chicago: Chicago University Press, 1980, 419-454.

Griliches, Z., "Data Issues in Econometrics," in Z. Griliches and M. Intriligator, ed., *Handbook of Econometrics*, Amsterdam: North Holland Pub. Co., 1986, 1466-1514.

Griliches, Z., "R&D Productivity: Econometric Results and Measurement Issues," in Paul Stoneman, ed., *Handbook of the Economics of Innovation and Technological Change*, Blackwell, Oxford, 1995, forthcoming.

Griliches Z. and Hausman, J., "Errors in Variables in Panel Data," *Journal of Econometrics*, 1986, 31, 93-118.

Griliches, Z. and Mairesse, J., "Productivity and R&D at the Firm Level," in Zvi Griliches, ed., *R&D, Patents and Productivity*, Chicago: University of Chicago Press, 1984, 339-374.

Griliches, Z. and Regev, H., "Firm Productivity in Israeli Industry, 1979-1988," *Journal of Econometrics*, 65(1), 1995, 175-204.

Griliches, Z. and Ringstad, V., *Economies of Scale and the Form of the Production Function*, Amsterdam: North Holland, 1971.

Hall, B. H., "The Manufacturing Master File 1959-87," NBER WP 3366, 1990.

Hall, B. H. and Mairesse, J., "Exploring the relationship between R&D and productivity in French manufacturing firms," *Journal of Econometrics*, 65(1), 1995. 263-294.

Hall, R. E., "The relation between price and marginal cost in U. S. industry," *Journal of Political Economy*, 96, 1988, 921-947.

Heady, E. O. and Dillon, J. L., *Agricultural Production Functions*, Aces, Iowa: Iowa State University Press, 1961.

Hellerstein, J. K. and Neumark, D., "Sex, wages and productivity: an empirical analysis of Israeli firm-level data," *Fall Institute for Economic Research in Israel, Discussion Paper 93.01*, Jerusalem, 1993.

Hoch, I., "Estimation of production function parameters and testing for efficiency," *Econometrica*, 23(3), 1955, 325-326.

Hoch, I., "Estimation of Production Function Parameters Combining Time-Series and Cross-Section Data," *Econometrica*, 30(1), 1962, 34-53.

Jorgenson, D. W., "Econometric Methods for Modelling Producer Behavior," in Z. Griliches and M. Intriligator, ed., *Handbook of Econometrics*, vol. 3, Amsterdam: North Holland, 1986, 1841-1915.

Jorgenson, D. W., and Lau, L. J., "The Duality of Technology and Economic Behavior," *Review of Economic Studies*, 41(2), 1974, 181-200.

Keane, M.D., and Runkle, D. E., "On the Estimation of Panel-Data Models with Serial Correlation when Instruments are not Strictly Exogeneous," *Journal of Business and Economic Statistics*, 10, 1992, 1-14.

Klein, L. R., *A Textbook of Econometrics*, Row, Peterson and Co., 1953.

Klette, T. J., "Estimating Price-Cost Margins and Scale Economies from a Panel of Microdata," Statistics Norway Research Department, Discussion Papers, 130, 1994.

Klette, T. J., and Griliches, Z., "The Inconsistency of Common Scale Estimators when Output Prices are Unobserved and Endogenous," Statistics Norway Discussion Papers, No. 127, 1994.

Krishna, K. L., "Production relations in manufacturing plants: an exploratory study," University of Chicago: unpublished Ph. D. dissertation, 1967.

Mairesse, J., "Comparison of Production Function Estimates on the French and Norwegian Census of Manufacturing Industries," in F. L. Altmann et al eds., *On the Measurement of Factor Productivities*, Gottingen: Vendenbeck & Ruprecht, 1975, 434-462.

Mairesse, J., "New estimates of embodied and disembodied technical progress," *Annales de L'INSEE*, 30-31, Special issue on *The Econometrics of Panel Data*, 1978, 681-719.

Mairesse, J., "Time-series and cross-sectional estimates on panel data: Why are they different and why should they be equal?" in J. Hertog et al eds., *Panel Data and Labor Market Studies*, Amsterdam: North-Holland Publishing Co., 1990, 81-95.

Mairesse, J., and Griliches, Z., "Heterogeneity in Panel Data: Are There Stable Production Functions?" in P. Champsaur et al eds., *Essays in Honour of Edmond Malinvaud*, 3, Empirical Economics, MIT Press, 1990, 192-231.

Mairesse, J. and Hall, B. H., "Comparing the Productivity of R&D in French and U.S. Manufacturing Firms," Paper given at the Conference on International Productivity Differences and their Explanations, Wissenschaftszentrum, Berlin, November 26-7, 1993.

Marschak, J., and Andrews, W., "Random Simultaneous Equations and the Theory of Production," *Econometrica*, 12, 1944, 143-205.

Mendershausen, H., "On the Significance of Professor Douglas' Production Function," *Econometrica*, 6(2), 1938, 143-153.

Mundlak, Y., "Empirical Production Function Free of Management Bias," *Journal of Farm Economics*, 1961, *43*, 44-56.

Mundlak, Y., "Estimation of Production and Behavioral Functions from a Combination of

Cross-section and Time-series Data," in Christ et al eds., *Measurement in Economics*, Stanford: Stanford University Press, 1963, 138-166.

Mundlak, Y., "Endogenous technology and the measurement of productivity," in *Agricultural Productivity: Measurement and Explanation*, S. M. Capalbo and J. M. Antle eds., Washington, D.C.: Resources for the Future, 1988, 316-331.

Mundlak, Y.,"Empirical production functions with a variable firm effect," Center for Agricultural Economic Research Working Paper 7005, Rehovoth, 1970.

Mundlak, Y., "On the (in)efficiency of empirical dual production functions and related issues," University of Chicago, unpublished paper, 1994.

Mundlak, Y., and Hoch, I., "Consequences of Alternative Specifications of Cobb-Douglas Production Functions," *Econometrica*, 33(4), 1965, 814-28.

Nerlove, M., "Returns to Scale in Electricity Supply," in Christ et al eds., *Measurement in Economics*, Stanford: Stanford University Press, 1963, 167-200.

Nerlove, M., *Estimation and Identification of Cobb-Douglas Production Functions*, Chicago: Rand McNally & Co., 1968.

Olley, S. and Pakes, A., "The Dynamics of Productivity in the Telecommunications Equipment Industry," in *NBER, Working Paper n. 3977*, 1992.

Pakes, A. and Griliches, Z., "Estimating Distributed Lags in Short Panels with an Application to the Specification of Depreciation Patterns and Capital Stock Constructs," *Review of Economic Studies*, 1984b, LI(2)165, 243-262.

Pakes, A. and Olley, S., "A limit theorem for smooth class of semiparametric estimators," *Journal of Econometrics*, 65(1), 1995, 295-332.

Reder, M. W., "An Alternative Interpretation of the Cobb-Douglas Function," *Econometrica*, 11(3 & 4), 1943, 259-264.

Ringstad, V., *Estimating Production Functions and Technical Change from Micro Data*, Central Bureau of Statistics of Norway, Economic Studies No. 21, Oslo, 1971.

Schultz, H., "Marginal Productivity and the General Pricing Process," *Journal of Political Economy*, 37(5), 1929, 505-551.

Smith, V. E., "An Application and Critique of Certain Methods for the Determination of a Statistical Production Function for the Canadian Automobile Industry, 1917-1930," unpublished Ph.D. dissertation, Department of Economics, Northwestern University, Evanston, IL., 1940.

Smith, V. E., "The statistical production function," *Quarterly Journal of Economics*, 59, 1945, 543-562.

Solow, R. M., "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics*, 1957, 39(3), 312-320.

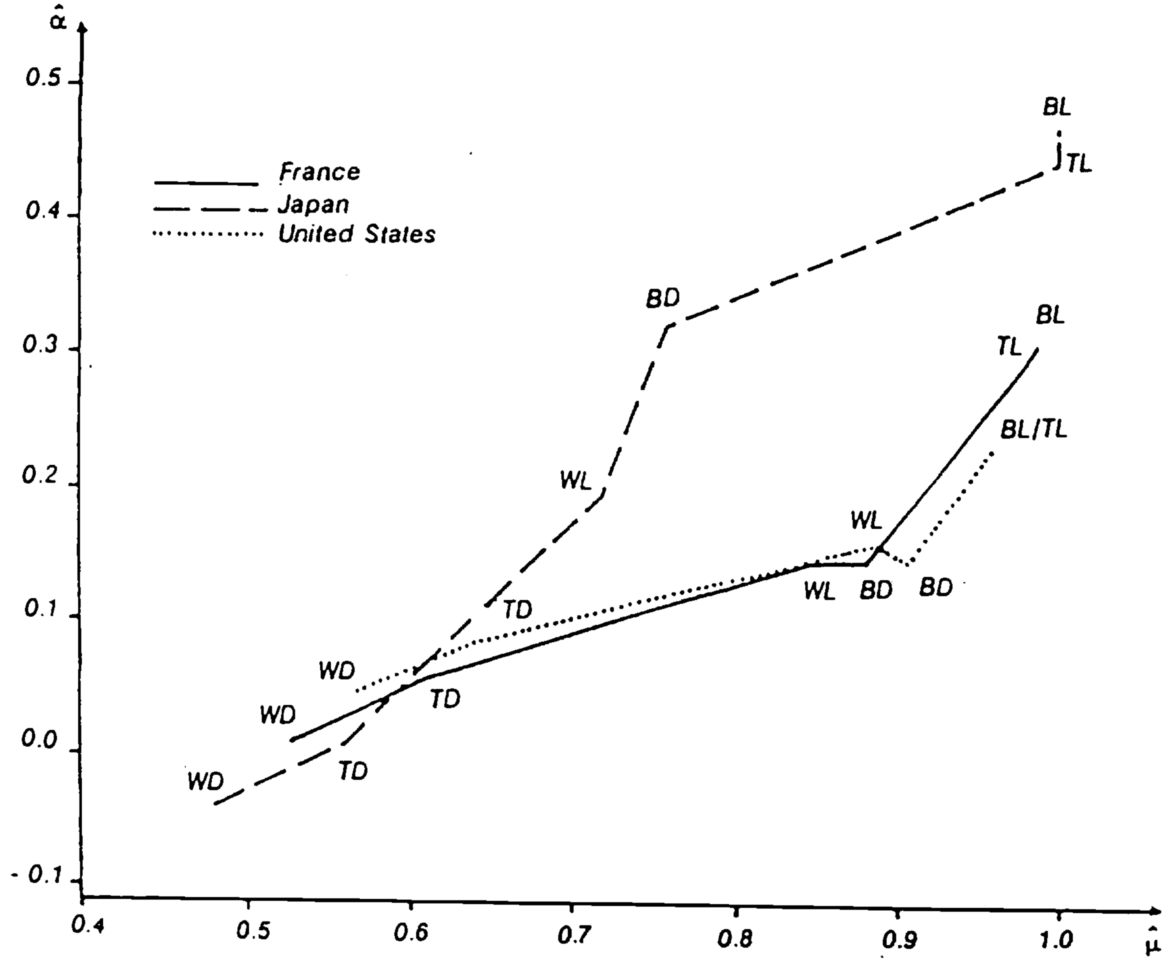
Tintner, G., "A Note on the Derivation of Production Functions from Farm Records," *Econometrica*, 12(1), 1944, 26-34.

Walters, A.A., "Production and cost functions: an econometric survey," *Econometrica*, 31(1-2), 1963, 1-66.

Zellner, A., Kmenta, J., and Dreze, J., "Specification and Estimation or Cobb-Douglas Production Function Models," *Econometrica*, 34(4), 1966, 784-95.

FIGURE 1

ALTERNATIVE ESTIMATES OF THE ELASTICITY OF CAPITAL $\hat{\alpha}$
AND THE ELASTICITY OF SCALE $\hat{\mu}$
FOR THE FRENCH, JAPANESE AND AMERICAN SAMPLES



From the upper left to the bottom right of the graph, the estimates are the following :

BL. Between in levels : $(y_{i\,.})$
TL. Total in levels : (y_{it})
BD. Between in differences : $(\Delta y_{i\,.})$
WL. Within in levels : $(y_{it} - y_{i\,.})$
TD. Total in differences : (Δy_{it})
WD. Within in differences : $(\Delta y_{it} - \Delta y_{i\,.})$

Source: Mairesse (1990)

Table 1: Alternative Estimates of α and Associated Data:
French Manufacturing Firms, 1967-79, N=441

Estimation Method	Estimated α	t_α	Net Variance of Capital	Measurement Error Bias Multipliers[1]	
	(1)	(2)	(3)	(4)	5=(4)/(3)
Total Levels	.304	34.8	.3023	1.	3.31
Between Levels	.315	10.1	.2851	.077	.27
Within Levels	.146	12.	.0245	.923	37.7
Total Differences	.061	3.4	.0085	2.	236.1
Long Differences[2]	.147	3.7	.1282	2.	15.6
Within Differences	.010	.5	.0074	1.986	267.7
GMM[3]	.052	.5	.0003	-	-

(1) Under the assumption of random measurement error in capital ϵ, the probability limit of the estimated α is given by

$$\text{plim}\ \hat{\alpha} = \alpha - \alpha m \sigma^2_\epsilon / \sigma^2_{zx}$$

where m is given in column (4), and σ^2_{zx} in column (3), and their ratio in column (5).

(2) This line is in terms of the longest difference (1979-67). For a parallel "between" differences calculation, the entries in column (3) and (4) should be both divided by 144 (12x12)..

(3) The GMM estimates are computed treating the first differences in labor and capital both as endogenous and using their two period lagged levels (and lagged output) as instrumental variables. The "net" variance of capital is computed implicitly from the final standard errors.

Table 2: Alternative Estimates of Production Function Parameters:
U.S. Telecommunications Equipment Plants, 1974 to 1986
(standard errors in parentheses)

Variables[1]	Sample					
	Balanced Panel		Full Sample[3]			
	(1)	(2)	(3)	(4)	(5)	(6)
	Total	Within	Total OLS		Nonparametric F	
Labor	.851 (.039)	.728 (.049)	.693 (.019)	.628 (.020)	.608 (.027)	
Physical Capital	.173 (.034)	.067 (.049)	.304 (.018)	.219 (.018)	.339 (.030)	.342 (.035)
Age	.002 (.003)	-.006 (.016)	-.0046 (.0026)	-.001 (.002)	.000 (.004)	-.001 (.004)
Time	.024 (.006)	.042 (.017)	.016 (.004)	.012 (.004)	.011 (.010)	.044 (.019)
Investment	-	-	-	.130 (.010)	-	
Other Variables	-	-	-	-	Powers of h	Polynomial in P and h
# Observations[2]	896		2592		1758	

(1) The dependent variable in columns 1 to 4 is the log value added, while in column 5 and 6, the dependent variable is the log(value added) - β*log(labor).

(2) The number of observations in the balanced panel for regressions in columns 1 and 2 are the observations for those plants that have continuous data over the period, with zero investment observations removed. Similarly, the 2592 observations in columns 3 and 4 are all observations in the full sample except those with zero investment. Approximately 8% of the full data set had observations with zero investment. The number of observations in the last two columns (5) and (6) decreased to 1758 because lagged values of some of the independent variables are needed in estimation.

(3) Consult the text for details of the estimation algorithm leading up to columns 5 and 6.

Source: Olley and Pakes (1992) Table 6.

Table 3: Alternative Estimates of Production Function Parameters[1]:
U.S. R&D Performing Firms, 1973, 1978, 1983, 1988
(standard errors in parentheses)

Variables[1]	Sample					
	Balanced Panel		Full Sample[3]			
	(1)	(2)	(3)	(4)	(5)	(6)
	Total	Within	Total OLS		Nonparametric F	
Labor	.496 (.022)	.685 (.030)	.578 (.013)	.551 (.013)	.591 (.013)	
Physical Capital	.460 (.014)	.180 (.027)	.372 (.009)	.298 (.012)	.321 (.016)	.320 (.017)
R&D Capital	.034 (.015)	.099 (.027)	.038 (.007)	.027 (.007)	.081 (.016)	.077 (.019)
Investment	-	-	-	.110 (.011)	-	
Other Variables[4]	-	-	-	-	Powers of h	Polynomial in P and h
# Observations[2]	856		2971		1571	

(1) The dependent variable in columns 1 to 4 is the log of sales, while in column 5 and 6, the dependent variable is the log(value added) - β*log(labor).

(2) The number of observations in the balanced panel for regressions in columns 1 and 2 are the observations for those firms that have continuous data over the period. Similarly, the 2971 observations in columns 3 and 4 are all observations in the full sample. (Only six observations had to be discarded because of zero investment.) The number of observations in the last two columns (5) and (6) decreased to 1571 because lagged values of some of the independent variables are needed in estimation.

(3) Consult the text for details of the estimation algorithm leading up to columns 5 and 6.

(4) The other variables in the equations are: Year, and Year x Industry 357 (i.e. computers) dummy variables.

Table 4: Estimation Results for Plants in Industry 384: Manufacture of Transport Equipment, Norway, 1983-89, N=409[1]

Variables	Basic Model		Augmented Model	
	OLS	IV	IV	Implied Elasticities[2]
Variable Inputs	.913 (.009)	.919 (.013)	.912 (.013)	1.059
Capital	.005 (.003)	.005 (.003)	.005 (.003)	0.006
Demand Shock[3]	-	-	.139 (.032)	-
Intercept	.001 (.003)	.001 (.003)	-.001 (.003)	-
RMSE	0.165	0.165	0.164	

(1) All estimates are based on first differences of the logarithms of the original variables.

(2) The estimated price elasticity is -1/.139 = -7.19 and the implied markup $\eta/(1+\eta) = 1.16$ multiplies the coefficients of the previous column.

(3) Aggregate industry output.

Source: Klette and Griliches (1994) Table 6.