LONG-RUN EXPECTATIONS AND CAPACITY

by

Robert H. McGuckin and Peter Zadrozny

CES 88-1 April 1988

Abstract

In this paper, we argue at a general level, that recent economic models of capacity and of its utilization are deficient because they do not adequately take into account firms' long-run expectations about conditions which are pertinent to their investment decisions, i.e., their decisions about altering productive capacity. We argue that the problem with these models is that they rely on the two conventional definitions of capacity which ignore these long-run expectations. Accordingly, we propose a third definition of capacity which incorporates these expectations and, thereby, corrects the problem. Furthermore, we argue that a correct, empirical analysis with the proposed definition -- indeed, any credible analysis of capacity or its utilization -- must take into account the demand for the output produced by the firms being studied. Finally, we apply the definition to clarify the meaning of surveys of capacity and, thus, show how it can be used to improve future surveys of capacity.

1. Introduction

Capacity has long been of interest to economists and public policy makers. Interest usually centers on the question: what output is attainable in a short period of time with given fixed resources?[1] For example, at the microeconomic level the purpose of defense mobilization analysis is to determine how much weaponry, ammunition, and other materiel can be produced under mobilization conditions. One of the earliest concerns of macroeconomists was to compute a measure of capacity output. Such a measure is of interest since once capacity is reached, increased demand for a product leads to an increase in its price. As another example, some recent work on capacity is motivated by the desire to improve methods for estimating multifactor productivity.[2]

In this paper we critically examine recent econometric models of capacity. Capacity constraints occur because one or more inputs are fixed in the short run. Usually, capital is considered to be the fixed input, although sometimes the labor input is also treated as fixed. Strictly, then, the subject of capacity overlaps substantially with capital theory and labor economics. To restrict the paper's size and to provide a sharp focus to the discussion, we only explicitly discuss a small part of this literature.[3]

Most econometric studies of capacity have proceeded with static models. Static models cannot fully capture the role of firms' expectations of future conditions in their current decisions. Moreover, the studies which have proceeded with static models have not fully taken advantage of the distinction between current and expected prices, even within the confines of a static analysis. The failure to make this distinction in models of capacity is an important limitation of their usefulness.[4]

The concern with inadequate specification of dynamic features in a model is a temporal specification problem. The problem of cross-sectional specification is, perhaps, an even larger topic than the problem of temporal specification. Thus, to keep the paper within reasonable bounds, on the matter of cross-sectional specification we shall limit our remarks to a problem closely associated with a temporal specification problem, the improper specification of variables as being exogenous.

For organizational convenience, the literature on capacity is divided into two branches: (i) studies which examine an array of production inputs with general production technologies, but which treat dynamics implicitly or not at all; and (ii) studies which explicitly consider dynamics, but focus on a small array of inputs and treat these with restrictive production technologies. Using this division, we heuristically discuss features which we think a model of capacity should have. We emphasize the critical importance of two features, the distinction between actual and expected prices of inputs and output and the need to treat output as endogenous by specifying a demand curve. As the structure suggests, a principal recommendation of the paper is the development of models which incorporate these features while simultaneously retaining the advantages of the static models. Such a research program is currently feasible and could produce substantially more convincing and useful economic models of capacity.

2. Economic Models of Capacity

Economists usually define, and correspondingly use, the notion of capacity output and capacity utilization in one of two ways.[5] Let -q(t) denote capacity output in period t in the sense of the maximum output which can be produced in the period with the given fixed inputs, and let q(t) denote actual output in period t. Then, in percentage terms, the first definition says that capacity utilization in period t, denoted by u(t), is defined by u(t) = q(t)/-q(t). Equivalently, capacity output is the output at which the short-run marginal cost production curve in period t, denoted SRMCC(t), becomes vertical.[6] A short-run is a sufficiently short planning horizon over which at least one input, usually capital, is fixed in the sense that it would be impractical to change its value within this horizon. A long-run is a sufficiently long planning horizon over which it is practical to vary all inputs to any desired values. An input is fixed (or quasi fixed) when time must pass and resources beyond purchase costs (adjustment costs) must be expended to change its value.

This first definition of capacity implies that output reaches capacity when substitutability between fixed and variable inputs is exhausted, or, equivalently, when marginal products of variable inputs fall to zero. The hallmark and difficulty of this definition is that it reflects extreme (e.g., wartime) and not to normal (peacetime) conditions. Most of the available economic data, to which this definition of capacity is applied, presumably reflect normal conditions in which all inputs are substitutable. Thus, some effort is required to make predictions for extreme conditions, in which fixed-to-variable-input substitutions have been exhausted, with inferences obtained with data for normal conditions. A further difficulty in applying this definition is that the econometric specifications of firms' production and cost functions are usually such that they imply SRMCCs which never become vertical.

These difficulties provide a justification for the second definition of capacity: capacity output is the output at which the short-run and long-run marginal cost curves intersect. Equivalently, capacity output is the output at which the short-run average cost curve is tangent to the long-run average cost curve.[7] Denote this second definition of capacity output by $\hat{q}(t)$. With this second definition of capacity output, capacity utilization is, then, defined by $u(t) = q(t)/\hat{q}(t)$.

This second definition represents an important step forward, because it involves the use of production technologies which allow any degree of substitutability -- as well as complementarity --

among the inputs. However, it also represents a shift of focus from associating capacity output with the maximum output which can be produced with fixed resources to associating capacity output with minimum-cost output under normal conditions. Each definition relates current output to a reference level of output, $\bar{q}(t)$ or $\hat{q}(t)$, which indexes the position of the SRMCC(t). That is, $\bar{q}(t)$ or $\hat{q}(t)$ serves as a measure, in terms of the horizontal (output) position of the SRMCC(t), of the available capital, whose fixity is the underlying source of capacity restrictions on output. Aside from allowing greater substitutability and complementarity, the second definition has the advantage that it emphasizes capacity as an economic concept. It views capacity as the result of the solution of a behavioral maximization problem and not simply as a technological phenomenon.

These common ways of defining capacity and capacity utilization are useful in situations where the planning horizon is short. Unfortunately, in many situations the planning horizon is sufficiently long to make it practical to vary capital. For such situations we introduce a third definition of capacity utilization, $u(t) = q^*(t)/\hat{q}(t)$, where $q^*(t)$ is the long-run expected or desired level of output, to be defined precisely below, and, as before, $\hat{q}(t)$ is the level of output at which the short- and long-run marginal cost curves intersect. The three definitions are illustrated in Figure 1 in the simplified case in which the long-run marginal- and average-cost curves are horizontal, hence, identical.

Figure 1

CAPACITY DEFINITIONS

1. $u(t) = q(t) \,/\, \bar{q}(t)$

"Engineering" - Focus on physical "maximum" output.

$\bar{q}(t)$ is maximum output.

2. $u(t) = q(t) \,/\, \hat{q}(t)$

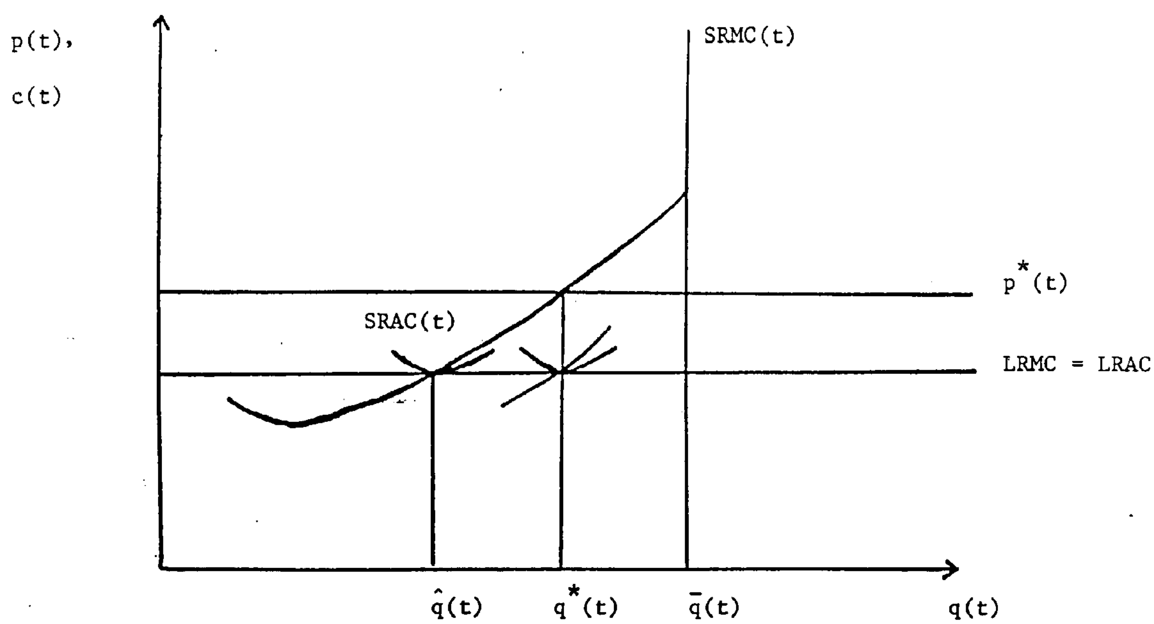
"Economic" - Focus on cost and capacity as an economic concept not simply a technological problem.

$\hat{q}(t)$ is the output at which SRMC = LRMC.

3. $u(t) = q^*(t) \,/\, \hat{q}(t)$

"Economic" - Focus on investment in capital and the importance of demand as well as cost.

$q^*(t)$ is the "long-run expected and desired" level of output.



The third definition has the following two advantages. First, it explicitly links capacity utilization to investment in capital -- to changes in actual capacity. Second, it provides a convenient way to illustrate several econometric problems in previous empirical work on capacity. Next, in order to precisely illustrate these points, we define and discuss notions of short- and long-run equilibrium.

Short- and Long-Run Equilibrium

For simplicity and without loss of generality for the present discussion, we assume the industry being studied is competitive. This means that firms can enter and leave the industry according to their long-run expectations of profitability and that firms take input and output prices as given. Moreover, we proceed in terms of a representative firm in the industry being studied and, thus, consider firm- and industry-level variables to be identical, except for differences in scale. Let $p_q(t)$ denote the price of output in period t and let k(t) denote the representative firm's capital stock in period t. We assume the current price of output, $p_q(t)$, is determined by a demand-supply equilibrium in the output market. The representative firm's current stock of capital, k(t), is determined by history and fixes the short-run marginal cost curve along which the firm operates.

The firm is said to be in short-run equilibrium (SRE) when it maximizes current (short-run) profits, $p_q(t)q(t) - p_l(t)l(t)$, with respect to $q(t)$ and $l(t)$, where $l(t)$ denotes the variable labor input and $p_l(t)$ denotes its price. In the maximization, the values of $p_q(t)$, $k(t)$ and $p_l(t)$ are taken as given and the restrictions of the production technology are imposed. The firm maximizes current profits -- is in SRE -- when it satisfies the condition $p_q(t) = SRMC(t)$, such that the position of the short-run marginal cost curve (SRMCC(t)) depends on the predetermined value of $k(t)$ and the exogenous value of $p_l(t)$. Since capital is fixed in this short-run maximization, its price, $p_k(t)$, does not figure in the calculations. If, as is usually assumed, there are no impediments to achieving the SRE, observations on output, inputs, and their prices are SRE values.

To define long-run equilibrium (LRE) values, we introduce price expectations. Let $E_t p_q(t+s)$ denote the (representative) firm's expectation at time t of the price of output at time t+s. Let $p_q^*(t)$ denote the firm's expected price of output. In particular, let $p_q^*(t)$ be a weighted average of the current price and expected future prices, namely $p_q^*(t)= \sum_{s=0}^{\infty} w_s E_t p_q(t+s)$, where the weights, w_s, are nonnegative and sum to one. The question of how firms forecast prices in an equilibrium setting is a deep and unsolved problem in economics. The major competing theories about this problem are the theories of adaptive expectations and rational expectations.[8] The implications of either adaptive or rational expectations can be captured by suitably restricting the w_s's and the forecasting mechanism, $E_t p_q(t+s)$. For simplicity, we assume that $w_0 = 0$ and that $E_t p_q(t+s)$ is the same for all s, so that $p_q^*(t)$ is simply the common value of $E_t p_q(t+s)$, for $s > 0$. We similarly define $p_l^*(t)$ and $p_k^*(t)$.

By working out an explicit dynamic optimization problem, it can be shown that future conditions become more important for the firm's current decisions as capital becomes more fixed. Capital becomes more fixed when its time-to-build or its adjustment costs increase. The greater importance of the future on current decisions manifests itself by the predominant mass of the distribution of the w_s's shifting to the right. In this regard, taking $w_0 = 0$ can be thought of as representing a high degree of capital fixity.

The LRE can now be defined in terms of the desired values of the variables under the firm's control. The LRE or desired values of output, capital, and labor are denoted by q*(t), k*(t), and l*(t). We also introduce the notion of long-run expected profits, defined as $p_q^*(t)q^*(t) - p_k^*(t)k^*(t) - p_l^*(t)l^*(t)$. Then, the firm is in LRE when its SRE values of output, capital, and labor, namely q(t), k(t), and l(t), are, respectively, equal to values of q*(t), k*(t), and l*(t) which maximize long-run profits. In the latter maximization, $p_q^*(t)$, $p_k^*(t)$, and $p_l^*(t)$ are taken as given and the restrictions of the production technology are imposed. The firm maximizes long-run expected profits when it satisfies the condition $p_q^*(t) = LRMC(t)$, where LRMC(t) denotes long-run marginal costs of production in period t. The position of the long-run marginal cost curve, denoted by LRMCC(t), depends on the given values of $p_k^*(t)$ and $p_l^*(t)$.

Thus, the firm is in a LRE when it satisfies the following three conditions: (i) the firm is in a SRE; (ii) actual input and output prices are equal to expected input and output prices; and (iii) the actual stock of capital, k(t), is equal to the desired stock of capital, k*(t).

Capacity Utilization and Long-Run Equilibrium

In the third definition of capacity utilization, $u(t) = q^*(t)/\hat{q}(t)$, the desired output, q*(t), is taken to be the long-run expected output of the firm. When current prices are equal to expected prices and the industry is in a long-run competitive equilibrium, this definition corresponds to the one commonly used in recent work.[9] Explicitly distinguishing between actual and expected prices as we have, may not be particularly important in some theoretical discussions, but it is crucial in an econometric analysis of capacity and related issues. A primary deficiency of many econometric examinations of capacity is that they fail to properly distinguish between actual and expected prices. Before stating this argument in detail, we conclude the discussion of firm behavior in our SRE/LRE framework, by specifying an investment decision rule.

Optimal Investment Decisions

Following the above discussion, in every period t and for given values of $p_q(t)$, k(t), and $p_l(t)$, the firm chooses values of

$q(t)$ and $l(t)$ which maximize its short-run profits. By definition, because $q(t)$ and $l(t)$ are variable, there are no impediments, in any given period t, to setting $q(t)$ and $l(t)$ in this way. Thus, the firm is in SRE in every period. In the case of investment in

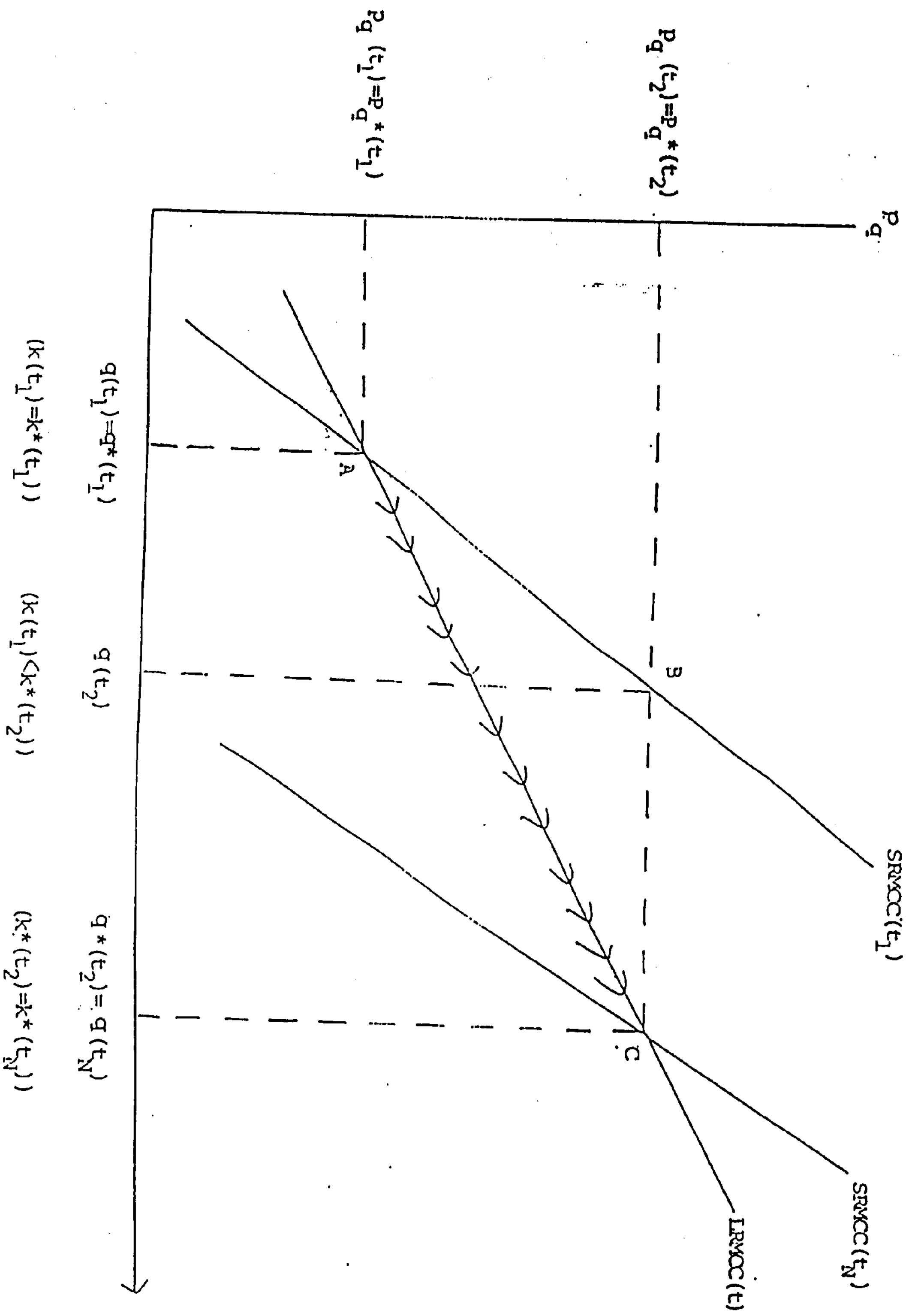

$p_d(t_1)=p_d^*(t_1)$
$p_d(t_2)=p_d^*(t_2)$
p_d
A
B
C
SRMCC(t_1)
SRMCC(t_N)
LRMCC(t)
$q(t_1)=q^*(t_1)$
$(k(t_1)=k^*(t_1))$
$q(t_2)$
$(k(t_1)<k^*(t_2))$
$q^*(t_2)=q(t_N)$
$(k^*(t_2)=k^*(t_N))$

Figure 2

capital, following standard practice, we assume that the firm adjusts fixed capital according to a flexible accelerator. That is, the firm sets $k(t+1) - k(t) = 1_k[k^*(t) - k(t)]$, where 1_k is a coefficient between zero and one. For simplicity, the depreciation of capital is ignored, so that $k(t+1) - k(t)$ denotes gross as well as net investment.

Used in this fashion, the LRE is simply a device for capturing in a simple way the implications of the solution of a proper dynamic specification of firm behavior. Such a specification is one in which the firm's decisions on all variables -- whether variable or fixed -- are given by the solution of the dynamic optimization problem, e.g., the maximization of expected present value.

In this view, 1_k is generally a function of all of the firm's structural parameters, although, it depends most critically on adjustment-cost and (or) time-to-build parameters which describe the degree of capital fixity. In this discussion, for simplicity, 1_k is taken to be a structural parameter in its own right. When capital is highly fixed (highly variable), the firm slowly (quickly) adjusts capital to its target value and 1_k is close to zero (one).[10]

3. Importance of Expectations

We are now in a position to show the importance of accounting for expectational effects in the analysis of capacity. There are two points to the argument. The first point is that when

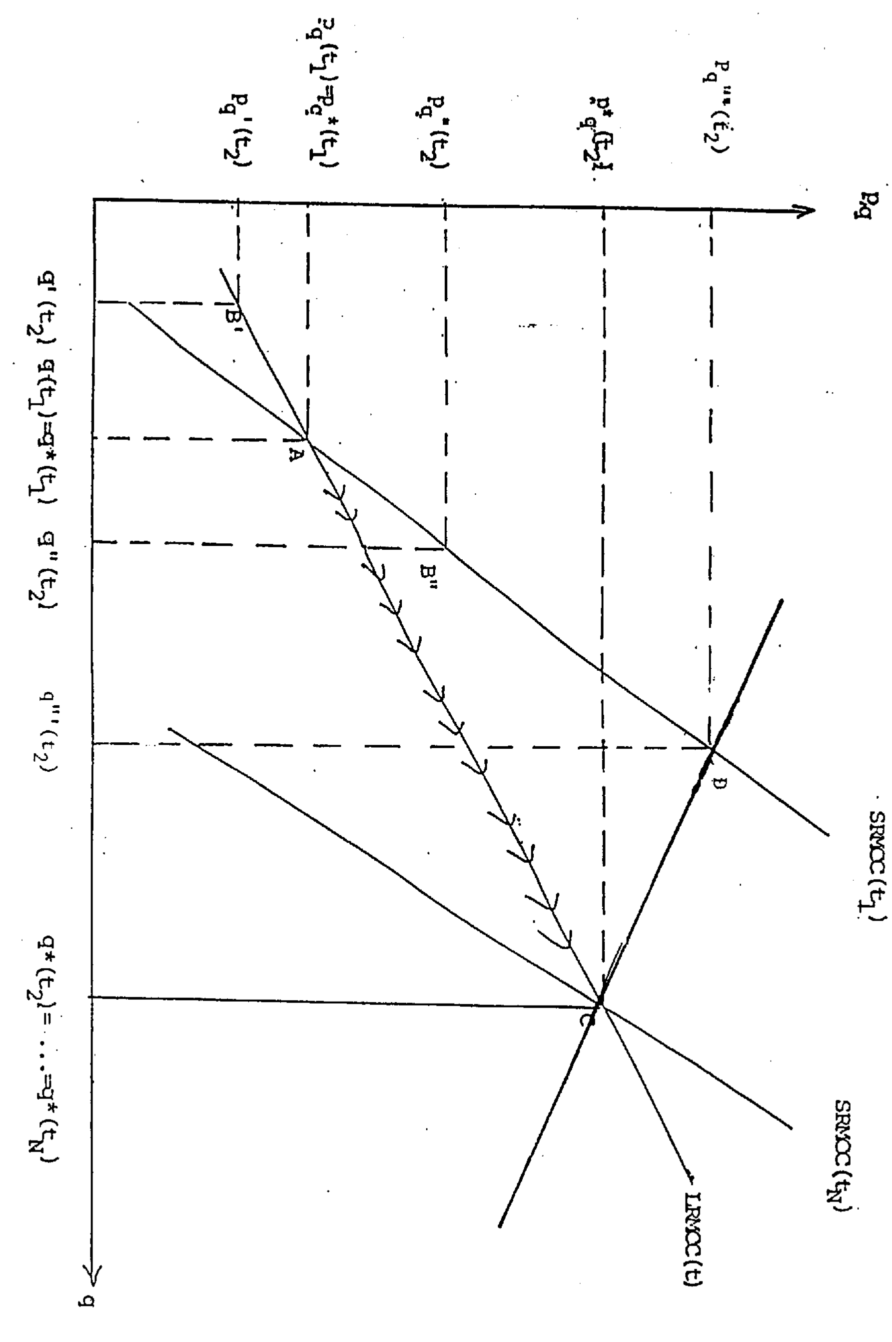


Figure 3

expectations are not incorporated in economic models of capacity, deviations between actual and expected prices will lead to inconsistencies between observed behavior and the behavior predicted by the model. This point is demonstrated by showing that a textbook account of the relationship between capacity utilization and investment leads to erroneous conclusions when the role of expectations in firms' decisions are ignored. As part of this discussion we also illustrate the second point of the argument, that in practice, actual and expected prices are likely to deviate. After explaining why actual and expected prices are likely to be different, in response to unexpected exogenous shocks, we conclude with a short section which argues that because of the way expectations are formed from observed prices, observed and expected prices are unlikely to be equal even when the LRE price is constant.

Capacity Utilization and Investment Behavior

To begin, we consider Figure 2. In this figure, the vertical axis measures the price of output and the horizontal axis measures the quantity of output. In addition, in parentheses, the horizontal axis indexes the stock of capital. The time frame of the analysis is $t = t_1, t_2, \ldots, t_N$, where t_1 is an initial period in which the (representative) firm is in an initial LRE and t_N is a final period in which the firm reaches a new LRE. The movement from the initial to the new LRE can be thought of as being caused by an exogenous shift to the right in the demand curve for the output of the industry. The location of each SRMCC(t) in Figure 2 depends on values of the current stock of capital k(t). It also depends on the production technology and the current price of labor, $p_l(t)$, although, these are held fixed throughout the analysis. SRMCC(t_1) and SRMCC(t_N) are the initial and final SRMCCs. The LRMCC(t) is fixed throughout the analysis. Its shape and position depend on the given production technology and on the given prices of capital and labor. In order to focus on the output decisions of the firm, actual and expected prices of inputs are assumed to be identical and to be fixed throughout the analysis. For example, in the case of capital, we assume that $p_k(t_1) = p_k^*(t_1) = . . . = p_k(t_N) = p_k^*(t_N)$.

In Figure 2, point A reflects the initial LRE in which actual and expected values are equal, $p_q(t_1) = p_q^*(t_1)$, $q(t_1) = q^*(t_1)$, and $k(t_1) = k^*(t_1)$. We assume that in the next period, t_2, the actual and expected prices of output increase by the same amount to $p_q(t_2) = p_q^*(t_2)$. The firm's immediate reaction is to increase its output from $q(t_1)$, associated with SRE point A, to $q(t_2)$, associated with SRE point B. However, B is not a LRE because at B the actual stock of capital is still at is previous value, $k(t_1)$, and, in response to the new permanent price increase, $p_q^*(t_2) - p_q^*(t_1)$, the desired stock of capital is now $k^*(t_2)$, which is greater than the actual and desired, initial, capital stocks, $k(t_1) = k^*(t_1)$. Consequently, the firm begins to invest in capital and continues to do so until it reaches point C in period t_N. The arrowheads along the LRMCC(t) reflect the steady shift to the right of the SRMCC(t) as the capital stock is increased from $k(t_1) = k^*(t_1)$ to $k(t_N) = k^*(t_N)$. By assumption, throughout the adjustment periods actual and expected prices remain at the same values, namely $p_q(t_2) = p_q^*(t_2) = \ldots = p_q(t_N) = p_q^*(t_N)$. Thus, at the new LRE point C, $p_q(t_N) = p_q^*(t_N)$, $q(t_N) = q^*(t_N)$, and $k(t_N) = k^*(t_N)$.

Investment and Conventional Capacity Utilization Measures

We now relate this transition in Figure 2, from LRE point A to LRE point C, to the conventional measure of capacity utilization $u(t) = q(t)/\hat{q}(t)$. Recall that $\hat{q}(t)$ denotes the capacity value of output in the sense of the intersection of the current SRMCC(t) and the LRMCC(t). This is the received economic definition of capacity.[11] According to this view, $u(t) > 1$ indicates a positive rate of investment in capital because there is "over-utilization" of capacity, $u(t) = 1$ indicates no change in the capital stock because capacity is being "optimally" utilized, and $u(t) < 1$ indicates disinvestment in capital because there is "under-utilization" of capacity, i.e., there is some unneeded capital. This standard argument is completely consistent with the situation in Figure 2 in which actual and expected prices are equal.

We now turn to Figure 3, which differs from Figure 2 only in that, after the initial LRE, the actual price of output is allowed to differ from the expected price of output. In fact, it is extremely unlikely that the actual price will equal the expected price. Thus, actual price in period t_2 could be $p_q'(t_2)$, $p_q''(t_2)$, or

$p_q'''(t_2)$, even though the expected price has increased, as in Figure 2, to $p_q^*(t_2) = ... = p_q^*(t_N)$. In the first case, where $p_q(t_2) = p_q'(t_2)$, $u(t_2) = q'(t_2)/\hat{q}(t_2) = q'(t_2)/q(t_1) < 1$ incorrectly indicates disinvestment. In the second case, where $p_q(t_2) = p_q''(t_2)$, $u(t_2) = q''(t_2)/\hat{q}(t_2) = q''(t_2)/q(t_1) > 1$ correctly indicates a positive rate of investment. However, the correct indication of investment behavior by u(t) in this second case is pure happenstance. The actual price of output in period t_2 could just as well have been below the initial value, at a point like $p_q'(t_2)$. In fact, $p_q(t_2)$ could just as well be above the new LRE value of $p_q^*(t_2)$ e.g., $p_q''(t_2)$.

In sum, the conventional measure of capacity bears no systematic relation to investment activity, because investment activity depends mostly on expected future conditions. By contrast, the conventional measure of capacity depends in large measure (through the appearance of actual output in the numerator) on the transitory component of current conditions. The conclusion is that it is critical to account for expectational effects in the analysis of firm behavior when

some input (capital) is fixed. In fact, this conclusion is a general one. Expectational effects will be significant in any economic situation in which there is some sort of "fixity," "friction," or "inertia."

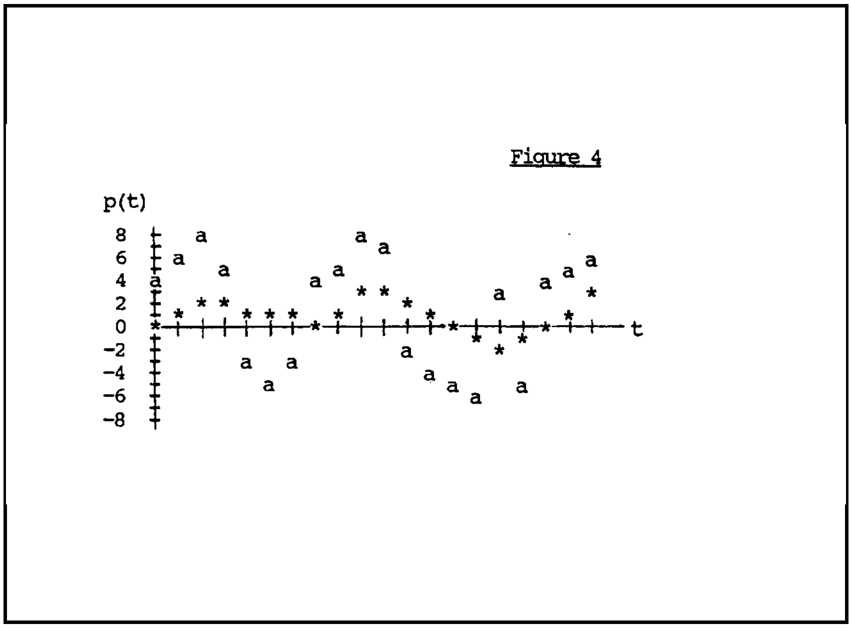
<u>Capacity Utilization Adjusted for Expectations</u>

It is this critique which leads to a third definition of capacity utilization, $u(t) = q^*(t)/\hat{q}(t)$. Compared with the standard definition, this third definition replaces output, q(t), in the numerator of u(t) with expected long-run output, $q^*(t)$. Accordingly, transient components of q(t), which are irrelevant to investment decisions, are stripped away. With this redefinition, u(t) will correctly indicate investment behavior according to the conventional rule. For example, as in Figures 2 and 3, after the initial LRE in period t_1, u(t) continues to be > 1, thereby indicating a positive rate of investment, until the new LRE is reached, and this is the case regardless where actual prices happen to fall, hence, regardless of where actual output happens to be on the SRMCC(t).

<u>Relation Between Actual and Expected Prices</u>

As we have illustrated in Figures 2 and 3, actual price movements reflect movements in their systematic and random components. Below, in Figure 4, we further illustrate the relation between these components with a representative time plot. Before proceeding with this illustration, however, we note that systematic variations in LRE price as described in Figures 2 and 3 will also cause a divergence between actual and expected prices. For example, abstracting from random (transient) fluctuations an outward shift in the expected demand for the output of the industry causes the price-quantity path to overshoot the new LRE value. This will occur because of the interaction of a positively sloped industry supply curve with a negatively sloped industry demand curve. For example, in Figure 3, if the line DC represents the new expected industry demand curve, then, the transitional price-quantity path will be ADC.

As before, let $p_q^*(t)$ be the common value of forecasts of the price of output. Suppose that $p_q^*(t-1)$ is an optimal, unbiased, one-step-ahead forecast of $p_q(t)$. Let $,_q(t) = p_q(t) - p_q^*(t-1)$ be the one-step-ahead forecast error. Equivalently, $p_q(t) = p_q^*(t-1) + ,_q(t)$, which is usefully interpreted as a decomposition of the actual price, $p_q(t)$, into a permanent component, $p_q^*(t)$, and a transitory component, $,_q(t)$. This interpretation is appropriate because, by their very nature, optimal forecasts are "smoothed" estimates of actual values and, therefore, fluctuate less widely and rapidly than the actual values which they are forecasting. This property is illustrated in Figure 4, in which "a" and "*," respectively, denote typical actual and expected price paths. A look at the figure shows that an observed or actual price is just



as likely to be above the expected price as to be below it.

4. Expectations and Survey Data

We have shown the importance of expectational effects, if capacity utilization is to be used as an indicator of investment activity. Expectations are similarly important in the analysis of capacity utilization data obtained directly from surveys, i.e., not manufactured with an economic model estimated with related observations.

The U.S. Bureau of the Census obtains survey responses on "actual," "practical," and "preferred" outputs of plants in a specified reporting period.[12] The instruction form defines actual output as the actual level of production. Preferred operations are those which "you would prefer not to exceed." The form then states that underlying this definition is a level of operations at which profits are maximized, namely the level where marginal revenue equals marginal costs. Following this the form states that "preferred capacity may equal but not exceed practical capacity," where the latter is defined as the "maximum" level of production that this establishment could reasonably expect to attain."

The meaning of "actual" output is universally understood. The quality of the obtained responses about actual output, thus, depends only on the accuracy of the respondents' accounting procedures and on their response rates in the survey. In the case of "practical" and "preferred" outputs there appears to be confusion about the meaning of these terms. In fact, it is not entirely clear -- especially in light of the previous discussion -- how "practical" and "preferred" outputs differ from themselves and from actual output. If firms are setting output every period by maximizing current profits, and there are no impediments to doing so, then, it would seem that actual, practical, and preferred outputs are identical. Indeed, some of the survey responses support this interpretation because in some instances the respondents give the same value for all three types of output. From a "short-run" viewpoint, it is natural to consider actual and preferred outputs to be identical, because, by the definition of the SRE, the firm is setting its actual output every period so as to maximize its current (short-run) profits. The extended description of "practical" output in the questionnaire conforms with what we defined as maximum output, -q(t). More than 50 percent of respondents equated "practical" and "preferred" output, although most respondents reported actual output below practical and preferred output. Thus, most respondents indicate that their actual operations are not at a profit maximizing level. This suggests that the preferred and practical concepts might be based on "long-run" considerations.

The primary role of our analysis is to clarify concepts. We showed that there is a useful distinction to be made between actual and expected prices and quantities. To the extent that "practical" output is meant to reflect the firm's current capacity constraints, i.e., the amount of fixed capital, then, capacity output in the sense of $\hat{q}(t)$ seems to be a candidate for making this notion precise. Similarly, if "preferred" is understood in the sense of the long run, i.e., where capital fixity is surmounted, then, capacity output in the sense of $q^*(t)$ seems to be a good candidate for making the idea of "preferred" output precise.

By clarifying concepts, one also clarifies their practical significance. One might conclude, e.g., following the above discussion, that there is a need for a concept like "preferred" output in the sense of expected long-run output, but that there is no need, in addition, for a notion of "practical" output. Once such clarifications are made, they will be useful for redesigning survey questionnaires. If, indeed, "preferred" output is to be understood in our sense of "expected" output, then, the role of long-run considerations and expectations about these considerations should be strongly emphasized in the definition of the concept in the survey questionnaire. Obviously, it is counterproductive to use specialized terminology, but commonplace words like "long-run" and "expected future," which are readily understood, would be an improvement over current descriptions.

In sum, we believe that the economic analyst has a major role to play in making such clarifications and that to do this properly one has to use a carefully articulated dynamic model. The heuristic model discussed in section 2 is simply a first step in this direction.

5. Expectations and Endogeneity

As we noted in the introduction, issues of temporal specification, such as the problem of treating expectations, are often thought of as being logically separate from cross-sectional specification problems. However, in this section, we note how, in one important respect, the problems of temporal and cross-sectional specification are intimately related. In particular, we note that the econometric problem of adequately treating expectations about the price of output in the industry being studied dictates treating output as being endogenous and necessitates bringing the demand curve for output into the analysis.

In general, the endogeneity or exogeneity of a variable depends on the level of cross-sectional aggregation of the data used in the analysis. For concreteness and clarity, we shall proceed in this discussion on the assumption that, as usual, the available data represent a higher-level aggregate such as a 2-digit SIC industry, not a lower-level aggregate such as a firm or plant. As in the previous discussion, we assume the industry in question is competitive. Thus, to individual plants and firms in the industry, all prices -- input as well as output prices -- are exogenous, i.e., are taken as given. However, since the data on output prices and quantities pertain to the industry as a whole and since the industry's output market determines these prices and quantities in a supply-demand equilibrium, the data on prices and quantities of this industry must be taken to be endogenous to the analysis. To do so, one must add a demand curve for the industry's output to the analysis.

In many econometric analyses of capacity and related issues, the price and quantity of output of an industry are both -- implicitly or explicitly -- incorrectly treated as exogenous.[13] Of course, the price elasticity of demand for the output may be close to zero or infinity, so that, respectively, the quantity or price of output of the industry is effectively exogenous. However, both the price and quantity of output cannot properly be simultaneously treated as exogenous.

It is more appropriate to assume that input prices are exogenous to the industry as a whole, hence, to the individual firms and plants in the industry, than to assume that the output price of the industry is exogenous. This is a more reasonable assumption, because it is likely that the prices of investment goods purchased and labor hired by the industry are determined in a wider market -- e.g., a national or even an international market -- of which the industry being studied is a small part.

Econometric Treatment of Expectations

The question of endogeneity-exogeneity of input and output prices is critically tied to the econometric treatment of expectations, i.e., to the way in which the notions of expected prices which we have introduced are quantitatively tied down to the actual data. In fact, the econometric treatment of expectations is completely different according to whether the price in question is exogenous or endogenous.

For prices which may properly be considered exogenous, all that is needed are some reasonable, simple, forecasting rules which track the data well. For example, one can develop one-step-ahead forecasting equations with routine time series methods[14] and, as in our previous discussion, one can, then, identify these forecasts with "expected" prices.[15]

When output prices and their quantities are endogenous, one must resort to other, more complicated, methods. One method is to proceed under the assumption of rational expectations. An output demand curve must be introduced into the analysis in order to make the rational expectations method operational. The technical details of how one proceeds according to the rational expectations hypothesis are too complex to be described here,[16] however, their essentials are easily conveyed in words.

The essence of the rational expectations approach is that it emphasizes that current prices and quantities -- and expectations of future values of prices and quantities -- must be consistent with market clearing, i.e., the condition that supply equals demand. Thus, in the case of the output price and quantity of an industry, this approach eliminates the arbitrariness -- and resulting biases -- of assuming that price and output are exogenous. The rational expectations approach uses fundamental economic concepts -- supply curves, demand curves, and market clearing -- to tie together expectations about prices in a market and the actual prices generated by the market.

The rational expectations approach has been severely criticized in some quarters.[17] There has, however, been much misunderstanding about what rational expectations means, how it should be implemented econometrically, and how one should decide its applicability in a given situation. One criticism which has been levelled against existing rational expectations models is that they unrealistically assume that prices and quantities fluctuate so as to instantaneously equilibrate markets. This criticism is legitimate. In fact, most rational expectations models of industry behavior have been strongly rejected by the usual statistical criteria. Undoubtedly this has in large measure been due to the assumption that prices equilibrate markets instantly. However, rational expectations and slow equilibration of markets due to, e.g. contractual rigidities in prices are not inconsistent.[18]

In sum, we believe that introducing expectations and operationalizing these expectations with the demand curve of the representative firm is an important and feasible next step in the development of econometric models of capacity.

ENDNOTES

1. Capacity is also of interest in the study of firms' strategic behavior. For example, it has been argued that incumbent firms in an industry hold excess capacity to deter entry into the industry by new firms. If firms do this then, in some sense, they are not minimizing costs of production in order to have a credible threat against possible entrants. However, we shall not be directly concerned with this literature, since it is tangential to our discussion.

2. See e.g., Berndt and Fuss (1982).

3. For the reader's benefit we have listed numerous articles in the bibliography, at the end of the paper, which we do not explicitly cite, either in the text or in the notes, but which are significant contributions to the study of capacity. A good starting point in the literature on capacity is the survey article by Winston (1974).

4. Some recent work, notably by Morrison (1985b, 1986), incorporates nonstatic expectations about exogenous output demand and input price variations. Unfortunately, treating the representative firms's output as exogenous is likely to result in simultaneous-equations biases in estimation. Input prices are, however, more likely to be exogenous.

5. Schnader (1984) provides a useful introduction to the definition of capacity and capacity utilization and discusses the leading sources of observations of capacity and capacity utilization. Christiano (1981) covers the same material but does so more thoroughly.

6. The fixed-proportions production function gives the simplest illustration of this definition of capacity output. Let production be governed by $q(t) = \min\ [\beta_k k(t),\ \beta_l l(t)]$, where β_k and β_l are positive fixed coefficients, capital $k(t)$ is the single fixed input, and labor $l(t)$ is the single variable input. Then, since $l(t)$ is variable, capacity output is proportionate to the available amount of capital, $-q(t) = \beta_k k(t)$. Let $p_l(t)$ be the price of labor in period t. Also, let SRMC(t) denote the short-run marginal cost of production in period t. Then, $\mathrm{SRMC}(t) = p_l(t)/\beta_l$ when $0 < q(t) < -q(t)$ and $\mathrm{SRMC}(t) = +4$ when $q(t) = -q(t)$. These two equations say that short-run marginal and average costs are identical and constant with respect to output up to capacity output, at which point they both become infinite.

7. Berndt, Morrison, and Wood (1983) trace this definition of capacity to Klein (1960) and adapt it to their work. Stigler (1966, pp.155-158) defines capacity output as that output at which short- and long-run marginal costs are equal. Equivalently, Berndt and Fuss (1982) and Berndt (1984) define capacity output as the level of output at which the short- and long-run average cost curves are tangent. Note that under conditions of perfect competition, capacity output will be at the minimum point of the long-run average cost curve. When long-run average and marginal cost curves are horizontal, this definition reduces to capacity output being defined as the output at which the short-run average cost curve is minimized.

8. Briefly, adaptive expectations forecasting rules are forecast updating rules which can be rationalized with time series models of the sort treated by Box and Jenkins (1976). Their hallmark is that they are simple to use. The hypothesis of rational expectations does not similarly propose a set of forecasting rules. Rather, it proposes methods for treating expectations formation in market-clearing equilibrium settings. The hallmark of a rational expectations equilibrium in a model is that the agents have unbiased estimates of future prices in the market which is being analyzed, but at the same time these expectations are consistent with market clearing -- that supply equals demand. By contrast, adaptive expectations rules are generally inconsistent with market clearing. The principle articles about the econometrics of rational expectations have been collected in the two volumes edited by Lucas and Sargent (1981). The articles by Lucas in (1981c) are also central to this literature.

9. See, e.g., the work cited in note no. 7.

10. Strictly speaking, q(t) and l(t) should also be slowly adjusted to target levels according to their own flexible accelerators. That is, q(t) and l(t) should be adjusted over time by $q(t+1) - q(t) = 1_q[q^*(t) - q(t)]$ and $l(t+1) - l(t) = 1_l[l^*(t) - l(t)]$, where 1_q and 1_l are parameters between zero and one. The reason why, strictly speaking, $1_q < 1$ and $1_l < 1$, is because interactions among inputs and output in the production technology lead to a "spill over" of the fixity of capital onto variable output and labor input. Moreover, it is this "spill over" which causes expected -- not just current -- prices of output and labor to be pertinent in the LRE, even though output and labor are individually variable. Since q(t) and l(t) are individually variable, 1_q and 1_l can be expected to be close to one,

especially when compared with 1_k. This is why the assumption made above, that q(t) and l(t) are set according to short-run profit-maximization (which is equivalent to setting $1_q = 1$ and $1_l = 1$), is a reasonable first approximation.

11. E.g., Berndt (1984, pp. 15-17).

12. The Census Bureau's "Survey of Plant Capacity" is described in the U.S. Department of Commerce publication listed in the references. This publication includes the questionnaire for the 1985 survey as well as definitions of terms.

13. This assumption is present in most empirical work, even in recent extensions of the literature which include dynamics explicitly. See, e.g., Morrison (1985b).

14. See, e.g., Box and Jenkins (1976).

15. See, e.g., Morrison (1985b, 1986).

16. See, e.g., Lucas and Sargent (1981).

17. See, e.g., Tobin's (1980) criticism and Lucas' (1981d) response to this criticism.

18. See, e.g., Taylor (1980).