The (Normal) Rate of Capacity Utilization at the Firm Level

by

Michalis Nikiforos*
Levy Economics Institute of Bard College

November 2012

*This paper is part of my recently completed Ph.D. dissertation at the New School for Social Research (NSSR). An earlier version of it has appeared in the working paper series of the economics department of the NSSR (16/2011). I would like to thank Duncan Foley, Peter Skott, Lance Taylor, Luca Zamparelli, Laura Barbosa de Carvalho, Christian Schoder, and Jonathan Cogliano as well as the participants in the 38th Annual Conference of the Eastern Economic Association in New York and the 15th Conference of the Research Network Macroeconomics and Macroeconomic Policies (FMM) in Berlin for useful comments and suggestions. The usual disclaimer applies. Financial support from the Greek State Scholarships Foundation is gratefully acknowledged.

The Levy Economics Institute Working Paper Collection presents research in progress by Levy Institute scholars and conference participants. The purpose of the series is to disseminate ideas to and elicit comments from academics and professionals.

Levy Economics Institute of Bard College, founded in 1986, is a nonprofit, nonpartisan, independently funded research organization devoted to public service. Through scholarship and economic research it generates viable, effective public policy responses to important economic problems that profoundly affect the quality of life in the United States and abroad.
ABSTRACT

This paper examines the endogeneity (or lack thereof) of the rate of capacity utilization in the long run at the firm level. We provide economic justification for the adjustment of the desired rate of utilization toward the actual rate on behalf of a cost-minimizing firm after examining the factors that determine the utilization of resources. The cost-minimizing firm has an incentive to increase the utilization of its capital if the rate of the returns to scale decreases as its production increases. The theory of economies of scale provides justification for this kind of behavior. In this manner, the desired rate of utilization becomes endogenous.

**Keywords**: Kaleckian; Long Run; Economies of Scale; Utilization

**JEL Classifications**: B20, B50, D21, E11, E12, E22, E25
INTRODUCTION

Capital utilization is an important—though usually neglected—economic variable. Many studies show that most firms keep their capital idle most of the time.\textsuperscript{1} It is thus worthwhile to understand the factors that determine the level of capacity utilization at the firm level. Two important studies on the theory of utilization of the firm are Robin Marris’s *The Economics of Capital Utilisation* (1964) and Roger Betancourt’s and Christopher Clague’s *Capital Utilization* (1981). Many of the insights of the present paper draw on these seminal contributions. Aside from this micro level, the concept of capacity utilization appears in various areas of economic research such as the economic cycle theory (e.g in the Real Business Cycle school), growth accounting, development economics, and taxation theory and policy.

Within the non-neoclassical paradigms of economic thought, capacity utilization is featured in recent debates because of its central role in what is often referred to as the Kaleckian model of growth and distribution.\textsuperscript{2} Within the Kaleckian analysis capacity utilization bears much of the weight of adjustment towards equilibrium. The Kaleckian framework has become increasingly popular because of its simplicity and its ability to accommodate different views and approaches. As is well known, apart from the standard Keynesian result that an increase in the propensity to save will lead to lower utilization of capacity and lower growth (what is usually called “the paradox of thrift”), the Kaleckian model, depending on the elasticities of the components of demand to distribution and utilization, can give rise to two different regimes: a wage-led and a profit-led regime. In the former an exogenous shift of distribution against profits leads to an increase in utilization and growth. The opposite happens in the latter.\textsuperscript{3}

The Kaleckian model has been criticized at many levels, like the specification of its investment function (see Skott, 2008a,b) and its pricing and distribution theory (among others Steedman, 1992). However, the most persistent critique is related to its inability to *equate in the long run* the actual rate of capacity utilization (the equilibrium rate of utilization resulting from solving the model) with the exogenously given desired—or normal or planned—rate (the utilization rate that a firm chooses based on a profit-maximization or cost-minimization process). Committeri (1986, p.170), referring to the contributions of Rowthorn (1981) and Amadeo (1986), writes that “there is the possibility of utilization being *different* from its normal degree, even in states of equilibrium (and indeed, actual and normal utilization would coincide only by a mere fluke).”

\textsuperscript{1}These studies include Foss (1963, 1981a,b, 1984), Taubman and Gottschalk (1971), Orr (1989), Shapiro (1986) and Beaulieu and Mattey (1998).

\textsuperscript{2}The fundamental ideas of the Kaleckian model of growth and distribution go back to the writings of the classical political economists, John Maynard Keynes, and, of course, Michal Kalecki (e.g.1971). In its contemporary form it has been developed by Joseph Steindl (e.g. 1952), Rowthorn (1981), Taylor (1983, 1990, 2004), Dutt (1984, 1990), Amadeo (1986), Bhaduri and Marglin (1990), and Kurz (1990). The onomatopoeia of the model also includes names such as Structuralist or Post-Keynesian.

\textsuperscript{3}A discussion of the Kaleckian model and the implication of non-linearities in distribution for the definition of wage- and profit-led economies can be found in Nikiforos and Foley (2012).
Auerbach and Skott (1988, p. 52) claim that “it is inconceivable that utilization rates should remain significantly below the desired level for any long period.”

The rationale behind this inconceivability is provided by Heinz Kurz (1986) in a paper clearly located within this debate. Quoting Sraffa (1960) he argues that the normal rate of utilization “will be exclusively grounded on cheapness.” In other words, each firm will choose its level of utilization based on the principle of cost minimization. He then explores how such a firm will determine its normal rate of utilization. Kurz uses the example of a firm that can produce a certain amount of output either by employing a certain amount of capital and labor in a single-shift mode of operation, or by employing a second shift and using only half the capital. The wage of labor in the second shift is higher than in the first because of social norms and conventions (e.g. the wage premium earned by workers who work during the night). The firm will choose the number of shifts—and thus the level of utilization of capital—that is more profitable (or less costly). Under Kurz’s setup the normal utilization rate will only change in response to technological changes or changes in the cost of labor and capital and the relative cost of labor between the two shifts, and does not respond to changes in demand. Using the classical metaphor of center of gravitation, he argues that demand causes gravitation around this center and in fact “it cannot be precluded that deviations of the actual situation from the ‘normal’ one, may become large, and remain so for a long period of time” (p. 40). However, the center of gravitation itself is not affected by demand forces.

In response to this critique, the Kaleckian side has argued in favor of an endogenous desired level of utilization in the long run; it is the desired level of utilization that converges towards the actual rate and not the other way around. However, they fall short in explaining how this endogeneity squares with a firm that minimizes its cost. In fact, the usual argument (found among others in Lavoie et al., 2004) is that the level of utilization is not the result of a cost-minimizing decision process on behalf of the firm, but rather a convention that can change in response to changes in demand. The argument that the utilization rate of the firm is merely a convention is not convincing. As we will show later, the level of utilization is one of the most important decisions for a firm.

In the present paper we examine how a firm that minimizes its cost chooses the utilization rate of its capital. The setup is similar to that of Kurz. The analysis is static and the firm can choose between a single- and a double-shift system of production. We start with a technology of only one technique of production (as in Kurz), the inclusion of additional techniques allows

---

4 The paper was published in *Political Economy*, the same year as the papers of Amadeo and Committeri mentioned above. Kurz explores the concept of the normal rate of utilization.


6 A detailed examination of the debate between the two sides is beyond the scope of this paper. Such a review can be found among others in Nikiforos (2011).
us to derive more results. Unlike Kurz we examine the role of economies of scale in the choice of the utilization level. Finally, another difference in our analysis is that we do not allow for any kind of “reswitching.”

The conclusions we make about the role of technology and the cost of capital and labor in the choice of the system of production are similar to those of Kurz, with one important exception. The economies of scale allow for an endogenous desired rate of utilization. We show that the firm will tend to utilize its capital more—adopt a double shift system—as the demand for its output grows, if the rate of the returns to scale decreases. We argue that the theory of economies of scale can provide justification for this kind of behavior of economies of scale.

In the next section we examine some early writings on the utilization of capital. In sections 3 and 4 the main principles behind the choice of the utilization of capital are stated. It is explained why this choice is similar in nature to the choice of the technique of production. In section 5 we lay down a formal model, at first with one technique of production (like Kurz), then with two, and finally with infinite techniques. In section 6 we show formally why the firm will tend to utilize its capital more as the demand for its output grows, if the rate of the returns to scale decreases. Finally, in section 7 we discuss if this condition is something more than a theoretical extravaganza. We resort to the theory of the economies of scale and show that there are reasons to believe that economies of scale behave in such a way.

EARLY WRITINGS ON THE UTILIZATION OF CAPITAL

Karl Marx was among the first to deal with the issue of the utilization of capital. In *Capital* he provides the main insights that still form the backbone of the theory of capital utilization. He analyzes the working day of capital within the framework of the labor theory of value and his theory of exploitation. In chapter 10 of Volume I Marx argues that since the (fixed) capital exists in order to “absorb labor and, with every drop of labor, a proportional quantity of surplus labor...Capitalist production drives, by its inherent nature, towards the appropriation of labor throughout the whole of the 24 hours in the day. But since it is physically impossible to exploit the same individual labor-power constantly during the night as well as the day, capital has to overcome this obstacle. An alteration becomes necessary, between the labor-powers used up by day and those used up by night”(Marx, 1976, p.367).

Thus, the shift system is the means which allows the capitalists to appropriate surplus value throughout the 24 hours of the day. Of course Marx was not so naive as to believe that factories would work for 24 hours per day. To paraphrase the title of the fifth chapter of Volume III (Marx, 1981, p.170), the shift system was a way for the capitalist to economize in the use of constant capital. However, (as he mentions in the first paragraph of the same chapter) “The lengthening of the working-day [thus] raises profits even if overtime is paid, and up to a certain point
even if overtime is paid at a higher rate than normal working hours.” In this sentence Marx introduces what we will call later the utilization differential, a wage premium that is paid to workers who work over the “normal” working hours and is the cornerstone of the modern theory of utilization. Moreover it becomes clear that, for Marx, the goal of the capitalist is not the economy of fixed capital (the high utilization rate) per se, but rather the maximization of profits. Thus, a careful interpretation would lead to the conclusion that according to Marx the working day of capital is determined by the capitalists—given the established norms of the working day and the technology—in their effort to maximize the surplus value or the rate of the surplus value.

The working day is a recurrent theme in Marx’s work, always in relation to his theory of production of surplus value.7 However, the distinction between the working day of capital and labor is not always explicit.

John Stuart Mill in his Principles of Political Economy (Mill, 1864, ch. IX, p.176) argues that “the only economical mode of employing” the machines is to keep them “working through the twenty-four hours.”8 Kurz (1986, p.42) interprets this statement as Mill arguing that the firm under the profit maximizing principle will run its machines around the clock and therefore “Mill’s opinion cannot be sustained.” It is more probable that the meaning of this short comment by Mill is that the machinery is wasted if it is not utilized and this waste is non-“economical.” A century later Georgescu-Roegen (1972, p. 284) stresses—along the same lines—that capital “idleness is the worst form of economic waste and a great hamper to economic progress.”

Finally, the young Alfred Marshall, in the essay The Future of the Working Classes (1925), argues that society would benefit from the “diminution of hours of manual labor.” He recognizes that “for every hour, during which his untiring machinery is lying idle, the capitalist suffers loss” (p. 113). The solution he proposes is the adoption of the multiple shift system, so that both the working day of labor is short and the capital does not sit idle. Note that the idea of idle capital as a form of economic waste is also present in this argument of Marshall. In his last book, Industry and Trade (1920) he reflects on British society and the “place of Britain in the world” after World War I. He argues in favor of extracting “as much work as possible out of his plant.” Again his proposal is “to work in shifts, so as to keep the plant at work for twice as long as the normal working day, then wages will be raised automatically far above their present level.”

7For example, in chapter 15 of Volume I of Capital he examines the working day in relation to the introduction of machinery in production. In the same chapter (p. 533-544) Marx deals with the speed of production, which as we will see below, is another constituent of the modern theory of utilization of capital.

8In this chapter Mill discusses the “Production on large and production on a small scale.” He comments on a passage of Charles Babbage’s On the economy of machinery and manufactures (1832, p.174) and explains that the larger the scale of production “the further the division of labor may be carried.”
RECENT CONTRIBUTIONS TO THE THEORY OF CAPITAL UTILIZATION

The insights provided in these early writing have become more explicit in recent contributions, which examine the determination of capital utilization.

The problem of the choice of optimal utilization on behalf of a firm is similar in nature to the more familiar choice of the capital-labor ratio. For example, the choice of the optimal capital-labor ratio is related (among others) to the trade-off between increasing labor productivity and decreasing capital productivity as we move from one capital-labor ratio to the other. Similarly, the choice of the level of capital utilization is related to the trade-off between the lower average capital cost of a higher level of capital utilization and the higher wage which is required as a compensation for working non-normal hours.

The similarity of the two problems was recognized by Marris (1964). He says (p.5) “We may therefore restate the production function by saying that annual output now depends on three variables in place of two: instantaneous employment, quantity of capital and the rate of utilization.” In the standard production theory the firm has to choose only the optimal capital-labor ratio. Implicitly this is tantamount to assuming that the utilization of capital is always maintained at a constant level. Capital utilization adds another dimension to the choice problem of the firm. This second dimension can be thought of either in continuous terms, i.e. how much time during a day capital is utilized, where time is a continuous variable (this is the approach used by Georgescu-Roegen, 1969, 1970, 1972), or in discrete terms, i.e. how a firm chooses the number of shifts it operates (the latter approach is followed by both Marris (1964) and Betancourt and Clague (1981)).

The question then is what are the factors that determine the level of utilization. A standard approach within the literature is to distinguish the factors that are related to cyclical fluctuation and others which are not. Winston (1974b) says “idle capital is explained as a consequence of unwanted accidents and adversities that occur after [ex post] a plant is built, or as a result of rational ex ante investment planning.” In Marris’s words it is “methodologically convenient to distinguish” between them. This approach, at least at first sight, is in accord with the critics of the Kaleckian model in that there is a set of factors that determine the long run, ex ante, desired level of utilization and another set of factors that explain the short run, ex post, actual level.

The ex post unwanted utilization is usually explained in terms of cyclical fluctuations in demand (Winston (1974b) is actually using the word “Keynesian”), unexpected shocks or simply to errors in judgment by the entrepreneurs. The literature is mostly concerned with the factors that determine the ex ante level of utilization.

However, as we will see later, no matter how convenient it is to distinguish the two sets of

9 A necessary condition for that is the set of techniques of production and the production function to be “well-behaved”, in the sense that higher capital productivity is accompanied by lower labor productivity.
factors, there is not always a clear cut way to do so. As a result, the ex ante normal level of utilization is also related to the ex post actual level.

THE EX ANTE DECISION ON CAPITAL UTILIZATION

The literature on capacity utilization is well summarized by Winston (1974b). In his article he describes a series of determinants of the optimal capital utilization, ex ante. A first reason for desired excess capacity is related to the market structure. The excess capacity is used as an entry deterrent by the monopolist or oligopolist. A low level of utilization of capacity would make the entry into the market unprofitable. Different dimensions of this argument are provided by Kaldor (1935), Chamberlin (1962), Spence (1977) and Cowling (1981).

Another reason for intended excess capacity is “rhythmic” variations of demand. Winston (1974b) and Betancour and Clague (1981) give the example of a pizza restaurant. Since the demand for pizza is usually limited around lunch and dinner time it is profitable for the owner of a pizzeria to leave its “plant” idle when there is no demand (e.g. after midnight until late in the morning).

Except for “rhythmic” variations in demand, intended unutilized capital can be explained with rhythmic variations in the prices of inputs. If the price of an input is predictably higher during a certain period it might be profitable for a firm to leave its capital idle during this period and produce when the price of this input is low. The first of these inputs that comes to someone’s mind is labor. For reasons related to several social norms people prefer to work during the normal working hours and the wage for working outside of these normal hours is higher. Firms have to pay a utilization differential, a premium over the regular wage to workers who work outside of these normal hours, during an afternoon or a night shift or during the weekend. In most countries this differential is established by law.\textsuperscript{10}

The rhythmic variation of the input price is not limited to the price of labor. For example, in agriculture one of the most important inputs is heat. Heat’s price is zero during the spring and summer months. During the winter producing an equivalent amount of heat—for example with the construction of a greenhouse—is costly. Therefore, for most farmers it is more profitable to have their machinery remain idle during the winter.\textsuperscript{11}

We could think of many more similar examples of input price variations. However, the essence of the problem can be understood as follows: the firm owns the capital whether it uses it or not. As a result, it has to choose between a lower average cost of capital and higher cost of the rhyth-

\textsuperscript{10}As we mentioned above, Marx was the first to realize the importance of these social norms for the working day of capital. The higher labor cost during “abnormal” hours may induce the firms to leave their capital idle during these hours and utilize it only when labor is relatively cheap.

\textsuperscript{11}Sometimes it is impossible to reproduce the climate conditions of the summer. In this case the idleness of the agricultural machinery is inevitable. This is stressed in Georgescu-Roegen (1969, 1970, 1972).
mic input(s) when capital is utilized a lot, and a higher average cost of capital but higher cost of the rhythmic input(s) when the utilization of capital is low. A cost-minimizing firm will determine its optimum level of capital utilization having this trade-off in mind.

Intuitively, then, it is not hard to list some of the determinants of the utilization of capital. The higher the relative price of capital to the rhythmic price inputs and the capital intensity of the production process, the higher the incentive of the firm to utilize its capital. Similarly, the bigger the “amplitude” of the rhythm of the input price (the higher the wage differential for the second shift, or the more expensive to build a green-house) the lower the incentive for utilization.

In the next section we employ a simple model to show explicitly how technology, the cost of capital and the wage, the “rhythmic” variations of the wage, and the level of demand determine the normal utilization rate.

THE CHOICE OF THE NUMBER OF SHIFTS

As we explained in the previous section, the choice of the number of shifts is another important dimension in the decision-making process of the firm. In most microeconomic analysis, implicitly or explicitly, the utilization of capital is assumed to be constant (and usually 100 percent). Standard microeconomic theory teaches us that a firm faced with a certain demand for its output will choose the technique of production (the capital-labor ratio) that minimizes its cost. This description of the decision process of the firm neglects the fact that the firm manager has to choose not only how much capital (vis-à-vis labor) she will employ but also how much she will utilize this capital (a real manager who would “commit” a similar neglect would surely lead her firm to bankruptcy sooner rather than later).

This two-dimensional decision process can become clear with the help of Table 1. The firm can choose between \( m \) techniques of production, presented in the different rows of the table. Every technique is more mechanized than the previous one, so that the required capital for the production of a unit of product is larger as we move from technique 1 to \( m \). We will assume that the techniques of production are “well behaved,” that is techniques where increased mechanization is accompanied by higher labor productivity, so there exists the trade-off we described above.\(^\text{12}\)

The columns of the table refer to different systems of production. By system of production we mean the number of shifts: the firm chooses if it will produce its output under a single shift system (System 1) or a double shift system (System 2). We can think of a single shift system as a 5-day morning 8-hour shift and the double shift system as a 5-day morning and evening 8-hour shift. Therefore, under the single shift system the working week is 40 hours, while under

\(^{12}\)A similar matrix is introduced by Marris (1964, p. 42). He makes the same assumption about the techniques of production.
Table 1: Matrix of activity programs

<table>
<thead>
<tr>
<th>Technique</th>
<th>System 1</th>
<th>System 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1, 1</td>
<td>1, 2</td>
</tr>
<tr>
<td>2</td>
<td>2, 1</td>
<td>2, 2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>m</td>
<td>m, 1</td>
<td>m, 2</td>
</tr>
</tbody>
</table>

the double shift it is 80 hours. The double shift system means higher desired utilization compared to the single shift system. Obviously, there are other systems of operation too. We can have a third system of operation with an additional night shift, so the firm would have a 5-day 24-hour per day working week (in terms of Table 1 we would have to add a third column), or even a firm with a continuous operation of 168 hours per week. The choice between only two systems is made for reasons of simplicity and clarity of exposition, and does not limit our theoretical predictions, since our results can be easily extended to the cases of more than two systems of operation.\footnote{A discussion of a firm with more than two systems of operation can be found in Marris (1964) and Betancourt and Clague (1981, ch. 2).}

The different techniques and systems of production define a set of activity programs (in the case of Table 1 the \(m\) techniques and the 2 systems define a set of activity programs with \(m \times 2\) elements). The firm, faced with a certain demand for its product, will choose the activity program that minimizes its cost.

In the next sections we discuss how the firm chooses the level of the utilization of its capital for different sets of available techniques of production. We start with the simplest case, a firm with only one technique available, then we move to a firm that can choose between two techniques and finally to a firm with an infinite continuum of different techniques of production available (\(m \rightarrow \infty\)). We will handle this latter case with the use of a production function.\footnote{It is well known that a smooth production function describes a “technology with an infinite continuum of techniques.” A nice discussion is provided by Foley and Michl (1999, ch. 3).} The choice of a production function comes in contrast to Kurz (1986), who adopts a Sraffian approach. In the present paper we will conveniently assume away any kind of reswitching either between techniques or systems of production.
One technique of production

Imagine a firm that faces a certain demand for the commodity it produces, say $\bar{Q}$. There is only one technique of production available. The firm uses only labor and capital as inputs. For the production of $\bar{Q}$ it needs an amount of capital services ($S$) and labor ($L$). In terms of Table 1, there would be two possible activity programs, described by a $(1 \times 2)$ vector.

The utilization of capital can vary only through the time the capital is used, that is, by adopting a workweek of 40 or 80 hours. The speed of operation is constant. That means that the instantaneous rate of services of capital varies proportionally to the capital stock in place, $S = vK$, where $v$ is a constant. For reasons of simplicity we can normalize $v = 1$, so $S = K$.

Thus, the firm has to choose between two systems of production. In a single-shift system a certain amount of capital stock, $K^1$, and labor, $L^1$, are combined to produce $\bar{Q}$. In a double shift system half of the product is produced in the first shift and half in the second shift. In this case an amount of capital $K^2 = K^1/2$ is combined with an amount of labor $L^{21}$ in the first shift and $L^{22}$ in the second to produce $\bar{Q}$. As is probably clear by now, the first number of the superscript refers to the system of operation and the second (if there) to the particular shift within each system. It is also clear that the amount of labor in the two shifts is equal, so $L^{21} = L^{22} = L^2 = L^1/2$.

As stressed before, the main characteristic of increased utilization is the higher cost of labor for the second shift. Firms have to pay a utilization differential which we can define as the ratio of the wage for working in the morning shift, $w_1$, and the wage for working in the evening shift, $w_2$. The utilization differential is equal to $\frac{w_2}{w_1} = 1 + \alpha$, where $\alpha > 0$. We can also define $r$ as the unit cost of capital. Following Marris (1964) and Betancourt and Clague (1981)—and in contrast to Kurz (1986)—we assume that the depreciation of capital is invariant to the system of operation. However, except for this change, which does not affect our main conclusions, our firm up to this point is similar to that in Kurz’s (1986) paper.

We can make one addition. We can assume that the production under the single shift system gives rise to some economies of scale vis-à-vis the double shift system. We can also assume that these economies of scale depend on the level of the production of the firm, $\bar{Q}$. We will discuss the issue of the economies of scale in more detail in the following sections.

The choice of the system of production “will be exclusively grounded on cheapness.” The total cost of production under the first system will be $C^1 = (rK^1 + w_1L^1)/\xi(\bar{Q})$, where $\xi(\bar{Q})$ refers to the economies of large scale production. Under the double shift system the cost of production is $C^2 = rK^2 + w_1L^{21} + w_2L^{22} = rK^2 + (2 + \alpha)w_1L^2$. We can define the cost ratio of the double shift system over the single shift system as $\Lambda = \frac{C^2}{C^1}$. It is easy to see that

$$\Lambda = \left[\pi + (2 + \alpha)\psi\right] \frac{\xi(\bar{Q})}{2}$$

(1)

where $\pi$ is the share of capital cost and $\psi = 1 - \pi$ is the share of wage cost to the total cost of
production under the single shift system, respectively. The firm will choose the “cheapest” system of production. The double shift system will be chosen as long as $\Lambda < 1$.

Using equation (1) we can make some conclusions about the factors that determine the utilization of capital. First, it is easy to see that $\partial \Lambda / \partial \pi = -(1 + \alpha)\zeta(\bar{Q})/2 < 0$. That is, the more capital intensive the technique of production is, the stronger the incentive of the firm to utilize it as much as possible. Second, $\partial \Lambda / \partial \alpha = \psi \zeta(\bar{Q})/2 > 0$. The larger the utilization differential, the higher the relative cost of the double shift system, and thus the firm will tend to use a single shift system of production.

The analysis so far validates the conclusion of Kurz: “The cost minimizing degree of capital utilization...[depends] on income distribution, i.e. a the basic wage rate (or alternatively, the rate of profits) and the wage differential to be paid for differently intensive work and work outside of ordinary working hours.” In other words, the rate of utilization depends only on technological (through $\pi$ and $\psi$) and distributional factors (through $\pi$ and $\psi$ and $\alpha$).

It is also worth examining the role of the economies of scale. We can see that $\partial \Lambda / \partial \zeta = [\pi + (2 + \alpha)\psi]/2 > 0$. The higher the returns to large scale of production, the higher the relative cost of the double shift system and the firm will have an incentive to underutilize its capital.

An important addition to our conclusions can be made when the economies of scale depend on the demand for the product of the firm, $\bar{Q}$. We can easily see that $\partial \Lambda / \partial \bar{Q} = [\pi + (2 + \alpha)\psi]/2$. That is, $\partial \Lambda / \partial \bar{Q} < 0$ if $\zeta'(\bar{Q}) < 0$. The entrepreneur will tend to choose a double shift system of operation over a single shift system of operation as the scale of production of her firm increases if the degree of the returns of scale decreases as the scale of production increases. This result is important because it shows that the level of utilization for a cost-minimizing firm depends on the demand for the product of the firm. Under these assumptions about the returns to scale, an increase in the demand for the product of the firm will tend, ceteris paribus, to increase the utilization of its capital. The next question is if there are reasons to believe that the returns to scale evolve in such a way. We discuss this later.

**Two techniques of production**

We can now imagine that there are two available techniques of production, say $I$ and $II$. Table 1 would now be a $(2 \times 2)$ matrix with 4 available activity programs. The addition of one more technique can help us understand how the choice of the system of production is related with the choice of the technique of production.

As before, under each of the techniques, the firm has to choose between two systems of production. Under technique $I$ in a single-shift system a certain amount of capital stock, $K^1_I$, and labor, $L^1_I$, are combined to produce $\bar{Q}$. In a double shift system an amount of capital $K^2_I = K^1_I/2$ is combined with an amount of labor $L^{21}_I$ in the first shift and $L^{22}_I$ in the second to produce $\bar{Q}$.

The related amounts under the second technique would be $K^1_{II}, L^1_{II}, K^2_{II} = K^1_{II}/2, L^{21}_{II} = L^{22}_{II}$. 

10
Technique II is more capital intensive than I, therefore $K_{II} > K_I$ and $L_{II} < L_I$.

A question we can ask is what technique of production the firm will choose for each system of production. For a single shift system the firm will choose technique II if $C^1_I > C^1_{II}$. With some simple algebra this can be restated as

$$
\frac{r}{w_I} \frac{K^1_{II} - K^1_I}{L^1_I - L^1_{II}} < 1
$$

Respectively, for the double-shift system the firm will choose technique II if $C^2_I > C^2_{II}$, which can be restated as

$$
\frac{r}{w_I} \frac{K^2_{II} - K^2_I}{L^2_I - L^2_{II}} < 2 + \alpha
$$

Equations (2) and (3) suggest that the utilization differential makes the choice of the more capital intensive technique more possible. Under the double shift system (when the utilization differential kicks in) the relative price of labor to capital is higher, so the firm will tend to adopt a technique of production which is more capital intensive.

**Infinite Techniques of production**

The results of the previous sections can be generalized for a technology with an infinite continuum of techniques. We can capture this with the use of a production function $Q = G(K, L)$.

For simplicity, as in Betancourt and Clague (1981), we shall assume that labor can be hired within eight hours shifts and we “shall define our instant to be the eight hour shift, that is, the eight hour shift is the unit of time.”

We will also assume that the ratio of capital to labor services in both shifts is constant after the plant is built\(^{15}\), so $L^{21} = L^{22}$. Therefore, under the double shift system half of the output will be produced during the first shift and half of it during the second shift. The firm will thus have to produce the demanded output choosing between the following systems of production

$$
\bar{Q} = Q^1 = G(K^1, L^1)
$$

$$
\bar{Q} = Q^2 = G(K^2, L^{21}) + G(K^2, L^{22}) = 2G(v K^2, L^{21})
$$

As before, the firm will choose the system of production with the lowest cost. The ratio of the cost of the double shift system over the single shift system can be now written as

$$
\Lambda = \left[ \pi \frac{K^2}{K^1} + (2 + \alpha) \psi \right] \frac{L^{21}}{L^1}
$$

\(^{15}\)Winston (1974a) and Betancourt and Clague (1981) show that this is equivalent to zero ex post (after the plant is built) elasticity of substitution.
where \( k^2 = \frac{K^2}{L^2} \) and \( k^1 = \frac{K^1}{L^1} \) are the ratios of capital to labor of the first shift under the single and the double shift system, respectively. The firm will choose the “cheapest” system of production. The double shift system will be chosen as long as \( \Lambda < 1 \).

For reasons of simplicity of exposition, and because we want to focus on the role of the economies of scale in the choice of the system of production, we will also assume that our production function is homothetic, therefore \( Q = G[F(K,L)] \), where \( F \) is a homogeneous function of first degree and \( G \) is a positive transformation of \( F \). Based on what we have said so far under the first system we have

\[
\bar{Q} = G[F(K^1,L^1)] = G(F^1) = G[L^1 f(k^1)]
\]

(6)

while under the second system

\[
\bar{Q} = 2G[F(K^2,L^{21})] = 2G(F^2) = 2G[L^{21} f(k^2)]
\]

(7)

From (6) and (7) we derive

\[
L^1 = \frac{G^{-1}(\bar{Q})}{f(k^1)}
\]

\[
L^{21} = \frac{G^{-1}(1/2\bar{Q})}{f(k^2)}
\]

(8)

Therefore

\[
\frac{L^{21}}{L^1} = \frac{G^{-1}(1/2\bar{Q})}{G^{-1}(\bar{Q})} \frac{f(k^1)}{f(k^2)}
\]

(9)

Based on (9), equation (5) is transformed into

\[
\Lambda = \left[ \pi \frac{k^2}{k^1} + (2 + \alpha) \psi \right] \frac{f(k^1)}{f(k^2)} \Phi(\bar{Q})
\]

(10)

where \( \Phi(\bar{Q}) = \frac{G^{-1}(1/2\bar{Q})}{G^{-1}(\bar{Q})} \). This function is related to the economies of scale. \( G^{-1}(1/2\bar{Q}) \) is the amount of inputs needed to produce \( 1/2\bar{Q} \) and \( G^{-1}(\bar{Q}) \) is the amount of inputs needed to produce the whole \( \bar{Q} \). If the production process exhibits constant returns to scale, the amount of inputs needed to produce half of a certain amount of the output is half of the inputs needed to produce the whole amount, therefore \( G^{-1}(1/2\bar{Q}) = 1/2G^{-1}(\bar{Q}) \), so \( \Phi(\bar{Q}) = 1/2 \). On the other hand if we have increasing returns to scale the amount of inputs needed to produce half of a certain amount of output are more than half of the inputs needed to produce the whole amount, therefore \( G^{-1}(1/2\bar{Q}) > 1/2G^{-1}(\bar{Q}) \), so \( \Phi(\bar{Q}) > 1/2 \). Mutatis mutandis for decreasing returns. It is also clear that \( \Phi(\bar{Q}) \) is analogous with the function \( \zeta(\bar{Q}) \) we defined in the previous sections.

Based on equation (10) we can derive the conclusions of the previous sections more for-
mally: i) the higher the share of the wage cost in the total cost under the single shift system \((\psi)\), the lower will be the cost of the double shift system relative to the single shift system; ii) the higher the utilization differential, the higher the cost of the double shift system relative to the single shift system; iii) the capital to labor ratio in each shift under the double shift system \((k^2)\) will be greater than the capital to labor ratio under the single shift system \((k^1)\) and finally; iv) the larger the returns to scale \((\phi(\bar{Q}))\), the higher the ratio \(\Lambda\), and, therefore, the higher the cost of the double shift system relative to the single shift system.\(^{16}\)

**SCALE OF PRODUCTION AND UTILIZATION**

A question of interest is what is the relation between the choice for the firm between the two systems and the scale of production, in other words what happens to the ratio \(\Lambda\) as the output of the firm \((\tilde{Q})\) increases. From equation (10) we can see that \(\Lambda = \left[\theta \frac{k^2}{k^1} + (2 + \alpha)(1 - \theta)\right] \frac{\tilde{F}(k^1)}{\tilde{F}(k^2)} \phi(\tilde{Q}) = c\phi(\tilde{Q})\) where \(c = \left[\theta \frac{k^2}{k^1} + (2 + \alpha)(1 - \theta)\right] \frac{\tilde{F}(k^1)}{\tilde{F}(k^2)}\). Because of the homotheticity of the production function \(c\) is invariant to changes in \(\tilde{Q}\), so \(\frac{\partial \Lambda}{\partial \tilde{Q}} = c \frac{\partial \phi}{\partial \tilde{Q}}\). The sufficient condition for the double shift system to become more attractive as production increases \((\frac{\partial \Lambda}{\partial \tilde{Q}} < 0)\) is \(\frac{\partial \phi}{\partial \tilde{Q}} < 0\). By the definition of \(\phi(\tilde{Q})\) we can see that this condition is met if\(^{17}\)

\[
\frac{G^{-1}(\tilde{Q})}{G^{-1}(1/2\tilde{Q})} \frac{G'[G^{-1}(\tilde{Q})]}{G'[G^{-1}(1/2\tilde{Q})]} < 2 \quad (11)
\]

The first fraction on the left hand side of the inequality is the inverse of \(\phi(\tilde{Q})\), while the second one is equal to the marginal product at the amount of inputs producing \(\tilde{Q}\) over the marginal product at the amount of inputs producing \(1/2\tilde{Q}\).

The sufficient condition for the satisfaction of this inequality is that the rate of the degrees of scale of the production function decreases. Our production function can be characterized for its returns to scale based on the level of \(j\) in the equation \(i^jQ = G(iF)\). For \(j = 1\) we have constant returns to scale, for \(j > 1\) we have increasing returns to scale, and vice versa. If \(F\) is the amount of inputs necessary for the production of \(\tilde{Q}\) and \(F'\) for the production of \(1/2\tilde{Q}\), then the LHS of inequality (11) can be written as \(\frac{F'}{G'(F)}\). Assuming that at the level of input \(F'\) scales of production \(j'\) prevail, while at \(F\) scales of production \(j\), we can rewrite this fraction using the Euler theorem as \(\frac{F'}{G'(F')} = \frac{j'0}{j0} = 2\frac{j'}{j}\). Obviously, if we have a constant degree of returns to scale (ei-

---

16 The results can be derived with some simple differentiation of equation (10), together with the results of the cost-minimization process of the firm. These results have already been demonstrated for one and two techniques of production and we skip them here for reasons of economy of space.

17 In more detail, 11 is derived as follows:

\[
\frac{\partial \phi}{\partial \tilde{Q}} < 0 \iff \frac{G^{-1}(1/2\tilde{Q})G^{-1}(\tilde{Q}) - G^{-1}(1/2\tilde{Q})G^{-1}(\tilde{Q})}{[G^{-1}(1/2\tilde{Q})]^2} < 0 \iff \\
\frac{1}{2} \frac{G^{-1}(\tilde{Q}) - G^{-1}(1/2\tilde{Q})}{G[G^{-1}(\tilde{Q})]} < 0 \iff \frac{1}{2} \frac{G^{-1}(\tilde{Q})}{G[G^{-1}(\tilde{Q})]} < 1
\]
ther if they are constant or increasing or decreasing) \( j = j' \) so \( \frac{\partial \omega}{\partial Q} = 2 \), therefore the scale of the output does not affect \( \Lambda \) and the choice of the system of production. On the other hand if \( j < j' \), that is if the rate of the degrees of scale decreases, \( \frac{\partial \omega}{\partial Q} < 2 \).

Therefore, the entrepreneur will tend to choose a double shift system of operation over a single shift system of operation as the scale of production of her firm increases if the degree of the returns of scale decreases as the scale of production increases. This result is important because it shows that the level of utilization for a cost-minimizing firm depends on the level of demand for its product. Under these assumptions about the returns to scale, an increase in the demand for the product of the firm will tend, ceteris paribus, to increase the utilization of its capital.

The next question is if this decrease of the returns to scale is something more than a theoretical sophistry. We discuss this in the following section.

**ECONOMIES OF SCALE**

In order to examine the behavior of the returns to scale as production increases we have to understand the causes of increasing returns. Within the literature there is a consensus that returns to scale are—at least to a certain degree—due to *indivisibilities*. Kaldor (1934) in a much cited passage mentions:

“It appears methodologically convenient to treat all cases of large-scale economies under the heading “indivisibility.” This introduces a certain unity into analysis and makes possible at the same time a clarification of the relationship between the different kinds of economies. Even those cases of increasing returns where a more-than-proportionate increase in output occurs merely on account of an increase in the amounts of the factors used, without any change in the proportions of the factors, are due to indivisibilities; only in this case it is not so much the “original factors,” but the specialized functions of those factors which are indivisible.”

Koopmans (1957, p.152) agrees with Kaldor and writes “I have not found one example of increasing returns to scale in which there is not some indivisible commodity in the surrounding circumstances.” Eatwell (2008) also explains why indivisibilities are the sole reason behind increasing returns to scale.

In the presence of indivisibilities, the returns to scale are created because production can be increased by, for example, \( t \) times without increasing all the inputs analogously. A simple example given by Georgescu-Roegen (1972, p.285) is the production of bread; compared to the baking of one loaf of bread “the baking of a batch of loaves does not require the amplification of all coordinates. The same mixing machine, oven, and building can take care of the multiple tasks.” In our terminology the “mixing machine, oven, and building” are indivisible inputs to
production. The baker needs at least one unit of each of these inputs whether she produces one or “a batch of loaves”. Obviously, the baker will use a smaller mixing machine, oven, and building if the demand for her production is small, compared to a baker that faces very high demand. However, usually capital is not divisible enough to justify the neglect of the benefits provided by a larger scale of production.

These benefits—and with them the rate of the returns to scale—will be exhausted as the scale of production increases. More formally, it can be shown that the rate of returns to scale is equal to the ratio of the average cost to the marginal cost; \( j(Q) = \frac{AC(Q)}{MC(Q)} \), where \( j \) as before is the rate of the returns to scale. The existence of indivisibilities means that \( AC > MC \) and (assuming constant prices of inputs and no substitution) that the average cost is decreasing while the marginal cost stays constant as production increases, thus the rate of the returns to scale decreases as production increases: \( j'(Q) < 0 \). As we showed in the previous section, if the returns to scale behave this way, the firm will tend to increase the utilization of its capital as the demand for its product increases.

Within the literature on capital utilization it has been Georgescu-Roegen (1969, 1970, 1972) who has paid much attention to the role of indivisibilities as a cause of increasing utilization when demand increases. He argues that during the production of any good there are inevitably some idle resources. The degree of this idleness can only be reduced if the demand for the output of the firm, and, as a result, the scale of production, increases. From a historical point of view Georgescu-Roegen goes as far as to argue that the Industrial Revolution was a result of increase in demand, which allowed the artisan shops of the time to start working with the “factory system” (a system with high utilization of capital), which in turn allowed the further division of labor with the results described by Adam Smith in the first chapters of *The Wealth of Nations*. This is his interpretation of the Smithian proposal that “the division of labor is limited by the size of the market.”

The role of increasing demand as a cause of increasing utilization is also highlighted by Marrs (1964). He argues that because of indivisibilities the entrepreneur either has to choose a different technique of production compared to the one that would be optimum if there were no indivisibilities, or underutilize his equipment (or both). He concludes that “the general tendency of effect of restraints on output must inevitably be in the direction of reducing optimum practicable rates of utilization.” As the demand for the product of a firm increases the “optimum practicable rates of utilization” will also increase. In other words, the desired rate of utilization will increase.

The causes of the returns to scale are not limited to indivisibilities. Kaldor himself (1966, 1972), three decades after his 1934 article, questions his thesis that all cases of large-scale economies can be treated under the heading of indivisibility. Among others, he refers to the returns to scale

---

18 This is the famous title of chapter III of Book I of *The Wealth of Nations* (Smith, 1999, p. 121).
caused by the “three-dimensional nature of space.” The usual example for this kind of economies of scale—it is given by Kaldor (1972) himself, Koopmans (1957), and Eatwell (2008)—is a cylinder, whose construction cost varies with its diameter \((2r\pi)\) while its capacity varies with the area of its top \((\pi r^2)\). In this case doubling the inputs (the radius) quadruples the output/capacity and we have a constant rate of returns to scale equal to 2. However we could have a decreasing rate of returns also in this case, if for example a larger diameter would require a thicker plate.

Young (1928), following Adam Smith, emphasized another cause of returns to scale: “with the division of labour a group of complex processes is transformed into a succession of simpler processes, some of which, at least, lend themselves to the use of machinery.” The returns to scale are a result of increased differentiation and the invention of new processes. This phenomenon can take place either at a micro level (division of the product of the same firm into simpler processes) or a macro level (division of production among different firms or sectors). Young (1928) and Kaldor (1966) seemed to favor the latter. They argue that the benefits of this kind of differentiation will be the “emergence of new subsidiary industries.” Their argument is verified by economic history. At the macro level it is not easy to conclude if the resulting returns to scale are exhausted or not, but we can make some safer conclusions for the micro level. The benefits of the division of labor within a firm are exhausted because the number of the simpler production processes is finite. In the famous Smithian pin factory it is very clear how the division of the tasks in the production of pins generates significant benefits, but it is also very clear that the number of the different tasks is finite and gets exhausted as the the extent of the market allows the firm to perform this division of tasks.  

In total, if we take into account all the above causes of returns to scale, together with reasons that would burden the firm with additional cost as production increases (e.g. increased bureaucracy, more difficult control of production, etc.) we can convincingly argue in favor of a rate of returns to scale which, at the firm level, is decreasing as the demand for the output of the firm and the scale of production is increasing. Therefore, the rate of utilization for the individual firm is not determined only by exogenous structural characteristics, like technology, the relative price of capital to the rhythmic price of inputs, or the degree of monopoly of the market, as it is usually assumed in the literature. On the contrary the cost-minimizing firm tends to adjust the level of its capacity utilization in the face of changes in the demand for its product.

Empirical work so far has not provided definite answers on the behavior of the returns to scale for different levels of production. Betancourt and Clague (1981) use data from the United

---

19Smith (1999, p.110) distinguishes eighteen different tasks: drawing out the wire, straightening it, cutting it, pointing it, grinding it at the top for receiving the head, etc. Smith (p.112) also highlights the “increased dexterity in every particular workman” as a possible source of returns to scale. More recently, it was Arrow (1962) who explained the benefits of learning by doing. Kaldor (1966, 1972) also includes it in his taxonomy of the causes of the returns to scale. It is hard to make a definite statement about the behavior of the rate of returns to scale due to learning by doing. However, it appears plausible that at the micro level, within an individual firm, if the rate of the returns is not decreasing, it is at least not increasing.
Nations Industrial Development Organization (UNIDO) for France, Japan, India and Israel and conclude that for the first two countries the rate of returns to scale is constant, while for India and Israel it is, decreasing. Other empirical efforts are also ambiguous. Haldi and Whitcomb (1967) and Moore (1959) examine engineering data and find a constant rate of returns to scale. On the other hand Bain (1956) and Pratten (1971) provide interview estimates which show that the benefits due to scale tend to be exhausted as the scale of production increases. Basu (2008) shows that more recent estimations also do not provide a clear answer. Marris (1964) is more straightforward and clear. He examines his theory of utilization for the UK firms. He confirms from his empirical research that scale restraints in the production of the firm lead—“through indivisibilities”—to underutilization of capital. In the context of our discussion, this supports a decreasing rate of returns to scale.

CONCLUDING REMARKS

In the previous sections we analyzed how a firm that minimizes its cost chooses the level of the utilization of its capacity. We highlighted how different technological factors and the cost of the production inputs affect this choice. Most importantly, we demonstrated that the choice of the system of production is related to demand. Higher demand can lead to an increase in the desired utilization and vice versa. As we explained, this conclusion is conditional on the behavior of the economies of scale for the individual firm. In the last section we discussed why the theory of the returns to scale provides justification for such a behavior.

Interestingly, such a behavior is also justified by the answers of the entrepreneurs themselves when they are asked why they do not utilize their capital enough. In the second page of his book Marris (1964) writes

“in business inquiries, one of the commonest reasons given for working shifts or not (as the case may be) relates to demand [emphasis added].”

Betancourt and Clague (1981) (also in page 2!) write

“interviews have shown that when factory managers have been asked why they are operating only one shift one of the most frequent answer given is that the firm would not be able to sell its product [emphasis added].”

These answers verify that the behavior of the actual firms is similar to our fictional firm, whose rate of returns to scale is decreasing as the demand for their product is increasing. Also interestingly, the literature tends to overlook the intuition provided by these answers and focuses on the “structural” determinants of the desired rate of utilization.

Some final remarks are in order here. First, this result does not settle the debate between the supporters of the Kaleckian model and its critics. Many more steps are necessary towards that
direction. The links between the micro and the macro level, as well as the empirical discussion of the Kaleckian analysis are some of these steps that come easily to mind. However, our discussion is important because it reconciles an endogenous normal rate of capacity utilization (a normal rate that reacts to the changes in demand) with a firm that minimizes its cost (or maximizes its profits).

Second, it would be interesting to see if our results hold under different extensions of the model. One could add more adjustment mechanisms for the firm to deal with increasing demand, like overtime labor. Another interesting extension would be a dynamic treatment of the decision process of the firm, in place of our static one, where for example the firm has certain expectations about the future demand of its product. We believe that our results carry through these extensions if someone does not assume perfect knowledge for the firm manager, but the formal demonstration of this result would be useful.

Finally, as in every theoretical issue, there should be a thorough check against reality, an empirical investigation of the issues, which are only addressed here at the theoretical level.
REFERENCES


