SHORT-RUN RETURNS TO SCALE AND THE PRODUCTION OF SERVICES

W. DAVID MAXWELL

Tulane University

In the following article it is argued that by relatively minor alterations of customary assumptions the nature of the short-run theory of the firm is changed quite significantly. By this alteration a form of output-maximizing or cost-minimizing entrepreneurial behavior is revealed that is not so apparent in traditional short-run theory. Instances of this behavior, and of behavior related to it, are particularly evident in the production of services. On this and other grounds it is argued that the alteration of assumptions is analytically desirable.  

CONVENTIONAL PRESENTATION OF THE LAW OF VARIABLE PROPORTIONS

Figure 1, below, represents a conventional presentation of the law of variable proportions, TP representing the various possible total outputs that would be yielded by combining various quantities of the services of the variable factor “A” with a fixed quantity of the services of some fixed factor of production. Over the range of output OA’ there are increasing average returns (the average product for the services of factor “A” is rising) and negative marginal productivity for the services of the fixed factor.

Assuming the supply of the services of the variable factor to the firm to be perfectly elastic and taking into consideration the cost of the fixed factor, the conventional textbook cost curves depicted in Figure 2 can be derived, OB being the output of the product that would require OA’ (Figure 1) of the services of the variable factor in combination with the services of the fixed factor. Thus given the shape of the production function shown in Figure 1 (and the other assumptions specified) the configuration of the cost curves in Figure 2 are predetermined (e.g., the point of inflexion of TP within the range OA’ would correspond to the minimum point on MC, etc.).

We have become accustomed, by analogies to automobile drivers and otherwise, to viewing the theory of the firm as descriptive of entrepreneurial behavior, and we assume such behavior to be rational. Thus Figure 2 indicates, for any given output, the least of the various types of cost that could be incurred in producing that output—in the view of the entrepreneur. These cost curves do not, of course, indicate the output at which the firm should produce but they do indicate the lowest costs at which any of the possible outputs could be produced—in the short run.

Thus the production function (Figure 1) from which the cost curves are derived must show the maximum outputs that can be secured if various quantities of the services of factor A are combined with the services of the fixed factor. As previously noted, within the range OA’ in Figure 1 the marginal productivity of the services of the fixed factor is negative. Thus if TP indicates the maximum outputs that can be secured, it must be true that the services of the fixed factor are indivisible. Otherwise larger outputs than those indicated by TP (over the

---

1 I am particularly indebted to Professors Ed Edwards of Rice University and Clark Edwards of Oklahoma State University for their helpful comments on an earlier draft. Any errors are, of course, my own responsibility.
range OA') could be secured by simply not permitting some of the services of the fixed factor to enter into the productive process. Such an action is precluded by assumption as well, however, for TP is drawn under the assumption that all of the services of the fixed factor are being combined with various quantities of the services of the variable factor (including those quantities within range OA').

At this point a dissenting view might arise. It might be argued that the traditional short-run production function does not necessarily assume indivisibility of the services of the fixed factor. It might be maintained that Figure 1 depicts the results of a hypothetical experiment in which various quantities of the services of the variable factors are combined with all the services of the fixed factor regardless of the divisibility of the services of the latter. There is no dispute that such an experiment could yield a curve such as TP even if the services of the fixed factor are perfectly divisible. Put otherwise, it is not contended that a section through the production surface could not have some such shape even if the services of the fixed factor are divisible. This would mean, however, that TP is not a curve of the maximum outputs that could be secured by combining various quantities of the services of factor "A" with those of the fixed factor. This would mean, in turn, that the cost curves of the firm would not depict the lowest possible costs of achieving various outputs in the short run.

Thus we must view TP as a curve of maximum outputs in the traditional short-run theory of the firm. We must also, however, assume the services of the fixed factor to be indivisible. Although the law of variable proportions is subject to a number of interpretations,2 in the "standard" case the services of the fixed factor are assumed to be indivisible.3

Of course, economists do not assert that the short-run cost curves of the firm must have the particular configuration possessed by those shown in Figure 2. The first important point, however, is that the choice of a particular shape or configuration for these curves is logically limited by the choice


\[^3\] Stigler, *op. cit.*, pp. 308-310. In the "standard case" the services of the fixed factor are considered "adaptable" as well as indivisible. "Adaptability" is a muddy concept, however, exemplified by the ability of the firm to change eleven shovels into twelve smaller ones in the short run but the inability to secure a twelfth shovel of the original size. The morass of "adaptability" need only be entered, however, if the previous assumption has been made that all of the services of the fixed factor are combined with any input of the services of the variable, no matter how small the latter (i.e., the assumption of indivisibility of the services of the fixed factor).
made concerning the shape and configuration of the short-run production function. The second important point is that the choice of the latter function is also a choice of assumptions concerning the indivisibility of the services of the fixed factor. In the immediately following section this choice of assumptions is examined.

THE DIVISIBILITY OF THE SERVICES OF THE FIXED FACTOR

In examining the question of the services of the fixed factor we must bear in mind that it is the divisibility of the services that is in question. Only factor services enter the production function. We must therefore avoid confusing, in Machlup’s terminology, “indivisible in acquisition” and “indivisible in use.” It may be possible to acquire a factor of production only in “lumps” or certain discrete sizes, yet the services yielded by this factor may be highly divisible.

Clearly, the indivisibility of the services of the fixed factor that is assumed in Figure 1 is indivisibility in use. TP represents the maximum outputs that can be secured by combining various quantities of the services of factor “A” with all of the services of the fixed factor. But is such an assumption of indivisibility the most plausible assumption to make about the services of a fixed factor in the short run? Consider again the implications of the range OA’ in Figure 1. By analogy, the indivisibility postulated is such that maximum output is yielded for the services of one “dose” of labor (the variable) by combining with this one “dose” all of the services yielded by the land of the King Ranch; the warehouseman’s maximum output, at an input of one bushel, is achieved by scattering that bushel over the warehouse floor; and the services of one laborer are combined with all of the services of an aircraft plant. There is no option at any

input of the services of the variable to use less than all of the services of the fixed. Such an action is precluded by the assumption of indivisibility in the services of the fixed factor.

When the time element in the nature of factor services is recognized, the assumption of indivisibility of the services of the fixed factor becomes doubly questionable. Although buildings or machines may be acquired in only discrete sizes their services are yielded through time and for this reason are highly divisible. It is seldom the case that an entrepreneur must use all of the services of the fixed factor at all inputs of the services of the variable. Again, by analogy, the manager of the King Ranch would not be precluded from using less than all of the services yielded by the ranch if he were combining these services with small quantities of labor, the warehouseman would not be forced to use all of his warehouse space to store a few bushels of potatoes, and the manager of the aircraft plant would not have to use all of the services of the plant at small labor inputs.

Reflection upon the physical nature of the productive process suggests that instances in which all of the services yielded by a factor of production must enter into the productive process regardless of the level of output are comparatively rare. An electrical generator may revolve at a constant speed throughout the twenty-four hours of the day, but the amount of power actually delivered into the productive process depends upon the load connected to the generator and may easily be varied from zero up to the physical capacity of the generator. A “thousand-watt” generator delivers only sixty watts to a sixty-watt bulb. Similarly, a train or motor truck that is transporting one small package does not deliver into the productive process more than a fraction of the transportation services that it could provide. And a constant-speed press which can stamp a thou-

---

sand items per hour from metal sheets can also stamp any quantity less than this.

Thus the more realistic view of the services of a fixed factor is that these services are sufficiently divisible to permit the entrepreneur to use less than all of them if by doing so he can increase output. In this view the limitation upon the entrepreneur in the short run is primarily that he can use no more than the total stock of services embodied in the fixed factor. He can use less.

If we accept this view, one of its implications should be immediately pointed out. Economists are fond, and rightly so, of the coincidence of dichotomies. In the view which we have rejected, the dichotomy of returns to proportion versus returns to scale corresponds to the short-run, long-run dichotomy. This coincidence can be maintained, however, only by the assumption of complete indivisibility of the services of the fixed factor. We have rejected this assumption and thereby the coincidence of these dichotomies.

But, the reader may well feel that he is being asked to choose between black and white and to deny the existence of the many shades of gray. An assumption of complete indivisibility of the services of a fixed factor underlies the traditional short-run analysis, whereas the assumption of complete divisibility of these services is apparently being suggested. Would not an assumption somewhere between these two “extremes” be more plausible?

Repetition is perhaps permissible on so fundamental a point. The concept of factor services is a highly abstract concept. The classical concept of capital as the “congealed” services of land and labor reflects this high degree of abstraction. The building or machinery used by the entrepreneur may be of given physical dimensions and may not in many cases be readily physically divisible. Even so, such a lack of physical divisibility would not imply a similar lack of divisibility in the services yielded by the building or machinery, for their contribution to the productive process consists of the stream of factor services they yield through time—not physical bits and pieces of themselves. To assume that a given stock of factor services is highly divisible (up to the limit of the amount previously acquired) is thus not an implausible assumption.¹

In Figure 2 we can retain AFC and its influence on ATC because we have not denied indivisibility in the acquisition of the services of the fixed factor. Our short run continues to be defined in terms of indivisibility in the acquisition of a stock of factor services. Thus our short run might be defined as the period of time during which the entrepreneur cannot change the size of his plant. But since he need not use all of the services yielded by the plant at all inputs of the services of the variable, variation in both types of factor services is permissible up to the constraint imposed by the total stock of services embodied in the fixed factor. Thus we have both returns to scale and returns to proportion in the short run. The primary purposes of the immediately succeeding section are to make this point more evident and to trace the effect of an assumption of divisibility (in use) of the services of the fixed factor upon the firm’s cost curves.

THE OUTPUT-EXPANSION PATH OF THE FIRM

We have agreed that short-run cost curves indicate the lowest cost at which

¹The failure to distinguish between units of a factor of production and units of the services yielded by a factor of production is understandable, but not excusable. The former is a less abstract and conceptually more readily measurable unit. But no matter what assumptions are made concerning the characteristics of the units in which a factor may be acquired, the characteristics of the service yielded by the factor are analytically distinct.

Those readers who object to the concept of a divisible fixed factor as a stock of factor services all of which need not enter into production may prefer the equivalent concept (for the context of this paper) of a stock of one factor which has an upper limit to the rates at which it can yield services.
various outputs can be produced. We have agreed that these curves are partially derived from the production function and must be compatible therewith. We have agreed that a curve depicting the total output that can be secured by combining factor services must indicate the maximum total outputs that can be secured by combining the particular quantities of factor services indicated. But we have rejected the assumption that the services yielded by the fixed factor are indivisible.

Since we will therefore permit variation in the quantity of the fixed factor's services that is used (i.e., that enter the production function) up to the constraint imposed by the total stock of the services that has been acquired, we must have recourse for further analysis to some type of representation of a three-dimensional surface—since three variables are now involved (services of the variable factor, services of the "fixed" factor, and output).

It should be reemphasized that we are not attempting to portray the equilibrium position of the firm. We are interested in deriving the short-run cost curves of the firm under our altered assumption concerning the divisibility of the services of the fixed factor. This means, in turn, that we are interested in determining the output-expansion path of the firm on the production surface. We are not at this time discussing the wisdom of operating at any particular point (or range of points) on the output-expansion path nor that of producing at any point (or range of points) on the firm's cost curves. In other words, we are seeking the equivalent of TP of Figure 1 under assumptions different from those that underlie TP and are also seeking to show the necessary implications of this equivalent for the short-run cost curves of the firm.

If we assume that the services of all factors are divisible, however, this precipitates us into the ancient argument concerning what type of production function is implied thereby. Without going into this argument, we shall side with Kaldor, Lerner, Knight, et al., against the express opposition of Chamberlin and at least doubtful attitude of Boulding by assuming that perfect divisibility of factor services would imply constant returns to scale.\(^6\)

Thus in the following analysis we shall assume a production function characterized by constant returns to scale. We shall also assume that there are only two inputs—labor services and capital services—and that while the former is variable in use and acquisition, the latter is variable in use only, a given stock of capital services having been previously acquired. Initially we shall also assume that the user cost of the services of capital is zero; i.e., that if the entrepreneur decides to use more or less of his given stock of capital services he incurs no additional cost on this account. We shall also assume that, having acquired the given stock of capital services for this particular productive process, the entrepreneur does not have the option, in the short run, of selling these services or, initially, of using them to help produce other goods (i.e., that the opportunity cost of capital services in the short run is zero).

Under these assumptions it is clear that the entrepreneur has an output-maximization or cost-minimization decision to make at every point along that portion of the output-expansion path in which he is not employing all of the given stock of capital services which he has previously acquired. In short, if he uses less than all of his capital services at some labor-services inputs, there must be some rationale by which he decides what portion of his stock of capital services to leave unemployed at these labor-

\(^6\) The argument, pro and con, is succinctly summarized by Professor Chamberlin in *The Theory of Monopolistic Competition*, sixth ed. (Cambridge, Mass.: Harvard University Press, 1950), pp. 235-244. The important point—that divisibility of the services of all factors implies, even for the short run, a problem of returns to scale—is unaffected by the position taken on this controversy.
services inputs. Under our assumptions the nature of his decision is clear. At each input of labor services he will combine with these services the services of capital until the marginal productivity of capital services becomes zero.

In terms of one of our earlier analogies, the manager of the King Ranch at each input of labor services combines with these labor services that quantity of the services yielded by the ranch such that the marginal productivity of the latter type of services is zero—so long as adding more of the latter type of services would cause a smaller output. If he had only a few laborers he would use only that amount of the ranch such that if he attempted to use more with this quantity of labor services he could secure only a smaller output by so doing.

Figure 3, below, depicts a portion of the firm’s production surface and the firm’s output-expansion or maximum total product curve (the line on which lie the points O, P, and M). Within the range OP, at each input of labor services (variable in acquisition and in use) a vertical section whose base is parallel to the axis representing capital services (variable in use only) is constructed, and the maximum point of the section located relative to variation in capital services. Then a line drawn to the capital-services axis from the base of a perpendicular to the maximum point on the section reveals the amount of capital services (e.g., OR) that are combined with that input of labor services (OD). In the early stages of production the maximum total product points (the points at which the marginal product of capital services equals zero) occur at numerical values of less than the given stock of capital services (OY) so that this entire stock is not utilized.

In this range in which capital services are less than fully employed, the output-expansion path consists of the straight line (OP) formed by these maximum total product points. Beyond this range labor-services inputs must be combined with the same amount of capital services (OY) so that the maximum total product curve switches direction abruptly at P, proceeding along the curve on which P and M are points.

Thus the range from O to P is one in which the quantity used of the stock of previously acquired capital services is adjusted to each labor-services input. Beyond this range, greater outputs result from combining larger labor-services inputs with the same quantity of capital services, so that returns to proportion succeed returns to scale and the total product curve becomes that on which P and M are points. In this second range the marginal product of capital is positive and—for a time—higher for successively larger labor inputs.

Within the range in which less than the total stock of capital services is employed, the marginal product of labor services is a total derivative and is constant. With a constant price for labor services the firm’s marginal cost is also constant within this range. The marginal product of capital services, on the other hand, is a partial derivative possessing a different level for each input of labor services and determining, by its equating to zero, the quantity of capital services utilized. Since the cost of the services of capital have already been incurred or contracted for, this cost does not affect the firm’s marginal cost curve. After all of the stock of capital services (OY) is employed, the marginal productivity of the services of labor becomes a partial derivative and declines, and the firm’s marginal cost curve rises.1

---

1 If L and C represent, respectively, the services of labor and capital and P represents total output, then \( dP = (\partial P/\partial L)dL + (\partial P/\partial C)dC \). In the range OP, in Figure 3 the marginal productivity of capital services \( (\partial P/\partial C) \) is a function of L and is equated to zero at each labor-service input. Beyond the range OP, dC is zero (for any positive changes). Consequently, in the range OP, the marginal productivity of labor services is a total derivative, and is a partial derivative beyond this range.

At least as early as Joan Robinson’s *Economics of Imperfect Competition* (London: Macmillan...
Thus in Figure 4, below, MP and AP are constant and equal for labor-services inputs over the range OS, input OS being the input at which all the stock of capital services enter into production. For labor-services inputs greater than OS, MP and AP decline. Figure 5 merely reflects the necessary implications concerning the firm's cost curves. Thus the firm reaches point P on the output-expansion path in Figure 3 at input OS of labor services (Figures 3 and 4). Also at point P (Figure 3) output OT (Figure 5) is secured.

Granting the argument thus far, the reader might still ask whether the range of outputs in which all of the services of the fixed factor are not being used constitutes a significant portion of the total range of outputs that the firm is capable of producing in the short run. While the answer depends partially on the relative magnitude of the stock of services in terms of which the short run is defined, it is largely an empirical question. Many empirical cost studies indicate that marginal costs tend to be constant very nearly up to capacity, suggesting that this range in which less than all of the stock of fixed-factor services is being utilized may indeed be a significant portion of the total range of outputs that the firm is capable of producing in the short run. The range OT of the short-run marginal cost curve (Figure 5) is, however, a reflection of returns to scale, whereas the remainder of the curve is a reflection of returns to proportion, whatever the relative lengths of the two ranges.

The process may be more readily apparent to some in terms of isoquants (Figure 6, below). The lettering employed in Figure 6 corresponds to that of Figure 3 (with the exceptions of F, N and U which will be discussed shortly) so that OY represents again the given stock of capital services, O, P, and M are points on the output-expansion path, P is the point along the path at which all the stock of capital services becomes employed.
and OS the labor-services input at which this occurs.

In effect the "price" of previously-acquired capital services is zero so long as some of these services are not entering into production (maintaining also the other assumptions of the basic model). Hence the "iso-costs" or "price" lines for labor services inputs of OS and less are horizontal, determining by their tangencies to the isoquants at points of zero slope the output-expansion path OP. If the assumption of zero user cost is replaced by an assumption that marginal and average user cost are equal and greater than zero, an output-expansion path such as OUM is generated, still yielding, however, a flat—though higher—marginal cost curve over the range of output in which not all of the stock of capital services (OV) is being utilized in production. If marginal user costs rise as output is increased, the marginal cost curve will be upward sloping even within this range.

Whatever the nature of user cost, returns within the early range of output are returns to scale, not proportion. Of even greater significance, this portion of the output-expansion path is a reflection of a particular type of entrepreneurial behavior and is not a mere technical datum given to the firm by an engineer. For at any given input of the services of the variable factor (within this range) the entrepreneur must choose that portion of the services of the fixed factor which should enter into production. Thus while the engineer may in effect provide the production surface, the output-expansion path on that surface results from the output-maximizing or cost-minimizing choices of the entrepreneur. It is this behavior in which we are primarily interested. In the immediately following section it is argued that there are many contexts in which we can look for such behavior.

A Sequence of Short Runs

In the analysis thus far it has been assumed that there is a unique short run definable in terms of the acquisition of the services of capital. Actually there are many short runs since a short run can be assumed with respect to the services of any factor of production or portion thereof. Thus there can be a short run for the period during which, in acquisition, the services of a building, of a certain type of machinery, or of any other component of "capital" are fixed. In essence, the device of the short run serves to limit the area concerning which the process of decision-making by the firm is to be analyzed, but the device itself is neutral. Thus the conventional short run in terms of plant says: "We know that a firm's decisions are influenced by all of its costs, but during the period in which the size of physical plant cannot change, the firm's decisions must be affected primarily by those costs which can vary. Therefore, let us analyze the firm's decision-making process during this period." However, there is nothing in the nature of this process which prevents its application to the period during which the services of a particular type of machinery are fixed, and, similarly, a period can be assumed during which the services of labor are fixed.

Stigler, op. cit.
SHORT-RUN RETURNS TO SCALE

By now assuming the factor which previously was considered the variable to be fixed (in acquisition) the attention of the analyst is directed to an area of decision making that is temporally of less duration. While it is true that conventional economic theory is “timeless” in the sense that the time necessary for many adjustments is often neglected, time is introduced into the analysis by the choice of long run or short run, and further, by the type of short run. By progressing from an analysis in which capital services are fixed and labor variable to an analysis in which both are fixed, time is being explicitly introduced, even though only relatively.

During the period in which the services of labor are fixed in acquisition but variable in use a great deal of the services of capital would be fixed in the same sense. If raw materials are considered the variable input we have the problem of two fixed factors—labor and capital. The analysis of the previous sections would suggest that, within the range of output in which less than the entire quantities of these services are utilized, the quantities used at each different input of raw materials are such that their respective marginal products are zero. This is still correct despite there being two fixed-in-acquisition but variable-in-use factors.10 It is only necessary to say that the firm’s output has changed. If a firm’s output increases, the quantity of factor services used has increased. Moreover, these services could all be fixed in acquisition but not in use. In our example, the raw materials could have been stockpiled and fed into the productive process at a rate chosen by the firm.

Thus there exists a sequence of short runs as a function of time, the choice of short run depending upon the area of decision making to be analyzed. In an extremely short time period all the firm’s factors are fixed in acquisition and in use so that the area of decision making is zero. As the periods increase in length, variation in use and then in acquisition become possible, so that the area of decision making increases so as to encompass questions of utilization of factor services already acquired, and finally, questions of factor acquisition. Thus there also exists a family of marginal cost curves (one for each short run), corresponding to the divisions made of entrepreneurial decisions, each such curve reflecting—to the extent that divisibility of the particular factor services defining it permits—a range of short-run returns to scale and the type of entrepreneurial behavior previously noted.

Application of the Analysis

The implication of the customary presentation of productivity theory is that the firm follows the law of variable proportions in the short run and the law of scale in the long run. It has been argued in the preceding sections that in the short run the firm more typically experiences returns to scale up to a constraint and thereafter returns to proportion. It has also been noted that the traditional short run is not unique, and that there exists a temporal sequence of short runs corresponding to the divisions that may be made of entrepreneurial decisions.

The fact that the earlier (in terms of output) reaches of the output expansion path are the result of a process that involves considerations of scale rather than of proportions will gain greatly in cogency for most readers if reasons are given for expecting output to occur in this range.11 And

10 To secure the maximum output for a given input of raw materials the partial derivatives of output relative to labor services and capital services would each have to equal zero. (The other conditions of a maximum would, of course, also have to be met.) There is, of course, no assurance that sufficient stocks of both types of services would be on hand to permit maximization of output relative to both. The firm’s labor services might all become employed, e.g., at an output which some of the services of capital did not cooperate in producing.

11 I.e., reasons more varied and specific than
the fact that there exists a temporal sequence of short runs will become more significant if we can point to situations which can most readily be analyzed in terms of short runs other than the traditional short run defined in terms of physical plant. In the present section it is argued that in the production of services are to be found many examples in which the problem of producing in the early reaches of the output-expansion path is implicitly or explicitly recognized and which can also best be defined in short-run contexts other than the traditional.

By "services" in this context we mean intangible economic goods. We select this category of goods because in many instances the production of services can only occur simultaneously with consumption or purchase. Since the output of services produced must in these instances coincide with the volume of sales transactions or consumption, and since the volume of sales transactions or consumption is often subject to wide fluctuations over relatively short periods of time, the output of the firm is subject to these fluctuations. As a result, output may often occur in the early reaches of the output-expansion path, i.e., the range of short-run returns to scale. In other instances in the production of services there exist devices or stratagems which can be interpreted as measures designed to avoid or compensate for the occurrence of output in the range of short-run returns to scale. Both types of cases attest to the plausibility of the assumption of divisibility, in use, of the services yielded by the factor of production in terms of which the short run chosen is defined.

Both the manufacturer and the producer of services have the handicap that they must purchase "inventories" of factor services many of which are automatically yielded through time at a relatively constant rate. The producer of services often has the additional disadvantage, however, of less control over his rate of output. Consider the retail firm. The retail firm's "output" is determined primarily by the number of customers it has, and since this number will vary temporally during the hours of a day, days of a week, and weeks of a month, an evenly distributed time frequency of sales transactions is not to be expected during these periods. If variation in utilization of factor services already acquired is a criterion for the selection of the type of short run by the analyst, it is clear that an important short run for the retail firm is the short run in terms of labor services.

In the labor short run the firm determines, at each output, the quantity of labor services to use (within the range in which not all of the firm's labor services are being utilized) by equating the marginal productivity of labor services to zero. Beyond this range, of course, all of the stock of labor services is employed. In both ranges, however, the level of the marginal physical productivity of labor services is different for each output.\footnote{In the labor short run the marginal product of labor services is a partial derivative of output relative to these services, and is a function of the quantities of other factor services utilized. Since the cost of labor services has already been incurred or contracted for, in arriving at any output in the range in which not all labor services are used these services are added until their marginal product is equal to zero. Larger outputs within this range permits more labor services to be added before this equality (or marginal product to zero) occurs than is true of smaller outputs.}

In the labor short run sales volume fluctuations determine the quantity of the services of this factor that are utilized. Since these factor services must typically be purchased or contracted for in advance, the employer, in a longer period, in arriving at the decision of the quantity of labor services to acquire must estimate and discount some average of the various levels of the marginal productivity function which will occur
during the period in which he is committing himself to provide employment. Since these levels vary with output his task is basically that of estimating some average future sales volume during the labor short run.

The producer of services is caught between Scylla and Charybdis. If he acquires too small a quantity of a factor whose services are yielded at a relatively constant rate through time, he will be limited in his ability to meet sales volume peaks by the necessity of combining his factor services in less than optimum proportions. If he acquires too large a quantity much of the total factor services yielded during the period may be lost. And it is noteworthy that even if sales volume peaks and troughs are of equal magnitude, the effect of the fluctuation cannot be removed by the fortunate acquisition of that quantity of the factor such that its services accrue at the "cycle average" rate. Having ten too few salesmen between the hours of one and two is not offset by the fact that between eleven and twelve an equal number were standing idle.

Thus the "peak off-peak" problem is a characteristic of the production of services which compels many entrepreneurial decisions to be made in a context other than the conventional short run in terms of physical plant, including the determination at some outputs of how many already-acquired factor services should be omitted from the production function. In many instances this determination is automatically made for the entrepreneur, but this renders the necessity of operation in a range of short-run returns to scale (with its concomitant of non-use of factor services previously acquired for a financial outlay) nonetheless painful.

The reactions of the entrepreneur to the fact that he has acquired some factor services which he cannot in his present circumstances utilize takes forms with which we are all familiar but which we may have been hindered in understanding by the conventional assumption that the services of fixed factors are fixed in use as well as in acquisition. Nonetheless, the fact that stocks of the several factor services must be acquired in discrete and to some extent incompatible "lumps" is a prime reason why the entrepreneur is confronted with the question of what to do about excess factor services. Many of the services of a building, once purchased, accrue throughout the twenty-four hours of each of a large number of days, while the services of the labor employed therein may be purchased by the entrepreneur for only eight hours of each of five days during each week. Moreover, it may be necessary to exclude from the production function some of the services of both of these factors during certain days or hours of the labor work week. The excess services of the building during a sixteen-hour period of each day may be an important factor in the decision to engage in multiple-shift operations and the excess services of the building and labor during the labor work week may be a determining factor in the decision to engage in multiple-product or multiple-service operations. If these factor services were not divisible in use their entry into the firm's production function could not be prevented and the thought of using them in some other manner could not be entertained.\footnote{We have assumed, in the analysis of the preceding sections, that user cost is zero and thus that the firm can not devote excess factor services to alternative uses during the period under consideration. Some, but not all, of the excess factor services could be diverted to other uses, however, without affecting the analysis so long as marginal user cost was zero. The range of output for which this was the case would, of course, be shortened and non linearity of marginal cost introduced at a smaller output. Thus the marginal cost curve of one of the products of a multi-product firm may more nearly resemble the "conventional" short-run marginal cost curve than that of the single-product firm.}

In terms of the analysis of the preceding sections, indivisibility in the acquisition of stocks of factor services leads often to the imposition of the conflicting constraints upon
the production surface. At the output at which all of one stock of factor services previously acquired is being utilized, some of the services of another factor, also previously acquired, may be unemployed. Furthermore, lesser outputs than this may mean that a portion of the services of both of these factors is unemployed.

The entrepreneur may have little recourse if he possesses excess labor services but is utilizing fully the services of his physical facilities, for the former have presumably been substituted for the latter to the extent made possible by the elasticity of technical substitution. But if some of the services yielded by his physical facilities are not being utilized in production and if he has used as his guide (on the cost side) to output and price the marginal cost curve appropriate to the short run in terms of the acquisition of these services, then when he observes the labor short run and finds unemployed labor services during this period he may well resort to the marginal cost function defined in terms of this short run and decide to lower his price and increase his output.

Thus the old charge that “excess capacity” leads to a downward pressure on prices is, in many instances, correct and the conventional refutation that since fixed costs cannot affect marginal costs the former cannot affect price or output is often wrong. Only if the stock of factor services, the acquisition of which occasioned the fixed costs, consists of factor services which are indivisible in use can it be safely asserted that “excess capacity” cannot lead to a downward pressure upon price. What is conceived as custom by one producer and thus priced in the context of a short run in terms of physical plant may be the opportunity of a second producer, acting in the context of a much shorter short run, to employ unused factor services, their “involuntary unemployment” stemming from the unevenness of sales transactions, indivisibilities in acquisition, or a miscalculation in the quantity of factor services acquired (including a failure to foresee that a third producer was going to “dump” excess services).

Since it is often difficult or dangerous to the stability of the existing price for the entrepreneur to attempt, by one stratagem or another, to devote his unemployed factor services to a market closely related to that of their gainfully employed counterparts, he may devote them to markets in which the cross elasticity of demand is very low. Important to his decision to do so is also, of course, the user cost of his excess services and the cost of any additional factor services that he would have to acquire. In the production of services the limiting case is perhaps that of the empty backhaul in transportation. In this case the producer on the return trip automatically leaves some of his factor services unemployed and is forced, therefore, to operate at a particular point in the early range (the range of returns to scale) of the short run defined in terms of the services that were acquired to make the round trip. In this limiting case it is ordinarily conceded that the relevant marginal cost for the pricing of any unforeseen return-trip traffic which the producer may secure is primarily the marginal user cost of the services which have already been committed to the return trip. The producer in this short run will charge a lower price for the return-trip transportation even if the elasticity of demand therefore (at a given price) is the same as that for his out-bound service.

The only uniqueness in the limiting case, however, is that very few additional factor services need be acquired, that user cost is very low, and that the market for return trip transportation is quite separate from that for outbound transportation. Whenever a producer has unemployed factor services he may devote them to production in a separate market, guided by the cost of services it is necessary to acquire, by user cost, and by the effect that production in the additional market is apt to have upon his primary market. Thus more frequently than we
might otherwise have supposed, price discrimination may arise not because of differing demand elasticities in the several markets at a given price but because of differences in the marginal costs that are relevant to these markets.

The cases described thus far have been treated as though the entrepreneur, after determining the quantity of previously-acquired factor services that are excess at a given output, has only the options previously indicated; i.e., to devote them to the production of other economic goods, to the production of the same service for other markets, to shift to another short-run context which would call for more output in the same market, or to lose the excess factor services. But this list of options (all of which require divisibility in the factor services, stocks of which have been previously acquired) is by no means exhaustive.

Another stratagem by means of which the producer of services compensates for sales volume fluctuations is exasperatingly familiar. Whereas the producer of tangible products can cushion himself against sales volume fluctuations to a considerable extent (and thus produce at a relatively constant rate) by the device of product inventories, the producer of services can—to a more limited extent—inventory customers in order to avoid production in the range of short-run returns to scale. Perhaps the limiting case in this respect occurs in the production of professional services. Each of us on occasion has been a part of a physician's "customer inventory" which helps to avoid loss of previously-acquired factor services during portions of the periods for which they have been acquired. By this device the physician minimizes that portion of the factor services yielded by his building, his equipment, his staff, and his own labor that do not enter into the productive process. A short run defined in terms of the acquisition of any of these productive factors would reveal that some of the factor services yielded by them may not enter into produc-

tion unless some such device as a "customer inventory" is employed. Such a possibility (of factor services not entering into production) could not occur if these services were not divisible in use.

Services other than professional services also have (to a perhaps lesser extent) "customer inventories," and some price concessions may be made in order to secure them. The occurrence of short-run returns to scale may also lead to an alteration in the nature of the service produced so that the factor service mix contains proportionally less of those factor services which, once acquired, are yielded at a relatively constant rate through time. (Witness the development of supermarkets in retailing, and auto, truck and trailer rentals in transportation, in which the service rendered is altered so as to require the firm to acquire relatively less labor services.) Such developments have often, of course, been spurred by rising factor prices but would have occurred, in many instances, at constant factor prices, the inability to use many of the factor services previously acquired being adequate incentive.

Thus in the production of services there are many instances of operation in the range of short-run returns to scale and of stratagems which can be interpreted as devices to avoid or compensate for this occurrence.

CONCLUSIONS

In the traditional theory of the firm the production function is viewed as a given technical datum. It is viewed as automatically depicting the maximum outputs that can be achieved by combining various quantities of factor services. If the services of the fixed factor are assumed indivisible in use as well as in acquisition these two points are quite consistent, for no enter-

prise-preneurial decision (other than the decision to use the specific technological process involved) affects the quantity of output that can be secured from a given set of inputs.

If for some outputs the marginal physical
productivity of the services of the fixed factor is negative the assumption of indivisibility in use implies that these services are nonetheless committed to production and that any entrepreneurial behavior to prevent this occurrence is ruled out, in the short run.

But if the indivisibility which characterizes the acquisition of factor services does not characterize their use, the production function does not automatically yield the optimum outputs achieved by combining various quantities of factor services—for the assumption of divisibility in use plus the customary assumption of rationality implies a specific form of entrepreneurial behavior if those optimum outputs are to be achieved.\(^{14}\) Thus for combinations in which some of the services of the fixed factor would otherwise have a negative marginal physical productivity the entrepreneur must determine the quantity of services of which this is true and prevent their entrance into the productive process.

\(^{14}\) The production function, of course, shows all outputs than can result if various factor service combinations are used. But a decision to alter a given factor service combination, although guided by the production function, is a managerial or entrepreneurial decision. Thus a decision to preclude from entrance into the productive process some of the services yielded by a fixed factor is a form of output-maximizing or cost-minimizing behavior that is not part of the production function, \emph{per se}. In other words, the production surface is comprised of points depicting all possible combinations, but a choice of any point is an entrepreneurial decision.

The more realistic view of the services of the fixed factor is that these services are divisible in use. Given divisibility in use (as distinguished from acquisition) the traditional short-run, long-run dichotomy does not correspond to the dichotomy of returns to proportion versus returns to scale. Instead, returns to scale precede returns to proportion, in terms of output. Given that divisibility implies constant returns to scale (and assuming constant factor prices) short-run marginal cost curves are flat over the first of these ranges of output. Thus the short-run output-expansion path of the firm is not merely the upper edge of a section through the productive surface at a given quantity of the services of the fixed factor but is instead (for a portion of its length) the result of an operation performed on that surface by an entrepreneurial decision. Thus the curve of maximum outputs which underlies the firm's cost curves is in part behavioral and is not merely information supplied to the firm by a technician. The choice of divisibility in use rather than indivisibility in use alters significantly the process by which the firm's short-run cost curves are derived.

Many examples of the output-maximizing behavior which characterize the early reaches of the output-expansion path exist and become particularly evident when it is recalled that there is no unique short run. Instances of this behavior and of behavior related to it are also particularly apparent in the production of services.