ECONOMIES OF SCALE IN THEORY AND PRACTICE

by

Richard G. Lipsey,

Emeritus Professor of Economics at Simon Fraser University

and

Fellow, Canadian Institute for Advanced Research,

January 2000

This paper replaces an earlier version called
"Capital and Scale"

Simon Fraser University

Harbour Centre

515 West Hastings Street

Vancouver, BC

V6B 5K3

Voice: (604) 291 5036, Fax: (604) 291 5034

rlipsey@sfu.ca

http://www.sfu.ca/~rlipsey

ECONOMIES OF SCALE IN THEORY AND PRACTICE

ABSTRACT

This paper considers a two-stage production process in which capital goods are produced in stage 1 and their services are used in stage 2 as inputs into a production function for final goods. Altering production in the second stage requires altering the rate at which identical units of some well-specified commodity are produced. However, altering production at the stage 1, in order to deliver an altered rate of flow of the capital service at stage 2, typically requires using a differently designed capital good. These design changes are a major source of variable returns to scale that are concealed when the production process is modeled as a one-stage process, as it is in neoclassical production theory. Empirical evidence shows that countless scale effects exist at stage 1 in real world production processes. These have nothing to do with indivisibilities and are consistent with constant returns to scale in the neoclassical, one-stage production function.

KEY WORDS AND PHRASES: Capital, economies of scale, two-stage production, indivisibilities, increasing returns.

TABLE OF CONTENTS

(*For readers' use only. Not for publication*)

ECONOMIES OF SCALE IN THEORY AND PRACTICE

Nineteenth century Austrian capital theorists stressed that capitalist production is two-stage production: First a capital good is made, then it is used to provide service inputs in the production of some final good. This paper argues that the neoclassical theory of scale effects obscures important relations by ignoring the Austrian two-stage process. Instead, it deals with capital goods only in terms of their service flows that enter into the production function for final output.

The paper distinguishes two production stages. In stage one, a capital good is produced and in stage two, the services of that good are combined with other inputs to produce some final output. When this is done, scale effects are seen to be common and to be located in the first-stage production of capital goods.

I. DEFINITIONS

It is conventional to distinguish three sub-components of capital: plant and equipment, residential buildings, and inventories. This paper is confined to plant and equipment, broadly understood as being everything that is not in the other two categories. *Returns to scale* refers to the response of output to a proportionate increase in all inputs with technology held constant. *Returns to outlay* refers to changes in the unit cost of output when that output is varied in the long run with input prices held constant.[1]

The capital goods that deliver capital services can be altered in two distinct ways. They may be r*eplicated,* which means creating units identical to those already in use. If the unit under consideration is one machine or one factory, another identical machine or factory may be installed. Replication applies to increases of the original capital good by integer multiples (and, where replication has already occurred, it also applies to reductions by integer multiples). Second, the capital may be reconfigured, which means altering its physical specifications. In response to the need to alter the rate of output, a differently constituted machine may be used, or a differently specified factory may be built. If the capital reconfiguration takes place by substituting among a stock of known technologies—in the Marshallian long run—this is called *long run reconfiguration.* If it takes the form of newly developed technologies it is called *very long run reconfiguration.*

II. SCALE EFFECTS IN THE NEOCLASSICAL PRODUCTION FUNCTION

The standard theory of production uses a single-stage production function showing output of the i^{th} product as a function of the flow of n inputs:

$$X_i = \psi_i(s_1, s_2, \ldots, s_n). \qquad (1)$$

These inputs include materials (what in the aggregate, the classical economists called “land”) and the services of various kinds of capital and of labour. In standard micro-theory textbooks, (1) is usually assumed to be a linear homogeneous function so that multiplying all inputs by the scalar, λ, multiplies the output by λ.

Sometimes constant returns are simply assumed because of its tractable characteristics. Sometimes, however, its existence is argued. The argument typically starts with some version of

the proportionality postulate. This was stated by Koopmans (1957: 76), one of the founders of the formal approach to production theory, as follows:

> "if an activity $a = (a_1, a_2, ..., a_n)$ is possible, then every activity $la = (la_1, la_2, ..., la_n)$ of which the net outputs are proportional to those of a, with a non-negative proportionality factor, l, is also possible."[2]

According to Koopmans the proportionality postulate implies that increasing returns to scale are impossible unless there is an *indivisibility* or lumpiness in one or more of the inputs. If there were no such indivisibility of inputs, the technique that proves superior at the higher scale could, in his words, always be "*subdivided proportionately*" to produce efficiently at the lower scale.[3] It seems clear that, in this context, Koopmans, means by "indivisibility" the impossibility of reducing all the inputs $a_1,...,a_n$ by the same proportion $0<\lambda<1$.[4]

Notice that this appears to deduce an empirical proposition—scale effects can only result from indivisibilities in capital goods—from a highly abstracted formulation of a production function. No empirical constraints are used in its formulation. Neither are capital goods explicitly modeled. The correct deduction is not that scale effects can only result from a lumpiness of inputs, but that this abstract formulation has removed all possible sources of scale effects other than those that are associated with some characteristic of the input flows, of which lumpiness is the obvious candidate.

Of the many problems with Koopmans' formal argument, the most basic one is in the high level of abstraction of the proportionality postulate. The reader is never invited to consider what "subdividing an activity proportionately" might mean in particular empirical circumstances. So let us take the unusual step of asking what the postulate means in the case of some specific real-world production activity.

Assume that activity *a* is the manufacture of a steam engine that will deliver 100 horsepower. The inputs are a specified set of materials, power, the services of various machine tools, and the services of various types of labour. Assume that one now wishes to produce a steam engine that will deliver 50 horsepower. "Subdividing the activity proportionately" might mean reducing all the dimensions of the steam engine in equal proportion. Or it might mean altering all of the inputs—materials, machine tool hours, power, labour, etc.—in equal proportion. Notice that these are alternative concepts of "subdividing the activity".

First, consider altering the inputs. Some inputs are three-dimensional solids (e.g., the base on which the engine is mounted), others are two-dimensional surfaces (e.g., the walls of the boiler), and yet others are effectively one-dimensional (e.g., the wiring). So building a working engine that uses exactly λ as much of each input is likely to be physically impossible and certainly would result in a technically inefficient engine, as pointed out in the basic engineering literature.

Next consider shrinking all of the dimensions of the engine in the same proportion. Because of the different dimensionality of the parts, this will alter the material contents of the engine in different proportions. Also scaling all of the engine's dimensions by some fraction λ and providing it with λ as much fuel will *not* alter its output of power in the proportion λ. For one reason, the ratio of heat loss through the engine's cylinders to the power delivered by those cylinders rises as the size

of the cylinder falls. (Since the time of Isaac Newton, practical engineers have known that a small body loses heat faster than a large body. See for example Cardwell (1995: 158).)

In summary, one can build a steam engine that will produce 50 rather than 100 hp, but this cannot be done either by scaling all the dimensions of the 100 hp engine by 50% nor (what is *not* the same thing) by reducing all the inputs that go to make up the engine by 50%.

A neoclassical theorist may object at this point that the above example does not deal with the type of situation envisaged in the neoclassical production function, which deals only with the flows of capital services and other inputs, not with the conditions under which these flows are produced. However, as is explored in the next section, the conditions under which capital goods having various output capacities are designed and produced lies at the core of many scale issues. One cannot know how the unit cost of a capital service behaves as the volume of service flow is altered in the long run unless one knows how the cost of producing a capital service varies as capital goods with different service-flow capacities are constructed.

II. AN EXAMPLE OF SCALE EFFECTS IN TWO-STAGE PRODUCTION

To go further, the two-stage production process needs to be modeled explicitly. The analysis begins with the simplest possible example of a wide range of cases where the scale effects depend on dimensionality rather than on physical relations. This example is chosen because its transparency allows the issues to be easily identified. It concerns a firm that is in the business of pasturing other people's horses. One square unit of fenced space is required to accommodate one horse. The grass is free and the only production cost is the fence, which is continuously variable. When the firm wishes to pasture more horses, it increases the size of its one fenced field.

This example contains no indivisibilities, since inputs can be varied continuously. The reconfiguration of the capital good is the simplest type, since output of the capital service is increased by using more capital with identical physical specifications.[5] Nonetheless, there is a scale effect in the two senses that when final output changes, there is a change in both the quantity required of the single input per unit of output and in the unit cost of output. When a second input is added, the optimal input ratio alters as the scale of output alters even though the second stage production function for the final good is linear homogeneous. This example thus refutes the generalization that scale effects *must be* rooted in input indivisibilities. It also shows that scale effects can occur even when the relevant neoclassical production function displays constant returns to scale.

A. A Single Input

First, write the firm's production function in terms of the input of capital services. Letting this input be *P*, measured in square acres of pasture, and the output be *H*, measured in the number of horses pastured, the function is:

$$H = P. \qquad (2)$$

This is the neoclassical production function relating inputs of factor services to a flow of output. It obeys constant returns to scale, just as it is supposed to do. The function is, however, the second stage in a two-stage production process.

Now write the production function for the capital service, pasture. The factor input, *F*, measured as the total number of feet of fence[6], produces a capital good, *P*, measured in square feet of pasture:

$P = (F/4)^2$. *(3)*

This production function is the first stage in the two-stage production process.[7] It displays increasing returns since

$dP/dF = F/8$ *(4)*

is increasing in *F*. So, as the length of the fence is varied by the multiple l, output of pasture acres varies by l^2.

In this two-stage process, the inputs that go into the first stage production of the capital good can be called the primary inputs —the fence in this case—and the outputs of the capital goods that is the input into the second stage production of the final good, the intermediate inputs—in this case the service of fenced pasture.

Next, substitute the capital service production function of (3) into the output production function of (2) to get

$H = (F/4)^2$. *(5)*

This is the production function for final output in terms of primary input, the fence that makes the capital good that provides the capital service of protected pasture. It displays increasing returns since

$dH/dF = F/8$

is increasing in *F*.

Finally, letting a foot of fence cost one unit of money, the total cost of fencing the pasture is *F*, and the cost per pastured horse (unit cost of output), is

$C = F/H$

$= 16/F$ *(6)*

which is monotonically decreasing in *F*, giving increasing returns to outlay.

All this is obvious and no doubt tedious. But the demonstration is important for that reason. There are *no indivisibilities* in this example. The physical nature of the capital good is unchanged and the area of the pasture is a continuous variable. The neoclassical production function, defined in terms of inputs of service flows, displays constant returns to scale. Yet there are scale economies. These are rooted in the geometry of our three-dimensional world. The fenced area increases with the square of the length of the fence, while the cost increases linearly with the length of the fence. These economies will, however, never be seen, if production functions are modeled only in terms of the second stage of production, and no account is taken of real physical relations that govern the production of capital services by the capital good.

B. Two Inputs

Let us now add a second input. Assume that each acre of land needs planted grass, *G*. There is a substitution between the two factors in the sense that a horse can be successfully pastured on a small amount of land that is heavily grassed, or a large amount of land that is sparsely grassed. In this two-stage process, the primary inputs are fence and grass, while the intermediate input is acres of pasture.

An example of a neoclassical, constant-returns production function for the final output that has these characteristics is:

$$H = P^{\alpha}G^{1-\alpha} \ , \quad 0 < \alpha < 1 \qquad (7)$$

Because this is a homothetic function, the ratio of the two inputs, *G* and *P*, is unchanged, as the scale of operations increases with input prices held constant. Letting *w* be price gives:

$$\frac{G}{P} = \frac{w_P}{w_G}\left(\frac{1-\alpha}{\alpha}\right) \qquad (8)$$

In this case, however, the prices that must remain constant are those of the primary input, grass, and the intermediate input, pasture.

Next substitute the production function for the intermediate capital service, *P*, from (3) into (7) to get:

$$H = [(F/4)^2]^{\alpha}G^{1-\alpha} \qquad (9)$$

which is the production function for final output in terms of the primary inputs. Because this function is homothetic, the ratio of these inputs also remains constant as output rises, provided that the prices of the two primary inputs, F and G, remain constant:

$$\frac{G}{F} = \frac{w_F}{w_G}\left(\frac{1-\alpha}{2\alpha}\right) \qquad (10)$$

Given the two production functions, however, it is not possible to hold the two ratios, G/P and G/H constant simultaneously. Wherein lies the contradiction? The problem is that the neoclassical one-stage model, assumes that the cost of intermediate service flows that are inputs into the second stage production function of the final good remains constant. But when the output of the final good is increased, the output of the capital service input must also be increased. Given the production function for producing the capital good, the unit cost of its service cannot remain constant. To obtain the implicit price of a unit of fence, notice that the total cost is:

$$TC(P) = w_f F = 4w_f\sqrt{P} \qquad (11)$$

The marginal cost of fencing is:

$$MC(P) = 2w_f P^{-1/2} \qquad (12)$$

Letting the price of the pasture inputs into the next stage of production be equal to the marginal cost of creating more pasture (i.e., $w_p = mc_p$), (8) becomes:

$$\frac{G}{P} = \frac{2w_f P^{-1/2}}{w_g} = \frac{(1-\alpha)}{\alpha} \tag{13}$$

or

$$\frac{G}{P^{1/2}} = \frac{2w_f}{w_g} = \frac{1-\alpha}{\alpha} \tag{14}$$

So the ratio of the inputs into the second stage production function does not remain constant. Instead, G falls relative to P as production rises. This substitution allows the firm to reap some of the scale economies in the production of the capital good.

The geometry of pastures is such that the cost of producing a given area of pasture is a diminishing function of the number of acres fenced. This scale effect is inconsistent with the typical neoclassical experiment in which the scale of production rises as the cost of service inputs remains constant in the second-stage production function for final output. But the inconsistency goes unnoticed when the conditions under which capital service flows are produced are not investigated. (This analysis has important implications for the measurement of productivity and technological change that is postponed for another paper.)[8]

IV. SCALE EFFECTS MORE GENERALLY

Four important sources of scale effects that can arise when output is varied in the long run are considered below: the geometrical consequences of reconfiguration, the physical consequences of reconfiguration, the cost of embodying capital services in capital goods, and the more intensive use of indivisible inputs.[9]

A. Scale Effects Arising from Geometrical Relations

We live in a three-dimensional world, which entails many scale effects, both increasing and diminishing. These effects typically require neither input indivisibilities nor violation of the neoclassical assumption that the production function for final output displays constant returns to scale.

The geometrical relation governing any container typically makes the amount of material used, and hence its cost (given constant prices of the materials with which it is made), proportional to *one dimension less than* the service output, giving increasing returns to scale over the whole range of output (at least with respect to the inputs of materials).[10] This holds for more than just storage. Blast furnaces, ships, and steam engines are a few examples of the myriad technologies that show such geometrical scale effects.

Costs of construction also often increase less than in proportion to the increase in the capacity of any container. Consider just one example. The capacity of a closed cubic container of sides s is s^3. The amount of welding required is proportional to the total length of the seams, which is $12s$. The amount of material required for construction is $6s^{2.}$ So material required per unit of capacity is $6/s$ while welding is cost is $12/s^2$. Not only are both of these falling as the capital good is reconfigured to increase its capacity; they fall at different rates.

The reason why these, and other similar scale effects discussed below, only show up when the first stage of the two stage process is modelled is that in the second stage the problem is to *replicate identical units of final output*. One seeks to alter the rate of production of some final product that has a specified and unchanging design. The ubiquitous real world scale effects only show up when one reconfigures some product to make it produce a different rate of service output. This occurs whenever one needs to reconfigure the capital good used to produce the service flows that are required for an altered rate of production of the final good.

B. Scale Effects Arising from Physical Laws

In most cases of long run reconfiguration, a different design of capital goods is required if a different capacity rate of service flow is required. The physical nature of the world in which we live typically implies non-constant returns to outlay: the cost of producing a unit of the capital service varies as the output capacity of the capital good is varied. If these effects are to be allowed for in the neo-classical one-stage formulation, they will show up as a change in the price of a unit of capital services produced from the reconfigured capital goods. Here are a few of the many possible examples.

***Containers*:** The geometrical reasoning given above cannot produce a final conclusion about scale effects; one needs to know some physics as well. It is imaginable, for example, that as the capacity of a container is increased, the walls would need to be thickened proportionally, making the volume of material increase linearly with the container's capacity. Physics tells us that in most cases this is not so. Although some thickening is often required, in many (probably most) cases, the thickening is less than in proportion to the increase in the surface area. In such cases, the volume of material increases less than in proportion to the increase in capacity (although more than in proportion to the increase in surface area).

***Light bulbs*:** To make a light bulb last longer, all that is required is to alter the strength of the filament without a proportionate change in all of the other materials that make it. To make a light bulb deliver a larger wattage of light, all that is needed is to change the resistance of the filament with no change in the glass or the socket of the light bulb. This gives increasing returns to reconfigurations designed to alter either the duration or rate of flow of the services of the light bulb over a wide range of duration and wattage.

***Ships*:** There are many scale effects associated with ships. First, the maximum speed that a displacement hull can be driven through the water is proportional to the square root of the length of the hull on the water line (planing hulls obey different laws). No amount of *a priori* reasoning could reveal this rather mysterious relation (Hiscock 1965:138). Second, while a ship's carrying capacity is roughly proportional to the cube of the surface area of its hull, geometrical relations plus the physics governing structural strength of a hollow body dictate that the ship's cost is related approximately linearly to the hull's surface area. Third, altering the ship's size also alters its handling and safety characteristics in complex ways. Fourth, as the size and other characteristics of a ship are changed, there is an alteration in the materials best used for its construction.[11] Thus building larger ships alters carrying capacity, construction costs, operating costs, speed, and other handling characteristics each in a different proportion.

Structures: If all the dimensions of a bridge are altered in the proportion λ, its structural strength is altered by $1/\lambda$ and its weight is altered by λ^3 (under the simplifying assumption that it is optimal

to use the same types of materials in bridges of all sizes). In other words, bridges and other similar structures, exhibit diminishing returns in the sense that as their size and the amount of materials used in their construction is increased, their strength increases less than in proportion (See Adams (1991: 81).) This is one of the most important sources of diminishing returns to reconfiguration that limits the extent to which other sources of increasing returns can be exploited by building larger versions of some generic capital good.

Blast furnaces: According to the physics of heat, the heat loss from blast furnaces is proportional to area of its surface, while the amount of metal that can be smelted is proportional to the cube of the surface sides. This is a source of increasing returns in the relation between fuel used and output capacity of such furnaces.

Steam engines and electric motors: As with blast furnaces, the heat loss from a steam engine's cylinder is proportional to the cylinder's surface area while the power it generates is proportional to the volume of the cylinder. This is one of the several reasons why the thermal efficiency of a steam engine is an increasing function of its size over a wide range starting from zero. This in turn is why steam powered factories were built much larger than the water powered factories that they displaced. There are no similar effects with electric motors, which is one reason why small-scale parts manufacturers (feeding into large-scale assemblers) became efficient when electricity replaced steam as the major power source of industry.

These examples illustrate that when the rate of output is altered in the long run, and capital is altered by reconfiguration rather than replication, the nature of the world in which we live will almost always produces a complex set of reactions, some tending to reduce the unit cost of output, some tending to increase it. Other reactions will alter the capital good's performance characteristics in ways that are only indirectly reflected in the relevant service flow. With reconfiguration, the experiment of multiplying all inputs by the same constant l is irrelevant. Engineers are never given a bundle of inputs and told to design a piece of capital. Instead they are given the output specifications and told to design the most efficient capital good to produce it. These examples also show that the neoclassical one-stage production function stated in terms of input services, may display constant returns while the whole production process has increasing returns, because the cost of a producing a required capital service flow falls as output increases. This is typically due to technical relations embedded in the nature of the capital goods themselves, which do not display constant returns to inputs when they are reconfigured (even within the confines of known technology). Another way of putting this is that the one-stage production function assumes constant something that is endogenous and cannot remain constant, the cost of producing inputs of the services of a capital good whose capacity must be changed in response to long-run changes in final output.

The entries in the *New Palgrave Dictionary of Economics* that are relevant to scale economies offer little help on the technological sources of scale effects. There is only one mention of a technological relation in all of the articles. Eatwell (1987 vol. 4:166) states that "There are some examples in which outputs are an increasing function of inputs for purely technological reasons" and goes on to cite the relation between the capacity of a pipeline and the material used in its construction as the diameter of the pipe is increased. One would hardly guess from Eatwell's article, let alone from the other *Palgrave* articles relevant to the subject in which not one technological relation is mentioned, that technological relations lie at the heart of so many

observed cases of scale economies (Baumol, Becattini, Oi, Silvestre and Vassilakis). For example, in his article "Economies and Diseconomies of Scale" Silvestre (1987: 80-83) mentions only indivisible inputs, set up costs, and Adam Smith's division of labour as sources of economies of scale.

C. Scale Effects Arising from the Technology of Producing Capital Goods

Early Austrian capital theorists asked: Why is roundabout production chosen when it requires waiting? They answered, without wholly convincing arguments, that the explanation lay in the superior productive power of indirect (capital using) over direct (non-capital-using) means of production. Eaton and Lipsey (1997) have sought to make this proposition more secure by deducing universal scale effects in the technology of producing capital goods on the basis of a very small input of empirical knowledge. What follows is a short intuitive version of the main points in their formal proof of this matter.

Assumptions: (1) There is a positive cost of waiting (i.e., the interest rate is positive). (2) Capital goods are needed to yield given flows of services over time and a decision must be made on the amount of durability to build into these goods. (3) The technology of building capital goods displays constant returns in the sense that the reconfiguration of a capital good to embody λ more or less capital services implies that the cost of producing that good changes in the proportion, λ.

Implication: Interest costs are minimized by minimizing the capital good's durability.

Empirical observation: Virtually all real capital goods could be made to be less durable than they now are. So if the amount of services embodied in a capital good is a measure of its "lumpiness", and if embodied services vary directly with durability, *endogenous lumpiness* must be created when capital goods are produced (i.e., the decision is made to embody more services in a capital good than is physically necessary).

Contradiction: The empirical observation of endogenous lumpiness is inconsistent with the implication drawn from the three basic assumptions: unit costs are minimized by minimizing durability.

Conclusion: Since assumption (1) is known to be true and assumption (2) may be taken to be to verified by observation of the great majority of capital goods (this author can think of no exceptions), the conclusion is that assumption (3) must be false. This assumption is then altered as follows: *There is a universal scale effect in embodying services in capital goods: as durability of the capital good is increased, there is some range starting from zero over which the services that it embodies rise faster than the cost of adding to the good's durability.*

This scale effect appears to be rooted in the physical nature of durable goods that yield their services over long periods of time. It seems likely that this is the scale effect that the Austrians were looking for when they tried to explain the efficiency of roundabout production.

D. Scale Effects Arising from Indivisibilities

Analysis based on the neoclassical production function is based on the implicit assumption that there is a real world counterpart to scaling all activities upwards or downwards proportionally by altering all inputs flows in equation (1) by some constant $\lambda>0$. As has already been observed, the

abstract nature of the usual treatment precludes asking about the real world counterpart of such an operation. Let us look into this matter further.

1. The meaning of indivisibility

To get further with the issue of indivisibilities, it is useful to distinguish three ways in which a capital good can be described. The first is its internal makeup, a blue print of how to make it. The second is the inputs that go into making the good. The third is the good's service output. The first may be called its physical descriptor, the second its input descriptor, and the third its output descriptor.

Now let us consider what can be meant by the term indivisible when applied to some capital good. To begin, notice that the terms divisibility and indivisibility can be used in each of the ways in which a capital good has been described, physical indivisibility, input indivisibility and output indivisibility.

First, consider the physical descriptor of a capital good. In the physical sense, a *divisible good* can be defined as one that can be divided into two or more parts, each of which is indistinguishable from the whole in any relevant aspect, except size. Two halves of a divisible good each working on its own, will produce exactly the same services as one whole good. Examples are seeds, fertilizer, and flour, where two x pound bags can do the same job as one $2x$ pound bag. Each of these items is divisible down to a single small unit such as a grain or a molecule.[12] Any good that does not have this property is indivisible in the physical sense, which includes any good with differentiated parts. As Baumol (1987: 793) points out, all manufactured capital goods are indivisible in their physical descriptors almost without exception.[13] (Baumol does not consider indivisibilities in either of the other two senses, input or output descriptors.) So the class of goods that are indivisible in the physical descriptor sense covers virtually all manufactured capital goods, including all plant and equipment, as well as all consumers' durables.

It follows immediately that there must be something wrong with any argument that makes physical indivisibility play a key part in explaining why scale economies exist for some types and sizes of capital goods and not for others. Koopmans provides an example of such an argument when he seeks to buttress his logical argument concerning indivisibilities (given earlier in this paper) with the following empirical observation:

> "I have not found one example of increasing returns to scale in which there is not *some indivisible commodity* in the surrounding circumstances. The oft-quoted case of a pipeline whose diameter is a continuous variable ...[requires] one entire pipeline of the requisite length ... to render the service. Half the length of line does not carry half the flow of oil...." (152 fn. 3, italics added.)

As has just been observed, however, virtually all produced capital goods are indivisible in the physical sense. So all Koopmans seems to be saying is that *in all cases of increasing returns to scale related to capital goods, there are capital goods in the surrounding circumstances.* True, but not very helpful.[14]

Is there any other sense in which a capital good, or a consumers' durable, might be called divisible? For a clue, look at the standard discussions of scale in which it is often argued that, in the absence of an indivisibility, output can be altered merely by "scaling up or scaling down", any given process, thus encountering constant returns. For example, Koopmans speaks of lowering output by "subdividing the process". Except where this means closing some of a set of identical production units ("de-replicating"), it must mean reconfiguring the capital used to produce at a smaller volume of output. First, consider the physical descriptor. We have already observed that reducing all of the physical dimensions of a piece of equipment in the proportion λ, this will not reduce all the inputs required to make the smaller machine in the same proportion (by virtue of three dimensional relations), nor will it scale down its strength or capacity to deliver services in the same proportion.[15] Second, consider the input descriptor. If all of the variables in the input descriptor are reduced in proportion, neither the physical descriptor nor the elements in the output descriptor will be altered in the same proportion. There are probably no physical capital goods that are divisible in the sense that they can be "scalded down" so that their physical, input, and output descriptors are all altered in the same proportion $0< \boldsymbol{l} <1$. This is not a matter of logic but a characteristic of the material, three-dimensional physical world that we inhabit.

In our world, manufactured capital goods having differentiated parts are not divisible in any of the senses defined above. They cannot be physically subdivided and have the parts perform as did the complete good; neither their physical dimensions nor the list of inputs that go into making them can be scaled up or down without altering their various performance characteristics significantly.

2. Altering production with indivisible capital goods

Let some capital good be indivisible in the senses that its physical descriptor is indivisible (differentiated parts) and that there is some minimum size below which it cannot be made. When the machine is operated at capacity, let it deliver services at the rate of s^* and let there be an optimum combination of all the other factor service inputs that will minimize the unit costs of production. There are two ways in which the indivisibility can manifest itself.

First, the output flow of capital services may be subject to an equality constraint so that the service delivered is fixed at $s = s^*$. In this case, there is an indivisibility in the output descriptor of the capital good. This would be the case of a machine that had to be operated full time at a constant speed, whether needed or not. In that case, reductions in output would require that the optimum combination of inputs be departed from as other inputs were reduced while s remained fixed at s^*. The law of diminishing returns (or variable proportions) then dictates that unit costs will rise as output is reduced.

Second, the capital good's output of services may be subject to an inequality constraint so that the actual service flow delivered is $s \leq s^*$. This would be the case if the machine (or factory) could be shut down when it was not needed. In this case, there is no indivisibility in the output descriptor of the physically indivisible piece of capital, only a constraint on the maximum value its output can take on. Now the optimum combination of inputs can be maintained whenever the level of output calls for an input of the capital goods services of no more than s^*. If it does call for more, either the output cannot be produced if factor proportions are rigid, or it can be produced at rising unit costs as the optimum portions are departed from when all other inputs are increased with s held constant at s^*. As it is with one machine, so it is with a whole factory. If it must be operated full

time, reductions in output will require departing from the optimum input combinations and so encountering rising unit costs. If it can be operated part time, optimum input combinations can be maintained up to capacity and costs will be constant over the range from zero to the plant's capacity. [16]

The fact that many "fixed factors" are subject to weak rather than strong inequality constraints explains why so many empirical short run cost curves are flat. Below full capital capacity, production is varied by varying all inputs in proportion, which is done in the case of capital by leaving some of it idle and holding employed input proportions constant (so that the law of diminishing returns does not apply). Only when the strong equality constraint applies to the "fixed" factor do "scale" effects occur, and only then will the short run cost curve display the falling portion so commonly seen in text book treatments of the firm (because the law of diminishing returns does then apply).

3. Set up costs

There are many once-for-all costs that create indivisibilities, such as the cost of developing a new product or a revised version of an old one, the cost of establishing a brand name, the cost of entering a new market, and so on. With many of today's consumer goods, production runs last no more than a few years and the cost of developing new products is a significant part of the total costs of producing that product. In these circumstances, firms face declining unit costs in terms of quantities and values of inputs as the sale of each product line is increased over that product line's lifetime.

An interesting historical case is in the reproduction of the written word. Until the introduction of the printing press, virtually all of the costs of reproducing a book were the variable cost of the scribe's time. After the printing press was introduced, the bulk of the costs were the fixed cost of type setting, and the marginal cost of another copy of a book or pamphlet was quite small. The early Protestants' appeal to the masses though the printed word would have been quite impossible without this cost structure (Dudley 1991: Chapter5).

This printing example illustrates the error in the commonly heard statement: "Since all costs are variable in the long run, long run must be characterized by constant returns to scale." Even when scale effects arise out of lumpy expenditures, these often need to be replicated over time, causing the firm to face falling unit costs of output at each and every point in time and over all relevant time intervals.

V. WHAT LIMITS SIZE?

It is a characteristic of most of the above examples that scale effects are embedded in the physical nature of the world in which we live. Many of these specific sources cause scale effects that are unbounded: the larger the fence, the lower the cost per unit of pasture enclosed; the larger the ship, the faster its hull can be propelled through the water, the larger the blast furnace, the lower are both the construction costs and the heat requirements per unit of ore smelted. If these were the only influences in operation, the size of each production unit would expand until only one unit existed, resulting in universal monopoly. But in practice there are many offsetting influences that limit size of production units.

A. Offsetting Scale Diseconomies

In some cases, physical relations limit size. It has already been noted that the structural strength of any three dimensional body diminishes as its dimensions are increased, *ceteris paribus.* For example, a small airplane can tolerate a hard landing that, scaled up proportionately, would destroy a 747. In many cases, turbulence arising from the motion of a body through a gas or a liquid increases more than in proportion to the increase in the dimensions of the body.

In other cases, complimentary technologies cannot support a larger size of the main technology.[17] For example, hand and animal driven bellows could only deliver an effective flow of air to quite small blast furnaces. In the Medieval period, water wheel driven bellows allowed air to be injected with more force so that furnaces could be increased in size, reaping the benefits of lower construction costs and lower heat requirements per unit of output. In the 19th century, steam pumps again increased the feasible size of the furnaces, reaping further scale effects. Later, the technique of preheating the air before injecting it allowed an even larger area of molten metal to be effectively bathed in oxygen, further increasing the efficient size of blast furnaces. In each of these cases, the scale economies in materials used and heat loss were balanced by diminishing efficiency of the air delivery system as the size of furnaces increased beyond some critical size. Successive innovations in the air delivery system allowed further exploitation of the scale effects in construction and heat utilization, but each time only up to some critical point.

Thus, increasing returns to some characteristics of specific technologies are often balanced by decreasing returns to other characteristics of the main technology or in complementary sub-technologies. The optimal size of productive unit is then the one at which the economies of scale in some aspects of the technology just balance the diseconomies in other aspects. The smallest workable size is seldom the most efficient. As size increases, most characteristics encounter favorable scale effects. However, many characteristics encounter decreasing returns which eventually dominate so that further increases in capacity result in higher rising costs per unit of capacity delivered. If the unit cost of delivering a capital service is plotted on the *Y* axis and the capacity output of the capital good delivering that service on the *X* axis, the resulting cost curve is U-shaped. As one moves along it from left to right, the capital good is being reconfigured to deliver increasing amounts of the services per unit of time. The precise shape of the curve, and whether it reaches a minimum at relevant output capacities, depend on technical relations which cannot be discovered without a detailed knowledge of the engineering characteristics of the specific technology in question. *A priori* reasoning cannot tell us about the existence and range of such scale economies and diseconomies.

B. "Scale" Effects in Economic History

In economic history, falling unit costs of output are often observed to accompany technological changes. These effects are associated with the very long run reconfigurations that accompany technological change and so are not the scale effects that economists typically define under conditions of constant technology. Nonetheless, they produce important changes in unit costs when new technologies allow increases in the scale of operations.

The reason for these "historical" effects is foreshadowed in the previous discussion of blast furnaces. They arise because the scale effects are permanently embedded in the geometry and physical nature of the world in which we live *but our ability to exploit them is dependent on the*

existing state of technology. Here are some important examples. Every time agents learn how to build and operate larger ships, they reap the unit cost reductions implicit in the relations laid out above. New materials associated with the 20th century materials revolution have allowed many processes to be carried on at a larger scale, so reaping more of the favorable scale effects inherent in the basic process. Chandler (1990) points out that the combination of the newly developed railway and the telegraph in 19th century US vastly increased the size of the market. This occurred in many industries that had technologically determined scale economies that could not be fully exploited in small local markets. Cigarettes, light machinery, electrical equipment, metal manufacturing, oil refining, rubber, paper, glass, aluminum, and steel, are just some of the many industries that grew in the last half of the 19th century to become technologically driven oligopolies, serving the US national market.

> "The steady high volume of throughput, needed to achieve and maintain potential economies of scale and scope could rarely be attained as long as the flow of goods depended on [local markets]…. The railroad provided the technology not only to move an unprecedented volume of goods at an unprecedented speed, but to do so on a precise schedule, that is, a schedule stated not in terms of weeks or months but of days and even hours…and the telegraph became a critical instrument in assuring safe, rapid, and efficient movement of trains" (Chandler 1990: 53-4)

Notice that to evaluate all of these cases, one needs empirical knowledge of technical relations that exist in the real world. No attempt is made to deduce anything about historical scale effects from the mere definition of some relation such as a production function.

C. Constant Returns to Scale

In the literature on scale effects, it is common to argue that, because of the possibility of replication, one should never expect to see diminishing returns to scale. This argument is an example of the common attempt in capital theory to deduce empirical behavior from *a priori* arguments. The correct statement is "where replication is possible, one should never expect to see diminishing returns to scale." This raises the empirical question "Under what circumstances is replication possible?" Writing in the *New Palgrave*, Eatwell answers this question without recourse to empirical knowledge. He writes: "…barring indivisibilities, there can be no barrier to replication… In other words, there can be no such things as decreasing returns *to scale*" (1987: 166, Italics in the original). In a similar vein Silvestre (1987:87) states "…an exact clone of the production process that exhaustively lists all factors of production should give exactly the same output. The failure to double the output suggests the presence of an extra input, not listed among the arguments of the production function, that cannot be duplicated."

This common style of argument against decreasing returns to scale implicitly assumes that replication is always possible in any relevant production process, as long as there are no input indivisibilities. But this is an issue that cannot be settled by a *priori* reasoning based on an abstract model of the production process—a model that omits many of the conditions that affect output other than factor inputs as usually understood.[18] To go further requires an appeal to empirical evidence. Here we can only illustrate the empirical possibilities. On the one hand, to produce more razor blades, a new plant identical to existing ones can be set up in a green field and

managed independently. This should yield constant returns to scale and to outlay. On the other hand, if more output is required at a point in space, it may not be possible to replicate.

Here is one spatial example that was important in economic history. As British coal mines went deeper and deeper in the 18th century, they went below the level of the water table and flooding became a problem. At first, horses were used—both to turn the capstans that drove pumps and to haul the water, bucket by bucket. This, along with some other techniques, sufficed for a while. But as more and more water had to be removed from the ever-deepening shafts, the number of horses used at a specific mine head increased. Although there is no limit to the amount of energy that can be obtained from horses if there is room for them to operate, there were physical limits to the number that could be applied to any one pithead. Long before that absolute limit was reached, costs per horsepower increased due to such problems as non-linear increases in difficulties of coordinating the operation of horses both when at work and when changing shifts. There was no input indivisibility only spatial problems in applying inputs to a point in space and coordination problems in dealing with larger quantities of inputs.[19] There was no point replicating by building another pit; more power was needed at each pithead. (The steam engine solved the problem by delivering far more power at a single point than horses could.)[20]

This example is only an illustration of the general proposition that one cannot deduce anything about scale effects if one knows nothing about the physical conditions under which production is occurring. For example, replication is possible in some circumstances, where no worse than constant returns is predicted, and not in others, where the possibility of decreasing returns arises whenever the physical conditions dictate decreasing returns to reconfiguration.

D. Replication and Scale effects

In all cases in which the cost curve relating the value of reconfigured capital to the unit cost of the capital good's service is U-shaped, the capital good that exists is expected to operate at capacity (unless the market can be served with one unit of the good operating at less than capacity). In other words, all relevant scale economies will be exhausted. Output will then be varied by replication. This was true, for example, when ore was smelted first in small furnaces with animal operated bellows, then in somewhat larger furnaces operated with waterwheel driven bellows, then in much large furnaces using steam power and later, and even later in larger ones using with preheated air. The shift to each of these methods of production produced scale economies that reduced the unit cost of smelting ore. But once the new technologies were installed, production was varied by replication at constant cost. For example, the introduction of steam driven bellows reduced the number of furnaces and lowered costs but, as production increased over time, the industry's capacity was expanded by replication at constant costs.

This observation has important implications for attempts to measure scale effects empirically. Many such attempts fit long run costs curves designed to show how unit costs vary with output when there is sufficient time to adjust capital equipment as well as inputs that are variable in the short run. Such long run cost curves will decline only if firms have monopoly power and forgo operating one unit of the required capital equipment at capacity. In most cases, even firms with market power have many units of each type of capital, e.g., many blast furnaces, many airplanes, or many drill presses. So one will see constant costs of production due to replication even though

each machine tool has large scale economies (economies which will have been exhausted by building each at its optimum capacity).

IV. CONCLUSION

To understand the causes and consequences of scale effects, one needs to model a two-stage production process. In the second stage, which is the only stage modeled in standard production theory, some final output is altered by replicating identical units at a faster or a slower rate. In the first stage, a capital good is produced to provide a capital service to be used in the second stage. If more capital service input is required in the second stage of production, the capital good produced in the first stage may either be replicated (where this is physically possible) or reconfigured. Reconfiguration alters the physical, input, and output descriptors of the required capital good. However, varying either the machine's physical descriptor or its input descriptor by some scalar, λ, are irrelevant operations. When efficient reconfiguration occurs to alter the capital good's outputs descriptor by some scalar, λ, there are no *a priori* expectations as to how this will affect the physical descriptor, the input descriptors, and the unit cost of delivering the capital service. Our world's three dimensional structure and its physical and chemical laws dictate many scale effects, some of which cause increasing returns to reconfiguration while others cause decreasing returns. Thus, one cannot understand either the sources of scale effects or their consequences without a great deal more empirical knowledge than is assumed in any standard micro theory text.

Two charges can be levied at the textbook version of the neoclassical production function. The first charge is one of excessive abstraction. Most of the interesting real world forces that create scale effects are assumed away in the standard treatment, leaving scale effects to be explained by a force, indivisibilities, that covers one of the least interesting of the many sources of actual scale effects.

The second charge is one of scholasticism. The medieval scholastic philosophers asked many interesting questions, which they thought they could answer by reason alone, unaided by empirical observation. The typical micro textbook treatment of the production function is scholastic in this sense. It deals with important economic relations but tries to deduce propositions about the behavior of production processes with virtually no input of technological facts, and without even modeling the stage of production where most of the scale economies occur, i.e., the design of capital goods.

To the defense that micro production theory is only a tool box that can be used to relate to detailed empirical observations even though it does not itself incorporate such knowledge, one may reply that a tool box that obscures the most interesting observations of scale effects can hardly be considered a good tool box. In so far as it hones the intuitions of those students who struggle through its difficult logical exercises, it conditions them, first, to think that scale effects are unimportant; second, to analyze important problems in terms of one-stage rather than two-stage production processes; and, third, to believe that important conclusions about the world can be reached from reasoning that abstracts from virtually all of the characteristics that differentiate our world from other imaginable worlds.

END OF TEXT

BIBLIOGRAPHY

Adams, James (1991), *Flying Buttresses, Entropy and O-Rings: The world of an engineer*, (Cambridge: Harvard University Press).

Baumol, W. J. (1987) "Indivisibilities" in Eatwell, Milgate, and Newman, 1987, vol. 2: 793-95.

Becattini, G., (1987) "Internal Economies", in Eatwell, Milgate, and Newman 1987, vol. 2: 889-91.

Cardwell, Donald (1995), *The Norton History of Technology*, (New York: Norton).

Chandler, Alfred D. (1990), *Scale and Scope: The Dynamics of Industrial Capitalism*, (Cambridge, Mass: Harvard University Press).

Dudley Leonard M., (1991), *The Word and the Sword*, (Oxford: Blackwell).

Eaton, B. Curtis, Richard G. Lipsey (1997), "Increasing Returns, Indivisibility and All That" in, *On the Foundations of Monopolistic Competition and Economic Geography* (Cheltenham: Edward Elgar).

Eatwell, J., M. Milgate, and P. Newman (eds.) (1987) *The New Palgrave, a Dictionary of Economics*, (London: Macmillan).

Eatwell, J., (1987) "Returns to Scale" in Eatwell, Milgate, and Newman, 1987, vol. 4:165-6.

Hiscock, Eric C. (1965), *Cruising Under Sail*, (Oxford: Oxford University Press).

Koopmans, T.C. (1957) *Three Essays on the State of Economic Science*, (New York: McGraw Hill).

Landes, David S., (1969) *The Unbound Prometheus*, (Cambridge: Cambridge University Press).

Lipsey, Richard G., C. Bekar and K. Carlaw (1998), "General Purpose Technologies: What Requires Explanation", Chapter 2 in *General Purpose Technologies and Economic Growth*, Elhanan Helpman (ed.), (Cambridge: MIT Press).

Oi, W., (1987) "Fixed Factors" in Eatwell, Milgate, and Newman, 1987: vol. 2: 384-5.

Rosenberg, N. and L. K. Birdzell (1986), *How the West Grew Rich*, (New York: Basic Books).

Silvestre, J., "Economies and Diseconomies of Scale" in Eatwell, Milgate, and Newman 1987: vol. 2: 80-83.

Vassilakis, "Increasing Returns to Scale" in Eatwell, Milgate, and Newman, 1987: vol. 2: 761-65.

ENDNOTES

[1] Silvestre (1987) refers to the cost concept as "economies and diseconomies of scale" and the physical concept as "increasing or decreasing returns to scale."

[2] In this activity analysis treatment, a_1 can be thought of for purposes of this paper as the output and a_2, ..., a_n as inputs.

[3] Koopmans goes on to make clear that he regards such indivisibilities as common in the real world and so does not believe that constant returns are typical of most real-world production functions.

[4] Vasillakis (1988: 761) invokes indivisibilities in his explanation of how increasing returns may be related to the division of labour.

[5] When the firm builds a larger fenced pasture it is reconfiguring its capital because the specification of its capital good, in this case the length of the fence, is altered. Replication would mean creating a series of fields with identical dimensions.

[6] We have assumed that the sides must be linear, in which case minimization requires that the field be square.

[7] The first stage consists of two separate operations. First, a capital good is produced; then it is used to provide service flows, which are used in the production of the final good. These two operations are collapsed into one by defining the capital good in terms of its service output. Thus equation (3) is simultaneously the production function for the capital good, fenced pasture, and for the flow of capital services of fenced pasture per unit of time.

[8] If the producer of the fence has market power (as is true for most real-world producers of capital goods), it will price its product so as to capture the rents resulting from the scale economies and there will be no measured change in total factor productivity (TFP) as output rises. But if the fence is sold at marginal cost, then costs will be falling for the second stage producer and measured TFP will rise as output rises.

[9] No claim is made that these categories exhaust all possible sources of scale effects. For example, the law of large numbers is responsible for many scale effects associated with risk management, in such areas as physical and financial inventories.

[10] For example, if the output is a two dimensional pasture, the input is a one dimensional fence; if the output is a three dimensional storage room, the input is two dimensional walls.

[11] For a full discussion of these important economies of naval reconfiguration see Rosenberg and Birdzell (1986: 83-4).

[12] This last qualification tells us that, in our world, no good is fully divisible. Sooner or later, an indivisible unit is encountered. For example, two atoms of hydrogen and the one of oxygen will not do the same job operating independently as will one molecule of water. Nevertheless, over a wide range, water is divisible in the way that a lathe is not.

[13] There are, however, a few divisible manufactured inputs, such as fertilizers and fuel oil.

[14] Silvestre (1987: 81) also quotes his passage and with apparent approval.

[15] In spite of science fiction writers' implicit belief to the contrary, if any piece of physical equipment, or animal, has all of its dimensions altered in proportion, its performance characteristics are altered drastically and not in the same proportion.

[16] Of course there may be other indivisibilities such as management services which will impart some scale effects as output is varied below full capacity, but these do not stem from the indivisibility of the capital good which is of concern to us.

[17] As emphasized by Lipsey Bekar and Carlaw (1998) most complex technologies have a fractal-like structure in that they are made up of many cooperating technologies each one of which is in turn made up of cooperating technologies and so on through layer after layer of complexity.

[18] Of course, if the list of possible missing inputs is defined as anything that might cause the neoclassical production function to display decreasing returns, such as climate, spatial conditions of production and varying physical ability to organize relevant factors to produce effective units of the required input, the proposition becomes tautological and hence uninteresting empirically.

[19] Some readers have pointed out that it is possible to "avoid" this problem by defining the input as an effective unit of horsepower delivered where it is needed. But this scholastic "solution" merely hides the problem by pushing it to the unmodelled first stage in which these units of horsepower inputs are produced. At that stage, difficulties of coordinating the activities of a larger and larger number of horses, and applying their effort to the required place cause the unit cost of an effective unit of horsepower to be rising giving decreasing returns to scale of the activity of pumping a volume of water out of a mine.

[20] For details of this example see Landes (1969: 96-7).