A list of Working Papers on the last pages

No. 189, 1987

WHAT CAN INPUT TELL ABOUT OUTPUT? Analyzing Productivity and Efficiency in the Absence of Output Measures

by

Erik Mellander* and

Bengt-Christer Ysander*

This is a preliminary paper. Comments are welcome.

July, 1988

* Financial support from the Bank of Sweden Tercentenary Foundation is gratefully acknowledged.

1. Introducing the problem

In many productive activities in the economy we encounter problems in defining and/or measuring output and thus in analyzing productivity development. Even in some manufacturing sectors – particularly those dealing with "tailor–made", highly differentiated goods – the measurement of quality changes, both across a given assortment of goods and over time, can be extremely problematic.

In the steadily increasing service sector of the economy, this problem of quality measurement tends to be nearly all–pervasive. Particularly in personal services, the only relevant output measures will often be defined in terms of different quality dimensions, some of which are practically impossible to measure in an objective way. There are also conceptual problems involved in deciding on the appropriate output dimensions for, say, health and education. Are only immediate results to be counted or should we also take into account the long–term capabilities they create? How much weight should be attached to the way in which results are achieved? Etc.

For private services, sold in competitive markets, the value of total sales can, however, be used as a measure of the value of production. Thus, we do not then need output measures to be able to carry out cross–section studies of e.g. substitution possibilities, returns to scale or inter–firm differences in efficiency.[1] For comparisons over time, e.g. studies of productivity changes, output measures have been considered necessary, however.

For services distributed by government, free of charge or at "nominal" user charges, no market valuation of service output is registered. We are thus denied also the possibility of using market values for cross–section studies. Without reliable output measures, we seem to be blocked, not only from analyzing productivity developments, but also from studying the efficiency problems, which could be expected to be especially urgent due to the lack of market competition.

The usual way to tackle the analytical problems arising from the paucity of measures for service output, is to use some indicator variable as proxy for output.[2] In most applications, however, this does not solve the problem in any satisfactory way. The choice of proxy remains arbitrary, and there is no way of knowing how well the chosen proxy will reflect the changes in the actual output.

What we try to do in the following is to explore how much of the problem that can be solved if, instead of a proxy, we use that part of economic theory known as duality theory. By working with cost functions instead of production functions we try to avoid altogether the use of an output measure. We investigate how much that can be learnt about productivity and efficiency from input data alone. The main condition for using this approach turns out – not surprisingly – to be that the technology should be homothetic, i.e. that factor proportions should be independent of the level of production

Given a homothetic technology it will be shown that we can learn surprisingly much about productivity and efficiency from simply analyzing the development of prices and cost shares for the various factors of production. In addition, we will argue that the assumption of homotheticity may in many cases be neither overly restrictive nor particularly unrealistic.

2. The meaning of homotheticity

Our main theoretical question can be framed in the following way. Under what conditions can a production activity be characterized by input data only? For answering this question we use the dual relation between production and cost functions, first shown by Shephard (1953): given a cost function, which fulfills certain regularity conditions, it is always possible to define a production function from which the given cost function can be derived, under assumptions of cost–minimizing behavior. A technology which can be characterized by a production function can thus equally well be described in terms of a cost function and vice versa.

To simplify the discussion, let us for the present disregard technical change and the possibility of inefficiencies in production. We will come back to these issues in Sections 3 and 4. For the time being we thus assume a static technology and cost–minimizing producers. Let the minimum cost function be $C_o = C_o(y,\mathbf{p})$, where y is output and $\mathbf{p}$ denotes the vector of input prices, $\mathbf{p} = (p_1 \dots p_n)$. Accordingly:

$$C_o(y,\mathbf{p}) \equiv \mathbf{p}'\mathbf{x}_o \quad \text{where} \quad \mathbf{x}_o \equiv \min_{\mathbf{x}} \mathbf{p}'\mathbf{x} \quad \text{s.t.} \quad \mathbf{x} \in V(y) \tag{1}$$

where $\mathbf{x} = (x_1,\dots,x_n)$ denotes a vector of inputs and V(y) is the input requirement set, i.e. the set of input bundles that can produce at least y. The subindex o is, henceforth, used to denote that the value of the corresponding variable is optimal, i.e. cost–minimizing.

The above mentioned regularity conditions can e.g. be formulated as in Diewert (1971, p. 489–90) and will, i.a., imply that $C_o(y,\mathbf{p})$ should be non–decreasing in both y and $\mathbf{p}$ and linearly homogeneous and concave in $\mathbf{p}$. If these conditions are all satisfied, the cost function will describe all economically relevant aspects of the production technology. The producer's input demands can be derived by means of Shephard's lemma, according to which:

$$x_{oi} = x_{oi}(y,\mathbf{p}) \equiv \frac{\partial C_o(y,\mathbf{p})}{\partial p_i}, \qquad i = 1,\dots,n. \tag{2}$$

Since the cost function is linearly homogeneous in **p**, it further holds – by Euler's theorem – that:

$$\sum_{j=1}^{n} p_j x_{oj}(y,\mathbf{p}) = \sum_{j=1}^{n} p_j \frac{\partial C_o(y,\mathbf{p})}{\partial p_j} = C_o(y,\mathbf{p}) \,. \tag{3}$$

The system (2) thus contains all information available in the original cost function, since the factor demands multiplied by the factor prices add up to $C_o(y,\mathbf{p})$, as shown in (3).

Using (2) the optimal shares of factor costs become:

$$s_{oi} \equiv \frac{p_i \cdot x_{oi}(y,\mathbf{p})}{C_o(y,\mathbf{p})}, \qquad i = 1,\dots,n \,. \tag{4}$$

For our purposes the system of input cost shares has an important advantage over the system (2) of input demand equations: in contrast to the x_i's the cost shares are not necessarily dependent upon the output level, y.[3]

Our problem can thus be reformulated in the following manner: Which constraints must be imposed on the cost function (or production function) to make the cost shares independent of the quantity of output?

Shephard (1953, p. 45–47) has shown that if, and only if, the production technology is homothetic, the cost function can be factored according to:

$$C_o(y,\mathbf{p}) = f(y) \cdot g(\mathbf{p}) \,, \tag{5}$$

where f is a continuous, monotonically increasing function of y. The form of this function is determined by the scaling properties of the technology, i.e. whether the production function exhibits increasing, constant or decreasing returns to scale. In general, for a homothetic technology the rate of returns to scale vary with the output level. In the special case when the rate of return is the same for all output levels, the technology is homogeneous. If, in particular, the technology is linearly homogeneous, i.e. if there are constant returns to scale, then y can be substituted for f(y) in (5).

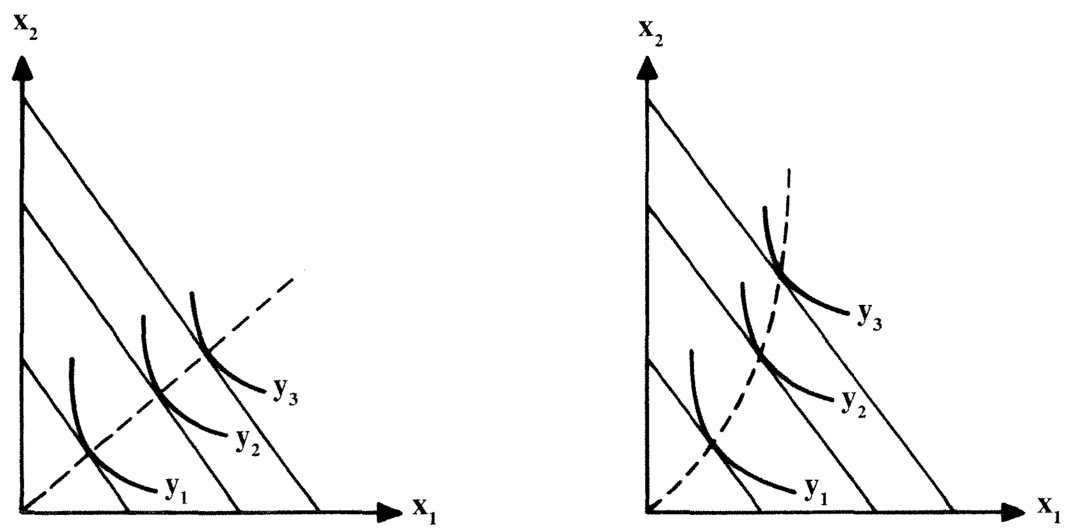
Given (5) the system of input cost shares becomes

$$s_{oi} = s_{oi}(\mathbf{p}) = \frac{p_i \cdot \frac{\partial g(\mathbf{p})}{\partial p_i}}{g(\mathbf{p})}, \qquad i = 1,...,n. \tag{6}$$

Specifying a homothetic functional form for C thus makes it possible to estimate the system (6), i.e. to estimate cost shares as functions of factor prices only, without having to take the level of output into account. This obviously provides us with a key to our main problem. Starting with input data – cost shares and factor prices – we can try to characterize and analyze production technology and producer behavior even in the absence of output measures.

The homotheticity assumption can of course be questioned. The characteristic feature of such a technology is that the optimal factor proportions, i.e. the ratios x_{io}/x_{oj}, $i \neq j$, are independent of the level of production, which is often a restrictive assumption. As shown by the isoquant diagram in Figure 1a, it implies that the expansion path, i.e. the dashed path describing the optimal input combinations at successively higher output levels $(y_1,y_2,...)$, is linear. For more general technologies the expansion path is non–linear as in Figure 1b. If, e.g., x_2 represents the capital input this figure can be taken to illustrate the often noted tendency to increase the capital intensity at larger scales of operation.

FIGURE 1: Expansion paths for homothetic and non – homothetic technologies in a static environment



a: homothetic technology b: non – homothetic technology

Homotheticity has also been decisively rejected in many studies concerned with, e.g., the manufacturing sector. However, it should be more easily motivated in the context of service production than in the production of goods. The reason is, of course, that services are more difficult to routinize and, hence, the scope for automatization more limited. Although this argument should be used with caution – for instance, one would expect factor proportions in large scale banking to differ substantially from those in provincial banks, whereas big and small kindergartens are presumably more alike in this respect – it appears as if the homotheticity assumption might be applicable where it is most needed, i.e. in contexts where no reliable output measures are available. In the case of government services homotheticity may, moreover, simply reflect centralized decision making, which tends to treat establishments of different size all alike. Furthermore, as we now introduce technical change we shall see that by taking technological developments into account we may, in effect, also allow for non–linear expansion paths, in spite of the homotheticity assumption.

3. Technical change and total factor productivity

Starting with the assumption of homotheticity we will now investigate what input data can reveal concerning technical change and productivity developments. We begin by considering the effect of technical change on the input requirements and the cost shares. Secondly, we discuss how the introduction of non–neutral, i.e. input specific, technical change into our homothetic technology can give rise to non–linear expansion paths, similar to those of non–homothetic technologies. Finally, we consider the connection between technical change and total factor productivity and show that it is possible to draw quite far–reaching conclusions about the rate of total factor productivity growth, even in the absence of an output measure.

Technical change (of a disembodied nature) can be incorporated in the model by augmenting the price function with a time index[4], t, resulting in the following cost function

$$C_o(y,\mathbf{p},t) = f(y)\cdot g(\mathbf{p},t) \ . \tag{7}$$

The system of cost shares corresponding to (7) is

$$s_{oi} = s_{oi}(\mathbf{p},t) = \frac{p_i \cdot \frac{\partial g(\mathbf{p},t)}{\partial p_i}}{g(\mathbf{p},t)} \ , \quad i = 1,\ldots,n \ . \tag{8}$$

By including the time index, input demands are allowed to shift over time not only in response to changes in relative factor prices but also because of exogenously determined technological developments. These developments affect the input requirements over time and, hence, also the input cost shares. In the following, we will use the letter τ to denote a relative time derivative. Accordingly, the rate of change in the usage of factor i resulting from technical change can be written

$$\tau_{x_{oi}} \equiv \frac{\partial \ \ell n \ x_{oi}(y,p,t)}{\partial t}, \quad i = 1,...,n. \tag{9}$$

Since in our case [cf. (2)]

$$x_{oi}(y,p,t) = f(y) \cdot \frac{\partial g(p,t)}{\partial p_i}, \quad i = 1,...,n, \tag{10}$$

the rate of technical change can be expressed in terms of only the input prices and the time index according to

$$\tau_{x_{oi}} = \frac{\partial^2 g(p,t)}{\partial p_i \partial t} \cdot \left[\frac{\partial g(p,t)}{\partial p_i}\right]^{-1}, \quad i = 1,...,n. \tag{11}$$

Further, the effects of technical change on the cost shares can be obtained in terms of the effects on the input demands as follows

$$\tau_{s_{oi}} \equiv \frac{\partial \ \ell n \ s_{oi}(p,t)}{\partial t} = \frac{1}{s_{oi}(p,t)} \frac{\partial}{\partial t} \left[\frac{p_i x_{oi}(y,p,t)}{\sum_{j=1}^{n} p_j x_{oj}(y,p,t)}\right]$$

$$= \tau_{x_{oi}} - \sum_{j=1}^{n} s_{oj} \tau_{x_{oj}}, \quad i = 1,...,n. \tag{12}$$

The technically induced rate of change in the i'th cost share will thus be equal to the difference between the rate of change in the demand for the i'th input and the corresponding cost–weighted average rate of demand changes, taken over all n inputs. However, due to the linear dependence among the cost shares the system of equations (8) – in contrast to the full rank system (10) – cannot provide us with all the n $\tau_{x_{oi}}$'s. Fortunately, the condition that the cost function be linearly homogeneous in the input prices implies one restriction on the $\tau_{x_{oi}}$'s. If this restriction is imposed there will only be n–1 independent measures of the rates of technical change, which is exactly what the cost shares are capable of generating.[5]

If $\tau_{s_{oi}} < 0$ technical change is characterized as factor i–saving and if $\tau_{s_{oi}} > 0$ it is said to be factor i–using. If $\tau_{x_{oi}} = \tau_{x_o} \neq 0$ for all i, so that $\tau_{s_{oi}} = 0$ for all i, then technical change is said to be neutral. This can only happen if the function g is multiplicatively separable in **p** and t, i.e. if (7) can be written in the form

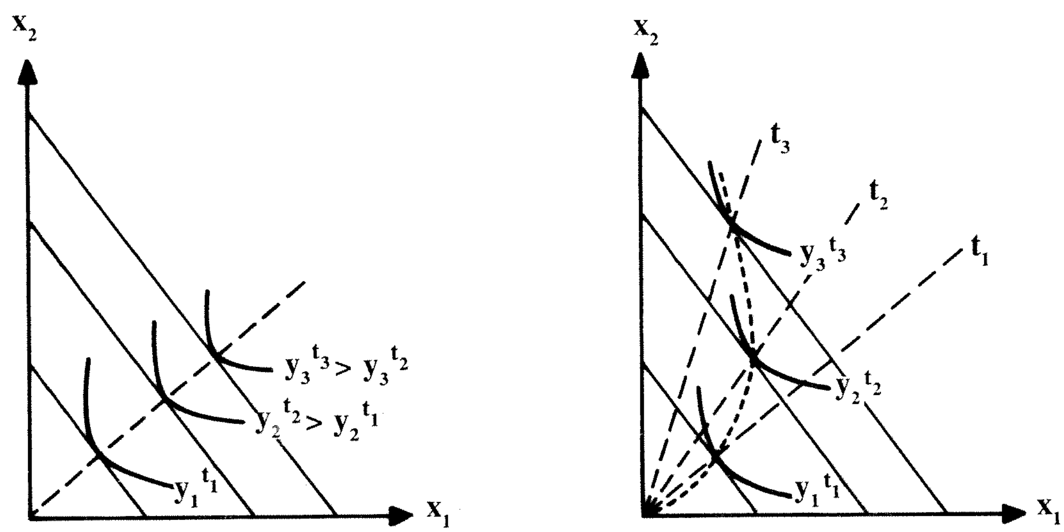
$$C_0(y,\mathbf{p},t) = f(y)\cdot g(\mathbf{p})\cdot h(t) \ .$$

It is easily verified by means of Shephard's lemma that in this case the system (8) degenerates to the system (6), i.e. that the cost shares are unaffected by technical change.[6]

Accordingly, to capture any effects of technical change through estimation of the system of input cost shares it is necessary to specify technical change as being non–neutral. This is no drawback as far as modeling is concerned; neutral technical change is presumably a very rare phenomenon. There is a negative consequence with respect to testing, however. This is due to the fact that the cost share systems corresponding to technologies undergoing neutral technical change are observationally equivalent to systems derived from completely static technologies. As a consequence it is impossible to prove that technical developments have had no influence on the production process and thus on the production costs. If the system (8) should prove superior to (6) on statistical grounds then, clearly, the hypothesis of no technical change must be rejected. If, on the other hand, statistical tests cannot discriminate between the two systems, technical changes might still have affected factor demand, albeit in a neutral fashion.

Regarding empirical implementation, the effect of allowing for non–neutral technical change might be very close to the effect of allowing for non–homotheticity, provided that there is a positive time trend in both input usage and output development. This is a relevant consideration here since most types of service production seem to have increased steadily over time. An illustration is givenin Figure 2, which illustrates the effects of neutral and non–neutral technical change between three points of time; 1, 2 and 3.

FIGURE 2: The observed expansion path of a homothetic technology under neutral and non–neutral technical change



a: neutral technical change　　　　b: non–neutral technical change

The impact of neutral technical change on the observed expansion path is illustrated in Figure 2a. As indicated by the unchanged slopes of the isocost lines, the relative input prices are here assumed to be fixed. Since the technical change is neutral it will by definition not alter the cost–minimizing factor mix, so the observed expansion path will remain linear. The only effect of technical change in this case will be to increase the output resulting from any given set of inputs. Thus e.g. the output obtainable with cost–constraint 2 at time t_2, $y_2^{t_2}$, will be higher than the corresponding production possibility at time t_1, $y_2^{t_1}$.

In analyzing the observed expansion path we are, however, faced with a difficult identification problem. The increases in productivity over time may be explained either by increased returns to scale or by technical change or by both. The figure thus illustrates the well known problem of separating the effects of technical change from the effects of returns to scale.

Figure 2b illustrates the consequences of non–neutral technical change. In addition to the identification problem just mentioned we will now also observe changes in the composition of the cost–minimizing factor mix, giving rise to a new problem of interpretation. As illustrated by the three rays through the origin at times t_1, t_2 and t_3, respectively, the expansion path is linear for any point in time. Over time the path itself shifts, however. Accordingly, if output grows over time the observed expansion path will look like the dotted line in the diagram, strongly resembling that of non–homothetic technology (cf. Figure 1b).

This means, on the one hand, that estimates of the effects of non–neutral technical change have to be interpreted with great caution. For instance, the observed expansion path of a homothetic technology undergoing non–neutral technical change may be almost indistinguishable from that of non–homothetic technology subject to neutral technical change.[7] On the other hand, it also means that from a practical point of view our maintained hypothesis of homotheticity need not be particularly restrictive. Given that there are time trends in inputs and output, effects of the production level, i.e. effects of non–homotheticity, may be captured by the time index variable.

Concerning, finally, the effects of technical change on the rate of total factor productivity a general duality result derived by Ohta (1974) can be applied. Let

$\psi(\mathbf{x},t)$ denote the production function to which the cost function (7) is dual, when $\mathbf{x} = \mathbf{x}_0$. The primal rate of total factor productivity is then defined according to

$$\tau_{\psi_o} \equiv \frac{\partial \, \ell n \; \psi(\mathbf{x}_0,t)}{\partial t}. \tag{13}$$

What Ohta has shown is that the following dual relationship holds

$$\tau_{\psi_o} = (-\tau_{C_o}) \cdot (\epsilon_{C_o y})^{-1} \tag{14}$$

where

$$\tau_{C_o} = \frac{\partial \, \ell n \; C_0(y,\mathbf{p},t)}{\partial t} \tag{15}$$

and

$$\epsilon_{C_o y} \equiv \frac{\partial \, \ell n \; C_0(y,\mathbf{p},t)}{\partial \, \ell n \; y}. \tag{16}$$

The first factor in (14), the negative of the rate of total cost diminution, is the dual representation of technical change. The second factor, the inverse of the elasticity of total cost with respect to output, is the dual form of the rate of return to scale. Returns to scale are increasing if $\epsilon_{C_o y} < 1$, constant if $\epsilon_{C_o y} = 1$, and decreasing if $\epsilon_{C_o y} > 1$.

Application of (15) and (16) to the cost function (7) shows that in the present context (15) becomes[8]

$$\tau_{C_o} = \frac{\partial \, \ell n \; g(\mathbf{p},t)}{\partial t}, \tag{17}$$

while the cost elasticity is given by

$$\epsilon_{C_oy} = \frac{f(y)}{y \cdot f'(y)}, \tag{18}$$

Inserting (17) and (18) into (14) we obtain the following expression for the dual rate of total factor productivity

$$\tau_{\psi_o} = \left[-\frac{\partial \ln g(\mathbf{p},t)}{\partial t} \right] \cdot \left[\frac{f(y)}{y \cdot f'(y)} \right]^{-1}. \tag{19}$$

Because of the occurrence of y in the last factor of (19), it is obvious that, in general, the system (8) of input cost shares does not provide all the information needed to calculate an estimate of the rate of total factor productivity. However, as long as production does not take place at negative returns to scale ϵ_{C_oy} will be strictly positive,[9] in which case the sign of (19) will be equal to the sign of the first factor on the right hand side, i.e. the dual rate of technical change. Accordingly, with this weak qualification the question of whether total factor productivity is increasing or decreasing can always be answered by means of (17) which can be obtained from the estimation of the system of cost shares.

If, further, the technology is homogeneous the rate of return to scale will be independent of the level of y and so (18) will be equal to a constant, instead of being a function of y. In that case the relations between the rates of total factor productivity at different points in time will be equal to the corresponding relation between the rates of cost diminution, leaving only the levels of total productivity growth undetermined. If, finally, the technology is linearly homogeneous, i.e. characterized by constant returns to scale, then (18) will be equal to unity and the negative of τ_{C_0} will be identical with the rate of growth in total factor productivity.

From a theoretical point of view our possibilities of learning about productivity growth from a study of cost shares will thus depend on how restrictive assumptions we are willing to make concerning returns to scale. In practical implementation, however, the data may well, at least partly, decide this issue for us. As explained above, positive time trends in the input and (unknown) output series are most likely to result in much of the returns to scale effects being included in the measurement of technical change. Accordingly, such time trends will have the effect of making our conclusions regarding productivity growth less contingent upon assumptions about returns to scale than indicated by a purely theoretical consideration.

Next we will try to discover what input data can reveal concerning possible inefficiency in production.

4. Inefficiency in production

The previous two sections have demonstrated that the dual approach makes it possible to obtain a quite extensive description of the structure of production, even if no output measure is available. However, by definition $C_0(y,\mathbf{p},t)$ denotes the smallest total cost attainable in time period t for input vectors yielding at least the output y. Accordingly, the dual representation of the production structure is, in general, valid only in the case of cost minimization.[10] Since there are applications for which the assumption of cost minimization can be questioned – e.g. the production of public services – it is important to investigate if this condition can be weakened.

Deviations from minimum costs (which, of course, must always be positive) can arise for several reasons. The one that most readily comes to mind is perhaps inefficient producer behavior. But there may be other causes as well, e.g. imperfections in the factor markets. These can take the form of regulatory constraints – such as rate of return regulations – or rationing schemes.[11] Restrictions in the goods market, e.g. production quotas, may be another reason. Similarly, if the exogenously given demand is highly variable it may be impossible to avoid some slack in off–peak periods in order to be able to cope with the peaks.[12]

In the following, we will be content with merely examining in what ways the existence of inefficiencies can be modeled and, secondly, how their cost–increasing effects can be estimated. We will consider two types of inefficiency; technical inefficiency and allocative or price inefficiency.[13] The easiest way to define these are by means of the corresponding efficiency concepts.

A producer is said to be technically efficient if, at a given level of production, he cannot reduce the amount of any input without at the same time reducing the amount of output. Accordingly, the production is technically efficient if it takes place at some point along the isoquant.

Price efficiency is directly related to the first order conditions for cost minimization. As is well known, these conditions require that the inputs be chosen such that their relative marginal productivity values, or shadow values, be equal to their relative prices. Denoting, as before, the production function by ψ and letting the product price be denoted by π, this can be formalized as requiring equality between the ratios

$$\frac{\pi \cdot \psi_i(\mathbf{x}_q, t)}{\pi \cdot \psi_j(\mathbf{x}_q, t)} \equiv \frac{w_i}{w_j} \qquad \text{and} \qquad \frac{p_i}{p_j}$$

for all $i \neq j$, where $\psi_i(\mathbf{x}_q, t)$ denotes the partial derivative of ψ with respect to x_i, evaluated at the point $\mathbf{x}_q = (x_{q1}, \ldots, x_{qn})$. The relative shadow values, i.e. the w_i/w_j, will be equal to the relative input prices if, and only if, the factor proportions at the points $\mathbf{x}_o$ and $\mathbf{x}_q$ are equal. Using the n'th input to normalize this requirement can be expressed according to

$$\frac{x_{qi}}{x_{qn}} = \frac{x_{oi}}{x_{on}}, \quad i = 1, \ldots, n-1. \tag{20}$$

If this equality holds the inputs are optimally allocated and the input mix is said to be price efficient.

Having thus defined our inefficiency concepts we proceed to modeling aspects. It will be practical to begin by discussing price inefficiency. The interest in price inefficiency has grown with the increasing popularity of the dual approach to applied production theory. The reason is, of course, that while the data required to estimate production functions are insufficient to obtain measures of price inefficiency there is no extra data requirement if the production activity is studied from the cost side – data on input prices will be collected anyhow.

A simple, yet quite powerful, specification is the following one, originally proposed by Toda (1976) for the two input case and subsequently generalized by Atkinson and Halvorsen (1980, 1984) to the n input case. The shadow prices, i.e. the marginal productivity values, are assumed to be proportional to the factor prices actually observed, the p_i's, according to

$$w_i = \lambda_i p_i \,, \qquad \lambda_i > 0 \,, \qquad i = 1,\ldots,n, \tag{21}$$

where λ_i is an (unknown) input specific proportionality constant.

Assuming, for the moment, that there is no technical inefficiency the realized cost shares – as opposed to the optimal, cost minimizing shares – can be derived as follows. First, notice that if the producer is technically efficient his choice of input levels can be regarded as the result of minimizing total shadow costs, $\Sigma_{k=1}^{n} w_k x_k$. Using (21) the minimum total shadow costs can be expressed in terms of the actually observed prices as

$$C_\lambda = C_\lambda(y,\mathbf{p},t) = f(y) \cdot g(\mathbf{\Lambda}_d \mathbf{p},t) \,, \tag{22}$$

where $\mathbf{\Lambda}_d$ denotes a n by n diagonal matrix with iith element equal to λ_i . This is the cost function which would have been the minimum cost function, had the prevailing prices been given by the vector **w** instead of the vector **p**. Application of Shephard's lemma to (22) yields the input levels which minimize total shadow costs, the $x_{\lambda i}$'s , according to

$$x_{\lambda i} \equiv \frac{\partial C_\lambda}{\partial(\lambda_i p_i)} = f(y) \cdot \frac{\partial g(\mathbf{\Lambda}_d \mathbf{p},t)}{\partial(\lambda_i p_i)} \qquad i = 1,\ldots,n. \tag{23}$$

Using (23), the <u>realized total cost</u>, C_r , can be written

$$C_r \equiv \sum_{k=1}^{n} p_k x_{\lambda k} = f(y) \cdot \sum_{k=1}^{n} p_k \cdot \frac{\partial g(\mathbf{\Lambda}_d \mathbf{p},t)}{\partial(\lambda_k p_k)}, \tag{24}$$

and the corresponding cost shares will be

$$s_{ri} \equiv \frac{p_i \cdot x_{\lambda i}}{C_r} = \frac{p_i \cdot \dfrac{\partial g(\mathbf{\Lambda}_d \mathbf{p},t)}{\partial(\lambda_i p_i)}}{\sum_{k=1}^{n} p_k \cdot \dfrac{\partial g(\mathbf{\Lambda}_d \mathbf{p},t)}{\partial(\lambda_k p_k)}}, \qquad i = 1,\ldots,n. \tag{25}$$

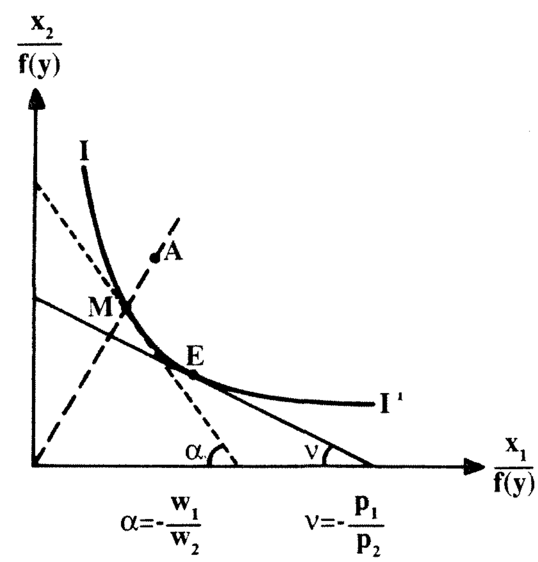
However, as C_λ is linearly homogeneous in the $(\lambda_i p_i)$'s, its derivatives, (23), must be homogeneous of degree zero in the same variables or, equivalently, in the λ_i's. Accordingly, the cost shares (25) must homogeneous of degree zero in the λ_i's, too. Therefore the absolute values of the λ_i's cannot be obtained. This property is consistent with the fact that the first order conditions for cost minimization only concern relative prices. The following normalization rule can thus be imposed without loss of generality

$$\lambda_n = 1\,. \tag{26}$$

Price efficiency can be tested by means of the hypothesis that not only λ_n but all the λ_i's are all equal to unity, in which case (25) is identically equal to the system (8) of cost minimizing shares.

Figure 3 can be used to illustrate the test in the two input case. In order to be capable of illustrating both price inefficiency and technical inefficiency the input demands measured along the horizontal and the vertical axes have been divided by f(y). Accordingly, all points lying on or to the northeast of the isoquant II' correspond to the same volume of output. For the sake of interpretation, notice that if there are constant returns to scale, i.e. if f(y)=y, this amounts to drawing the diagram in the space of input/output–coefficients. Figure 3 can thus be viewed as a generalization of such a diagram.

FIGURE 3: Price inefficiency



The isocost shown by a solid line corresponds to the factor prices actually observed, p_1 and p_2 . Since we are assuming technical efficiency, production is taking place somewhere along the isoquant II' . Given the observed factor prices the producer will minimize costs by operating at the point E. Now, assume that production is actually taking place at the point M. With the input prices at the observed levels this point is obviously price inefficient. However, it would have been a price efficient location if the isocost had been given not by the solid but by the dotted line. The slope, α, of this latter isocost equals the ratio of the shadow prices since, given that production occurs at M, this is the relative price corresponding to cost minimization. The hypothesis to be tested is thus whether the slope of the hypothetical isocost, α, is significantly different from ν, the slope of the actual isocost. In the two input case this simply means testing if $\lambda_1 = 1$ [λ_2 is equal to unity from the beginning, in accordance with the normalization rule (26)].

If the hypothesis of price efficiency is rejected, we can define a measure of the increase in the total costs caused by the non–optimal factor mix – in spite of the fact that our lack of output measure would seem to prevent us from obtaining a measure of minimum cost. The relative increase in total costs, κ say, can be obtained in the following way

$$\kappa \equiv \frac{C_r}{C_o} - 1 = \frac{\sum_{j=1}^{n} p_j \cdot \frac{\partial g(\mathbf{\Lambda}_d \mathbf{p}, t)}{\partial(\lambda_j p_j)}}{g(\mathbf{p}, t)} - 1 . \tag{27}$$

The numerator involving the partial derivatives is equal to the denominator in the cost shares (25). Moreover, due to the linear homogeneity of the function $g(\mathbf{p},t)$ in $\mathbf{p}$, the value of this function can be calculated by setting all the λ_i's in the numerator equal to one.[14] From the definition of κ it is also clear that once we have a numerical value for κ we can use this figure and the realized total costs, C_r, to solve for an estimate of the absolute cost increase caused by the price inefficiency, according to

$$C_r - C_o = C_r \cdot \frac{\kappa}{1-\kappa} . \tag{28}$$

Quite a large amount of information can thus be extracted by means of the simple specification (21). However, a relevant concern is to which extent this information is contingent upon the assumption that the production process is technically efficient. Since both Toda (op. cit.) and Atkinson and Halvorsen (op. cit.) discuss only price inefficiency they leave this question open. We will go on, however, and consider technical inefficiency, too.

Technical inefficiency can be either neutral or non–neutral in nature. In the former case the relative waste, i.e. the degree of overutilization, is the same for all inputs, while in the latter it may vary among the inputs.[15] We will first investigate the effect of introducing of neutral technical inefficiency.

Diagrammatically, neutral technical inefficiency can be illustrated by a movement from the point M to the point A in Figure 3. Since A lies northeast of the isoquant II' it cannot be a technically efficient point. Further, as A and M lie on the same ray through the origin (dashed) and, accordingly, are characterized by the same factor proportions they must suffer from the same price inefficiency. And the test for price efficiency described above is equally applicable to A as to M or, indeed, to any other point along the dashed ray (above M). The reason is that the hypothesis tested only concerns the relative factor usage, employing no information about the absolute amounts used of the various inputs.

Neutral technical inefficiency thus does not affect the formulation of the test for price inefficiency. However, it is obvious that if technical inefficiency exists then the expressions (27) and (28) do not provide information about all the extra costs arising from inefficiencies in production. To enable computation of the economic consequences of technical inefficiency its occurrence has to be explicitly modeled. Figure 3 suggests a simple way to do this: the distance between the actual point, A, and the technically efficient point, M, can be used to obtain a measure of the degree of technical inefficiency. Formally, let x_{ai} denote the actually observed usage of input i. Technical inefficiency can then be taken into account by means of the following specification:

$$x_{ai} = (1 + \zeta)\cdot x_{\lambda i}\,, \quad \zeta \geq 0, \quad i = 1,\ldots,n, \tag{29}$$

where ζ represents the degree of overutilization. The production process is thus technically efficient if $\zeta = 0$. Notice that one and the same ζ applies to all inputs; a movement along the dashed ray in Figure 3 from M to A, i.e. from (x_{m1},x_{m2}) to (x_{a1},x_{a2}) implies that the usage of both inputs is increased by the same percentage. It is this property that makes (29) a specification of neutral technical inefficiency.

Unfortunately, in our context this neutrality leads to an identification problem, stemming from the fact that the system of input cost shares is unaffected by the formulation (29). This can be seen as follows. Given (29), the actually observed cost shares can be written

$$s_{ai} = \frac{p_i\cdot(1+\zeta)\cdot x_{\lambda i}}{\sum\limits_{k=1}^{n} p_k\cdot(1+\zeta)\cdot x_{\lambda k}} = \frac{p_i\cdot x_{\lambda i}}{\sum\limits_{k=1}^{n} p_k\cdot x_{\lambda k}} = s_{ri}\,, \tag{30}$$

$i = 1,\ldots,n$. The cost shares taking into account both price inefficiency and neutral technical inefficiency are thus equal to the cost shares computed with regard to price inefficiency only. Accordingly, the system of input cost shares is invariant to neutral technical inefficiency. Hence, the specification (29) cannot provide a solution to the problem of estimating the costs of technical inefficiency.

It should be noted that this negative conclusion has a positive counterpart. It implies that in the presence of neutral technical inefficiency the system of cost shares (25) can still be used to study productivity developments and price inefficiency; neutral technical inefficiency will not introduce any biases into these analyses.

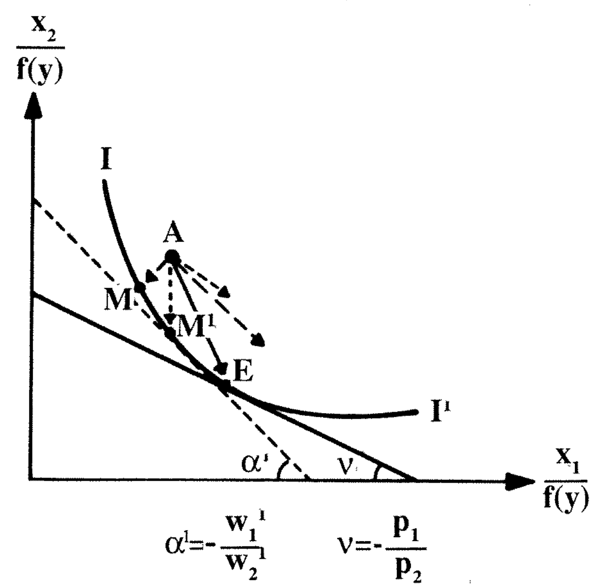
To be able to take technical inefficiency explicitly into account it is, however, necessary to allow it to be non–neutral, i.e. to let the degree of overutilization vary among the inputs. This could be accomplished by means of several different specifications, all of which would have the following two properties in common

$$x_{ai} \geq x_{\lambda i} \qquad i = 1,\ldots,n, \tag{31a}$$

$$\frac{x_{ai}}{x_{\lambda i}} \neq \frac{x_{aj}}{x_{\lambda j}} \quad \text{for at least one} \; = 1,\ldots,n\,. \tag{31b}$$

We will show, however, that non–neutral specifications of technical inefficiency also lead to – or rather call attention to – a kind of identification problem, namely that of discriminating between technical inefficiency on the one hand and price inefficiency on the other. In principle, there is no unambiguous way to partition the sum of technical inefficiency and price inefficiency into its components. As soon as one wishes to consider both technical and price inefficiency simultaneously the problem illustrated in Figure 4 will emerge.

FIGURE 4: Equivalent combinations of price inefficiency and technical inefficiency



The isoquant and the points A, M and E have been reproduced from Figure 3, as well as the isocost corresponding to the observed factor prices, p_1 and p_2. We now make the thought experiment that the producer operating at the point A moves to the efficient point, E. This movement, illustrated by the solid arrow, can be considered as the sum of two vectors, representing movements towards technical and price efficiency, respectively. The problem is that, in principle, the sum can be decomposed in an infinite number of ways. Two of these are shown in the diagram.

The dashed vectors illustrate the special case in which the adjustment towards technical efficiency is neutral, i.e. when (29) holds. This adjustment is shown by the vector from A to M, whereas the other one is equivalent to the movement from M to E, i.e. the movement for price efficiency. Of the dotted vectors the one pointing due south, to M', corresponds to a non–neutral adjustment towards technical efficiency where the amount of x_1 is held constant while decreasing the use of x_2. By elimination, the other vector must then show the movement yielding price efficiency.

The usefulness of the neutral specification (29) lies in the fact that it imposes a restriction which enables us to pick out exactly one of all the conceivable decompositions. Of course, in principle this restriction is no less arbitrary than any other one. It has two advantages, however. One is that it lends straightforward economic interpretations to the two inefficiency concepts, separating, as it does, between pure quantity effects of inefficiency (technical inefficiency) and effects relating to factor prices and, accordingly, factor proportions (price inefficiency). If non–neutral technical inefficiency is allowed this distinction becomes blurred; cf. the movement from A to M' in Figure 4 which involves a considerable change in the factor mix. Secondly, as shown above, neutral technical inefficiency has no effect on the analysis of price inefficiency. In contrast, any attempt to quantify non–neutral technical inefficiency will have an impact on the measures of price inefficiency.[16]

In regard to technical inefficiency we thus seem to be caught in a dilemma. If we assume it to be neutral we cannot estimate it but we will be able to get unambiguous measures of the price inefficiency. If, on the other hand, we specify technical inefficiency as being non–neutral it may be possible to quantify it but if we succeed in doing so its magnitude will be dependent upon the measure of price inefficiency.

To get out of this dilemma it is first necessary to decide on the relative importance of properly measuring price inefficiency and of having the possibility to estimate the combined effects of price and technical inefficiency. It might be argued that the potential advantage of the last alternative – i.e. that it, possibly, would enable computation of the effects on total costs not only of price inefficiency but also of technical inefficiency – is so important that the question of discriminating between these inefficiencies can be disregarded.

One way to implement this strategy would be to introduce non–neutral technical inefficiency by means of the generalized input/output–coefficients that we employed in the construction of Figure 3 and 4. Letting ζ_i represent the non–neutral technical inefficiency connected with input i this could be formalized in general terms according to

$$\frac{x_{ai}}{f(y)} = \phi\left[\frac{x_{\lambda i}(y,p,t)}{f(y)},\zeta_i\right], \quad \zeta_i \geq 0, \quad i = 1,\ldots,n \tag{32}$$

where ϕ is a function, taken to be increasing in both of its arguments. As indicated by the subindex λ, the first argument allows for price inefficiency [cf. (23)] and the second for non–neutral technical inefficiency. Moreover, the right hand side of (32) is independent of the output level. This property will carry over to the actually observed cost shares, since these can be written

$$s_{ai} = \frac{p_i \cdot \frac{x_{ai}}{f(y)}}{\sum\limits_{k=1}^{n} p_k \cdot \frac{x_{ak}}{f(y)}}, \qquad i = 1,\ldots,n. \tag{33}$$

One problem still remains, however, arising from the fact that whereas price inefficiency concerns the relative utilization of the various inputs, technical inefficiency relates to the absolute amounts used. This means that there are as many measures of non–neutral technical inefficiency as there are inputs, i.e. n. But, since the system of input cost shares only contains n–1 independent equations it can only provide at most n–1 estimates of non–neutral technical inefficiency. At least one restriction must thus be imposed on the ζ_i's in (33).[17] Unfortunately, economic theory can be of no guidance in the choice of a suitable constraint.

Moreover, specifications of non–neutral technical inefficiency cannot be given the same general applicability as the specification (21) of price inefficiency. For this reason, and because of the dependency between the measures of price inefficiency and non–neutral technical inefficiency, it is not possible to say if the parameters of (33) are identified without considering the functional forms for the cost function and the stochastic specifications to be employed in the estimation of cost share system.[18]

Although neither the requirement concerning the constraint on the ζ_i's, nor the appropriate choices of functional form and stochastic specification seem to pose insurmountable problems a detailed discussion of these issues are beyond the scope of this paper. We will thus be content with concluding that given a suitable restriction on the (non–neutral) technical inefficiency characterizing the production process it should in principle be possible to obtain a measure of the combined effect of price inefficiency and technical inefficiency. The details concerning implementation are, however, left as a task for further research.

5. Summing up

What can you learn about productivity and efficiency when no reliable output measures are available? This is the question we have tried to answer with the help of duality theory. Our results indicate that the possibilities to characterize the production process by means of only input data are indeed much greater than could be expected intuitively. As long as homotheticity can be assumed, knowledge of cost shares and input prices enable us to analyze both productivity effects of technical change and efficiency in producer behavior. Moreover, we find it that in many of the economic activities constituting natural candidates for application of our results – e.g. private and, in particular, public services – the assumption of homotheticity need not be either particularly unrealistic, nor very restrictive.

In summary, our analysis leads to the following three conclusions:

1. Technical change can be modeled, provided that it is specified as being non–neutral, i.e. affecting the various inputs differently; neutral technical change will have no impact on the cost shares. This means, i.a., that with constant returns to scale we will be able to measure the rate of growth in total factor productivity. Further, non–neutral specifications of technical change may, in practice, weaken the seemingly restrictive implication of homotheticity that the cost minimizing factor proportions are independent of the output level.
2. Price inefficiency can be explicitly taken into account in the analysis, to show if the composition of the factor mix is optimal, given the relative input prices. If price inefficiency is ascertained measures of the resulting increase in total costs can be obtained. Moreover, these measures will not be distorted by the occurrence of neutral technical inefficiency – i.e. wasteful input usage, characterized by the degree of overutilization being the same for all inputs – since neutral technical inefficiency will not affect the cost shares.
3. Input specific, or non–neutral, technical inefficiency can in principle be

captured by the cost shares. Further, by specifying simultaneously both non–neutral technical inefficiency and price inefficiency it should be possible to estimate their combined effect on total costs. However, any specification involving non–neutral technical inefficiency will have the disadvantage of making the separation between price inefficiency and technical inefficiency ambiguous.

There is thus hope that our dual approach can make a contribution to the pressing problem of measuring productivity and inefficiency in the many sectors of the economy for which there are no satisfactory indicators of the production result. It should be pointed out, however, that even though we manage in this way to escape the quality problem involved in measuring output, we are still left with the, sometimes, almost equally difficult quality problem inherent in input measurement.

NOTES

[1] Of particular interest in this context are the US studies, carried out in recent years, of the relative efficiency of private versus public electricity production. There are no problems with measuring output in this special case – one simply counts the numbers of produced KWh during a certain period of time. Both econometric and non–parametric, linear programming approaches have been used in these studies (Cf. e.g. Pescatrice and Trapani (1980), Atkinson and Halvorsen (1984), Färe et.al. (1985)). The perhaps counterintuitive conclusion seems to be that the publicly run electricity plants are at least as efficient as the private (Cf. Färe et.al. (1985), p. 89–90).

[2] Early, well–known, examples of this approach are provided by the attempts of Kiesling (1967) and Feldstein (1967) to estimate production functions for education and health respectively. A recent study, which exemplifies the later methodological developments in this area, is the estimation by Bjurek et.al (1986) of frontier production functions for local public insurance offices in Sweden, based on combined cross–section and time–series data.

[3] There is a price to be paid for this property, however. Using (3) and (4) it can be seen that the cost shares sum identically to one. Accordingly, whereas the rank of the system (2) of input demands is equal to n, the rank of the system (4) of cost shares is only n–1. Because of this linear dependence the original cost function cannot be recovered from the cost shares. The cost shares thus contain less information than the input demands, cf. McElroy(1987, p. 743). Still, plenty of information can be obtained by means of (4), e.g. the substitution and price elasticities which are commonly used to characterize the production technology and the producer behavior.

[4] For examples, see Parks (1971), Woodland (1975), Berndt and Khaled (1979), and Nadiri and Schankerman (1980).

[5] For instance, if the technology is of the translog type proposed by Cristensen, Jorgensen and Lau (1973) the restriction implies

$$n^{-1} \cdot \Sigma_{j=1}^{n} \tau_{x_{oj}} = \Sigma_{j=1}^{n} s_{oj} \tau_{x_{oj}}$$

i.e. equality between the simple arithmetic mean and the cost–weighted average. This restriction makes it possible to solve for one of the τ_{xoi} 's in terms of the other and of the n–1 independent cost shares.

[6] This proves that $g(\mathbf{p},t)=g(\mathbf{p})\cdot h(t)$ is a sufficient condition for neutral technical change. To prove necessity, notice that neutrality requires the right hand side of (11) to be equal for all i. Since we know that, in general, $[\partial g(\mathbf{p},t)/\partial p_i] \neq [\partial g(\mathbf{p},t)/\partial p_j]$ this implies that the function $g(\mathbf{p},t)$ must have the property that

$$\frac{\partial^2 g(\mathbf{p},t)}{\partial p_i \partial t} = \frac{\partial g(\mathbf{p},t)}{\partial p_i} \cdot \varphi(t) , \quad i = 1,\ldots,n,$$

where φ is a (non–zero) function of t. To have this property the function $g(\mathbf{p},t)$ must, however, be multiplicatively separable in **p** and t.

[7] A comparison of these two cases in terms of overall goodness–of–fit (likelihood values) can be found in Berndt and Khaled (1979, Table 2). They find the alternative with non–homothetic technology and neutral technical change slightly superior according to this criterion. They make no comparisons in terms of, e.g., price and substitution elasticities, however.

[8] Since it can be shown that

$$\tau_{C_o} = \frac{\partial \ln g(\mathbf{p},t)}{\partial t} = \Sigma_{j=1}^{n} s_{oj} \tau_{x_{oj}}$$

the computation of the dual rate of technical change does not require any extra effort; the sum on the right hand side will be of use in the computations of the effects of technical change on the cost shares, too [cf. (12)].

[9] Notice that this does not exclude the possibility that decreasing returns to scale prevail; the inverse of $\epsilon_{C_o y}$ may lie in the open interval]0,1[.

[10] Intuitively, the qualification "in general" would seem redundant. It is not, however, as will be shown below.

[11] This potential difficulty in correctly attributing the causes of deviations from cost minimization is also noted by Atkinson and Halvorsen (1984, note 3).

[12] Cf. Fuss and McFadden (1978)

[13] Among the inefficiency concepts relevant here, i.e. those relating to the input side [see, e.g. Försund and Hjalmarsson (1974)], we are thus not considering scale inefficiency. The reason is that, by construction, our approach does not permit any analysis of scaling properties; the cost shares are invariant with respect to both the level of production, y, and the scaling function, f. This invariance property also has a positive implication, however. It means that, if present, scale inefficiency will not have any effect on the analysis of other types of inefficiencies, or of the productivity effects of technical change.

[14] Linear homogeneity of $g(\mathbf{p},t)$ in **p** implies that

$$g(\mathbf{p},t) = \Sigma_{j=1}^{n} p_j \cdot \frac{\partial g(\mathbf{p},t)}{\partial p_j},$$

which is precisely the expression we get if we put all the λ_i's in the numerator of (27) equal to one.

15 For discussions of neutral technical inefficiency, cf. Debreu (1951) and Farrell (1957). Non–neutral technical inefficiency has been considered by Färe (1975).

16 Strangely enough, this does not seem to have been generally recognized in the literature. For instance, Lovell and Sickles (1983) consider both price inefficiency and non–neutral technical inefficiency without even mentioning the problem.

17 This difficulty illustrates one limitation of analyzing the cost shares as compared to the input demands. Since the system of input demand equations contains n independent equations it is capable of providing unrestricted measures of non–neutral technical inefficiency for each of the n inputs, thus eliminating the need for a constraint on the ζ_i's.

18 See, e.g., the discussion in Lovell and Sickles (1983)

REFERENCES

Atkinson, S.E. and Halvorsen, R. (1980), "A test of Relative and Absolute Price Efficiency in Regulated Utilities", Review of Economics and Statistics LXII: 4–80.

——— (1984), "Parametric Efficiency Tests, Economies of Scale and Input Demand in U.S. Electric Power Generation", International Economic Review 25: 647–662.

Berndt, E.R. and Khaled, M.S. (1979), "Parametric Productivity Measurement and Choice Among Flexible Functional Forms", Journal of Political Economy 87: 1220–1245.

Bjurek, H., Hjalmarsson, L. and Försund, F.R. (1986), "Productivity Change and Productive Efficiency in Swedish Social Welfare Offices", Paper presented at the European Meeting of the Econometric Society in Budapest.

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

Debreu, G. (1951), "The Coefficient of Resource Utilization", Econometrica 19: 273–292.

Diewert, W.E. (1971), "An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function", Journal of Political Economy 79: 481–507.

Färe, R. (1975), "Efficiency and the Production Function", Zeitschrift für Nationalökonomie 35: 317–324.

Färe, R., Grosskopf, S. and Logan, J. (1985), "The Relative Performance of Publicly–Owned and Privately–Owned Electric Utilities", Journal of Public Economics 26: 89–106.

Farrell, M.J. (1957), "The Measurement of Productive Efficiency", Journal of the Royal Statistical Society, Series A 120: 253–281.

Feldstein, M.S. (1967), Economic Analysis of Health Service Efficiency: Econometric Studies of the British National Health Service (vol. 51 of Contributions to Economic Analysis), Amsterdam: North–Holland.

Försund, F.R. and Hjalmarsson, L. (1974), "On the Measurement of Productive Efficiency", Swedish Journal of Economics 76: 141–154.

Fuss, M. and McFadden D. (1978), "Flexibility versus Efficiency in Ex Ante Plant Design", in Fuss, M. and McFadden D. (eds.): Production Economics: A Dual Approach to Theory and Applications, Amsterdam: North–Holland.

Kiesling, H.J. (1967), "Measuring a Local Government Service: A Study of School Districts in New York State", Review of Economics and Statistics XLIX: 356–367.

Lovell, C.A.K. and Sickles, R.C. (1983), "Testing Efficiency Hypotheses in Joint Production: A Parametric Approach", Review of Economics and Statistics LXV: 51–58.

McElroy, M.B. (1987), "Additive General Error Models for Production, Cost and Derived Demand or Share Systems", Journal of Political Economy 95: 737–757.

Nadiri, I. and Schankerman, M. (1980), "The Structure of Production, Technological Change, and the Rate of Growth of Total Factor Productivity in the U.S. Bell System" in Cowing, T.G. and Stevenson, R.E. (eds.): Productivity Measurement in Regulated Industries, New York: Academic Press.

Ohta, M. (1974), "A Note on the Duality between Production and Cost Functions: Rate of Returns to Scale and Rate of Technical Progress", Economic Studies Quarterly 25: 63–65.

Parks, R.W. (1971), "Price Responsiveness of Factor Utilization in Swedish Manufacturing, 1870–1950", Review of Economics and Statistics LIII: 129–139.

Pescatrice, D.R. and Trapani, J.M. (1980), "The Performance and Objectives of Public and Private Utilities Operating in the United States", Journal of Public Economics 13: 259–276.

Shephard, R.W. (1953), Cost and Production Functions, Princeton, New Jersey: Princeton University Press.

Toda, Y. (1976), "Estimation of a Cost Function When the Cost is Not Minimum: The Case of Soviet Manufacturing Industries, 1958–1971", Review of Economics and Statistics LVIII: 259–268.

Woodland, A.D. (1975), "Substitution of Structures, Equipment and Labor in Canadian Production", International Economic Review 16: 171–187.