ELSEVIER

Structural Change and Economic Dynamics 6 (1995) 485–504

STRUCTURAL CHANGE AND ECONOMIC DYNAMICS

Why the interpretation of the Cobb–Douglas production function must be radically changed

Paolo Sylos Labini

Università di Roma 'La Sapienza', Via Nomentana 41, 00161, Roma, Italy

Abstract

Despite several stringent criticisms, the Cobb–Douglas function has not been abandoned and, recently, a number of growth models have been presented that make use of it. This function, it has been said, can be retained 'at a first approximation'; in any case, empirically it has had 'an apparent success'. I maintain that the said success is an illusion and that, as a rule, the two exponents, α and β, have values incompatible with the distributive shares that the Cobb–Douglas claims to explain. I also maintain that we have to allow for technological changes and, therefore, assume a dynamic – not a static – substitution. The two exponents can then be unified in one exponent, γ, which is irrelevant in the interpretation of income distribution, but can have some interest in growth theory. However, if it is true that growth is, by its very nature, an uneven process, then we have to adopt a multi-sector dynamic approach.

Keywords: The Cobb–Douglas production function
JEL classification: E25; O40

1. The Cobb–Douglas production function: requisites for the neoclassical interpretation

The production function known as the Cobb–Douglas production function – $Y = AL^{\alpha}K^{\beta}$ – can be interpreted in terms of marginalist theory of distribution if it complies with a certain number of requisites. The list of these requisites is impressive, as is their complete divorce from the characteristics of contemporary economic reality. The list is as follows.

(I) It is necessary to assume conditions of atomistic competition in all markets, so that the prices are parameters.

Elsevier Science Publishers B.V.
SSDI 0954-349X(95)00025-9

(II) The returns must be constant: only with constant returns is the sum of the two exponents, α and β, equal to one and can Euler's theorem be applied. However, since this requisite conflicts with the previous one – constant returns and widespread atomistic competition are not compatible – it is assumed that all firms are at the point of minimum cost, where, for an instant, constant returns are achieved: in the hypothetical movement logically preceding the point of minimum cost, returns are still increasing (and costs decreasing) and, in the logically successive movement, returns are beginning to decrease (and costs increase). If one considers shifts of the production function – the isoquants moving gradually higher – one assumes that production increases not as a result of the expansion of already-operating firms, but merely as a result of the entry of new firms, all of which are tiny and all with the same cost curve (U-shaped).

(III) One assumes that the various capital goods are malleable and adaptable at will, so that the aggregate capital can be treated as though it were a single good (capital like jelly).

(IV) On the basis of the preceding assumptions, one uses the notion of the partial derivative of production; i.e. the notion of marginal productivity of a factor which, on the one hand, allows the definition of the notion of elasticity of substitution – often assumed to be constant – and, on the other hand, permits us to establish the equivalence between the values of the two exponents and those of the two macro distributive shares.

(V) To render dynamic the aggregate production function, other requisites were added, which were not less disputable than the preceding requisites: it is assumed that technical and organizational progress is 'neutral'; i.e. it pushes the isoquant to the right, leaving its form unchanged. In other words, the efficiencies of capital and of labour increase but their reciprocal substitutability does not change with respect to variations in the relative prices of the same factors.

(VI) It is assumed, finally, that the value of aggregate capital can be measured independently of its return.

2. Theoretical and empirical criticisms of the neoclassical interpretation

The criticisms of the aggregate production function have related to the absence of realism or to the logical indefensibility of the above requisites. I refer, in particular, to the criticisms put forth by Robinson (1953–54), Pasinetti (1959) and Sraffa (1960). I also refer to the excellent critical survey of the theory of capital worked out by Harcourt (1972). It is true that these criticisms go back many years, but they have never proved wrong: they have simply been ignored. A few defenders of the aggregate production function, however, have recognized the substantial validity of the said criticisms, but some (such as Ferguson, 1972) have claimed that, 'at first approximation', the notion of production function could be retained: an untenable position, as Roncaglia has correctly observed, given that one (the sixth) of the requisites

undermines the logical base of the construction (the work that includes this quotation (Roncaglia, 1978) also has a concise critical review of the marginalist theory of capital). One of the defenders (Fisher, 1971), after having honestly recognized that the notion of aggregate capital must be abandoned, still observed that "there is a genuine empirical phenomenon which needs explanation", adding "the question of what is behind the apparent success [of the production function] in the explanation of the distributive shares is not banal".

3. Explanation of the distributive shares according to the Cobb–Douglas function

The 'apparent success' seems to consist of the fact that, in empirical testing, the sum of the two exponents, α and β, almost always seems close to unity, as required by traditional marginalist theory, and the pair of values – for example, 0.7 and 0.3; 0.8 and 0.2 – seems to be compatible with the values of the two main distributive shares imputable to labour and to capital. The conviction that these were, as a rule, the results of empirical testing was so generalized that when, a few years ago, a brilliant young economist named Paul Romer found that the "exponent relating to labour can be substantially inferior to its share in income, possibly of the order of 0.1 or 0.2", he spoke of a "suggestive puzzle" (Romer, 1987, pp. 166–167). He does not appear to have even slightly suspected that the puzzle depends simply on the fact that the theoretical construction that lies behind those empirical tests does not work. In fact, this is precisely the case, as should already have appeared in the criticisms formulated on the theoretical and on the empirical levels, by quite a few economists (Mendershausen, 1939; Phelps Brown, 1957; Robinson, 1953–54; Pasinetti, 1959; Sraffa, 1960; Simon, 1979). It is worth recalling these criticisms, since an increasing number of young and talented economists seemingly do not know them, or do not take them seriously, and continue to work out variants of the aggregate production function, and include, in addition to technical progress, other phenomena, for example, human capital. In this paper I limit myself to recalling briefly three criticisms, integrating them into a point that I maintain – correctly, I hope – to be essential.

Let us start with the 'apparent success' noted by Fisher with respect to empirical testing. In reality, this 'success' of empirical testing does not exist: in truth, one must speak of failure. More precisely, Cobb–Douglas empirical tests belong to two categories: (I) those founded on the time series, and (II) those derived from cross-sections that refer (1) to industries of a particular country, or (2) to various countries with significantly different prices of production factors. It seems that the sum of the two exponents is often close to unity only in the case of (II, 1)-type testing (cross-sections); we shall shortly see why this has nothing to do with the marginalist theory of distribution. I know of only one case (Arrow et al., 1961) that comes under the category (II, 2) (cross-sections between countries), and it is a case in which the empirical test under discussion makes no sense, as we shall also see. *The sum of the two exponents is, instead, rarely close to unity in the tests based on the time series.* There are, it is true, many tests of this type in which that sum is equal

to unity, but this occurs simply on the basis of an ad hoc assumption. In other words, the validity of the marginalist theory of distribution is taken for granted. Therefore, the sum of the marginal productivities of the two factors multiplied by the respective quantities exactly exhausts total production.

"Today" – Solow wrote in the opening sentence of his 1957 paper – "it takes something more than the usual 'willing suspension of disbelief' to talk seriously of the aggregate production function" or, more bluntly, as Ferguson put it (1969, p. 169), to do so is necessarily a "statement of faith". In truth, when $\alpha + \beta = 1$ is not taken as an assumption, then the sum of the econometric estimates of these two exponents often gives results that, compared with the expectations created by traditional theory, appear completely absurd. Worse is to come: even if one imposes the said constraint, the results are often ridiculous, as is the case when one exponent turns out to be greater than unity, so that the other exponent is negative. It is worth reflecting on Table 1, which collects estimates that relate to four sectors in seven countries: among the 17 unconstrained estimates, only in four (i.e. 1, 6, 8 and 9 in the first group) do the values of the exponents seem reasonable and is their sum not too far from unity. Among the 14 constrained estimates (third and fourth groups), only half present reasonable values for the two exponents. To the estimates in this table should be added those to which Romer alludes without specifying them. Faced with this picture, it is difficult for Solow's agnostic not to become a declared unbeliever!

Herewith, the procedure elaborated by Solow and then adopted by various other economists.

One excludes increasing returns to scale and assumes that $\alpha + \beta = 1$, where $\alpha = \partial Y/\partial L \cdot L/Y$ and $\beta = \partial Y/\partial K \cdot K/Y$ and, following marginalist theory, that factor prices are equal to the marginal products of the two factors. Given the increases in output and in capital, and given β, it is calculated, year by year, how much of the output increase is due to capital increase, and how much to a trend factor, which indirectly can also explain the improvement in human capital. Solow estimates β year by year, on the basis of different sources and assuming that $\alpha + \beta = 1$ calculates by difference the increase of output imputable to the trend factor – to 'technical progress'. Guarini and Tassinari (1990) follow a similar criterion, but estimate β's value on the basis of the marginalist relationship and calculate α by the difference. According to Solow, in the non-agricultural sector in the USA in the period 1909–1949, the average annual output increase, equal to 1.8%, can be imputed by 87% to technical progress and by 13% to the increase in capital per employed person; according to Guarini and Tassinari, in the Italian manufacturing sector in the period 1976–1983, the corresponding figures are 2.5%, 83% and 17%.

4. Static and dynamic substitution

The Cobb–Douglas tests that I have recalled are those based on time series: they are the most significant ones, since we find in this category the original estimates of Douglas, as well as those of Solow, and here we find the recent models that include

human capital and other quantities that enter into the growth process. As I noted, the sum of the two exponents in cross-section testing is often close to unity (cf. Douglas, 1948) but, in this case, the explanation has nothing to do with the marginalist theory of distribution or with the relative cohorts of requisites. As Phelps Brown (1957, p. 553) correctly observed, the explanation is almost obvious: that the sum of α and β is often near unity "is the outcome, not of an analytical relation really governing the variations of labour, capital and output, but simply of the fact that between one industry and another these variates change for the most part in the same proportions as one another". Phelps Brown also put forward – or, rather, reproposed, in Mendershausen's footsteps (1938) – a critique of the traditional interpretation of empirical tests based on time series; a critique similar to that elaborated here and that I consider statistically correct but theoretically incomplete. It is statistically correct, since it is true that "each exponent simply explains a relationship between two different rates of growth" (Phelps Brown, 1957, p. 550). However, in my view, it is incomplete, because it does not consider the problem of the forces that condition the evolution of the different quantities: the contribution that I intend to make with this paper deals with precisely this problem.[1]

I maintain that the traditional interpretation of the Cobb–Douglas production function, which refers to the marginalist theory of income distribution, should be abandoned, and another – completely different – interpretation should be adopted. The point is that traditional theory given technology and is founded on the concept of substitution, whereas it is necessary to consider *dynamic substitution*. The first is defined with reference to a given level of production that can be obtained with different combinations of labour and capital; combinations that express the substitutability of the two factors, depending on the hypothetical variations of their relative prices, given the technology, already known and available. If the ratio increases between the prices of the two factors – labour and capital – one moves, under certain very restrictive conditions, from a less capital-intensive technique to a more capital-intensive technique. It is reasonable to hypothesize that no new technique has yet been discovered for the new price ratio, higher than the preceding one; a hypothesis that is anything but arbitrary, since at any point the available techniques for producing a given good are relatively few and, if examined carefully, are not wholly interchangeable but imply the production of different baskets of goods, or of similar but not identical goods – the differentiation of the goods characterizes ever more the economies of our times. When the ratio between the prices of labour and capital, therefore, reaches a level never previously achieved and there is no technology available that permits labour saving – labour having become more costly than capital – one then assumes that a new and more suitable technology will be developed. Generally, the passage from one type of technology to another is stimulated by the trend of output to increase, so that the hypothesis of an increase in

[1] It is fitting to point out that Douglas was so impressed by Mendershausen's critique as to write (1967, pp. 178): "Ragnar Frisch [in a work that I have been unable to find] and Horst Mendershausen, as I remember, said that our study should be thrown in the waste basket and all future research on it discontinued".

Table 1
Econometric estimates for four sectors in seven countries.

Equation	Country	Period	Sector	Alpha	Beta	$\alpha + \beta$	$r\%$	*Authors*
A	1 Italy	1922–1939	Non Agr. Firms	0.77	0.30	1.07		EEC
Without	2 Italy	1970–1984	Manu. Ind.	0.11	1.35	1.46		Guarini/Tassinari
constraint	3 Canada	1870–1938	Industry	1.07	0.18	1.25		EEC
on trend	4 Norway	1900–1955	Economy	0.92	0.91	1.83		EEC
	5 UK	1870–1912						
		1924–1938	Economy	1.73	0.18	1.91		EEC
	6 USA	1899–1922	Manu. Ind.	0.81	0.23	1.04		Douglas
	7 USA	1909–1949	Non Agr. Firms	1.85	0.69	2.54		EEC
	8 Australia*	1907–1924	Manu. Ind.	0.84	0.23	1.07		Douglas
	9 Australia**	1902–1927	Manu. Ind.	0.78	0.20	0.98		Douglas
	10 New Zealand	1915–1935	Manu. Ind.	0.42	0.49	0.91		Brown
B	1 Italy	1922–1939	Non Agr. Firms	0.91	0.48	1.39	0.85	EEC
Without	2 Italy	1970–1984	Manu. Ind.	1.33	0.05	1.38	3.00	Palazzi
constraint	3 Canada	1870–1938	Industry	0.75	0.09	0.84	1.20	EEC
and with	4 Norway	1900–1955	Economy	0.30	−0.39	−0.09	3.50	EEC
trend	5 UK	1870–1912						
		1924–1938	Economy	5.03	−0.74	4.29	1.30	EEC
	6 USA	1899–1922	Manu. Ind.	0.91	−0.53	0.38	4.70	Palazzi
	7 USA	1909–1949	Non Agr. Firms	1.51	0.06	1.37	1.50	EEC

C	1 Italy	1922–1939	Non Agr. Firms	0.67	0.33	1.00		EEC
With the	2 Italy	1970–1984	Manu. Ind.	−0.30	1.30	1.00		Palazzi
constraint	3 Canada	1870–1938	Industry	0.62	0.38	1.00		EEC
$\alpha + \beta = 1$	4 Norway	1900–1955	Economy	−0.29	1.29	1.00		EEC
and without	5 UK	1870–1912						
trend		1924–1938	Economy	−0.05	1.05	1.00		EEC
	6 USA	1899–1922	Manu. Ind.	0.75	0.25	1.00		Douglas
	7 USA	1909–1949	Non Agr. Firms	0.54	0.46	1.00		EEC
D	1 Italy	1922–1939	Non Agr. Firms	0.72	0.28	1.00	0.13	EEC
With the	2 Italy	1970–1984	Manu. Ind.	1.012	−0.012	1.00	3.00	Palazzi
constraint	3 Canada	1870–1938	Industry	0.90	0.10	1.00	0.90	EEC
$\alpha + \beta = 1$	4 Norway	1900–1955	Economy	0.84	0.16	1.00	1.80	EEC
and with	5 UK	1870–1912						
trend		1924–1938	Economy	−0.35	1.35	1.00	−0.30	EEC
	6 USA	1899–1922	Manu. Ind.	0.84	0.16	1.00	4.70	Douglas
	7 USA	1909–1949	Non Agr. Firms	1.12	−0.12	1.00	1.90	EEC

Notes: * Victoria; ** New South Wales. The last column gives the authors of the estimates, including the group of EEC experts (CEE, 1961) and my colleague Palazzi, who did not limit himself to calculations but also provided useful suggestions for the whole analysis.

Apart from the C-6 estimate, Douglas elaborated another estimate, imposing the constraint $\alpha + \beta = 1$ using deviations from the trend: for the two exponents he finds the values 0.84 and 0.16. Solow follows the peculiar procedure described in the text, in which he assumes that $\alpha + \beta = 1$ and adopts a predetermined value of β, which is on average equal to 0.32, so that $\alpha = 0.68$. Introducing the said constraint but estimating the equation on data of the whole period, and for non-agricultural firms, the EEC group found completely different values (D-7): $\alpha = 1.12$ and $\beta = -0.12$. Guarini and Tassinari, who follow a similar but not identical method to Solow, obtain the following average values, $\alpha = 0.76$ and $\beta = 0.24$, while in their estimates without constraint and without trend (A-2) the values are completely different: $\alpha = 0.11$ (Romer's puzzle!) and $\beta = 1.35$.

I invite the reader to compare the list presented here – 31 estimates – with the shorter list – 13 estimates – presented in Table 1 by Walters (1963, p. 26); the four estimates concerning the United States, Victoria, New South Wales and New Zealand (the most 'favourable' ones to the Cobb–Douglas function) appear in both lists, whereas the other nine estimates (with above mentioned constraint or with a trend) do not appear here.

An anonymous referee has pointed out to me a paper by Heathfield in which the author illustrates a number of additional failures, on the empirical plane, of the Cobb–Douglas production function (Patterson and Schott, 1979, pp. 226–245).

the ratio between the prices of the two factors is naturally linked to the hypothesis of an increase in output; though the two hypotheses do not necessarily go together. With this concept, technology can no longer be assumed as 'given' and technological change comes to depend on a variation in the ratio of relative prices and (generally) on an increasing output trend. This process requires time. In this case, there is also a problem of factor substitution, but it is a dynamic, not a static, phenomenon. In my opinion, this kind of factor substitution is actually much more important than the static substitution.

Two caveats should be kept in mind. First, for the price of workers' services, we can take – as is done generally – an index of wages and salaries. For the price of 'capital', we should not take the interest rate alone – for the firms, the interest rate is the price of monetary capital necessary to acquire all production means – but an index that combines the interest rate and the prices of capital goods; in a first approximation, however, we can limit outselves to considering only an index of the prices of capital goods (Sylos Labini, 1993, pp. 33–34). Secondly, if the price of labour increases compared with the price of capital goods, then labour is gradually saved, but this does not necessarily imply the dismissal of a certain number of workers: the 'saving' remains a potential if output expands.

Static substitution is seen through the isoquants, which imply the concepts of marginal rates of (static) substitution, marginal productivity of factors and elasticity of substitution. Dynamic substitutability does not imply these notions and is based on a completely different conception. Traditional theory has seriously exaggerated the importance of static substitutability, to the detriment of both the complementarity between factors and dynamic substitutability that – as we whall see – takes place mostly between capital and labour.

5. Productivity equation

Once we admit that the Cobb–Douglas function has nothing to do with the distribution of income and that the introduction of an exogenous rising trend cannot contribute to understanding the dynamics of modern economies, we have to ask: what is the meaning, if any, of that function? In posing the problem in terms that I deem correct, I was recently helped by an observation by Zaghini, which led me to reconsider one of the versions of the productivity equation that I had put forward some years ago (Sylos Labini, 1984, pp. 101–119) as follows:

$$\hat{\pi} = a + b\hat{Y} + cW/P^{\mathrm{ma}}_{-m} + dI_{-n} \tag{1}$$

where π is labour productivity, Y is income, W/P^{ma} is the ratio of wages to machinery prices, I is investments, and the circumflex accent indicates a time rate of change. The variations of income express what I call the Smith–Verdoorn effect (market expansion stimulates both short- and long-term productivity increases). One assumes that the productivity increase is motivated by an increase in investments, in turn motivated by the increased ratio of wages to the prices of machinery, with the warning that W/P^{ma} and I undergo lags – m and n, where $m > n$.

Let us start from a proposition on which economists of all persuasions can agree: productivity variations depend mainly on those of the ratio K/L, that, accepting Pasinetti's point of view, we define a degree of mechanization (Pasinetti, 1981, pp. 188 ff.). As a rule, however, there is no stable proportionality between the two variations. Why? One might think it depends on the type of investments that gradually increase capital stock. As a rule, all investments involve an increase in productive capacity and in productivity; the proportion between the two effects, however, varies according to the type of investment and the impulses that propel them. When the investments that increase productive capacity prevail, the degree of mechanization tends to grow more rapidly than productivity, whereas the opposite occurs when the investments of the second type prevail. It is reasonable to maintain that the investments of the first type are particularly stimulated by the increase in income (Y), and those of the second type by the increase in the relative cost of labour (W/P^{ma}). We can, therefore, state the relationship between labour productivity (Y/L) and the 'degree of mechanization' (K/L):

$$Y/L = \gamma K/L \tag{2}$$

where $\gamma = a + bY + cW/P^{ma}$. (I omit investments, since their variations are implicit in the variations in the degree of mechanization.) To establish a parallel with the neoclassical production function, we can formulate the relationship as:

$$Y/L = (K/L)^{\gamma} \tag{3}$$

which becomes $\log Y/L = \gamma \log K/L$. This relationship corresponds formally to the Cobb–Douglas production function when the constraint is posed that $\alpha + \beta = 1$ and one assumes that $A = 1$.

Two observations can be made here. First, the productivity increase discussed above is that deriving from endogenous technical progress, directly propelled by impulses from the economic system occurring almost uninterrupted; exogenous technical progress is not considered, since it is fostered by innovations that may be of great importance from the scientific and economic standpoints, but which are discontinuous. Secondly, in the neoclassical version, the equality $Y/L = (K/L)^{\beta}$ presupposes that $\alpha + \beta = 1$. This does not apply to the concept adopted here, which is completely different from the neoclassical version; thus, γ can be greater than unity and α does not even appear.

Exponent γ varies over time according to the variations of the two impulses noted above. This is confirmed by the econometric estimates that relate to the manufacturing sector of the USA in the period 1899–1920, which largely coincide with those considered by Douglas (1934), and of Italy in the period considered by Guarini and Tassinari (1990). Herewith, the estimates.[2]

[2] I have referred to Eq. (3) and calculated γ as the ratio $\log Y/L/\log K/L$; then I estimated γ as a function of Y and W/P^{ma}_{-m}. Data sources: USA: Y and K: Douglas (1934) and Kendrick (1961); prices (deflators) of investment goods: Kendrick, Table A-11b. Italy: Y and K: 1970–1984 Guarini and Tassinari (1990, p. 161); 1985–1991 data of K kindly supplied by the Istituto Nazionale di Statistica; prices of investment goods: Banca d'Italia, Relazione annuale – Appendice (various years). The figures below the coefficients of explanatory variables indicate t statistics. All R^2 of econometric estimates are 'corrected'.

USA (1899–1922):

$$\gamma = \underset{}{0.033} - \underset{11.38}{0.003Y} + \underset{4.50}{0.112W/P^{ma}_{-m}}, \quad R^2 = 0.917, DW = 1.468. \tag{4}$$

Italy (1970–1991):

$$\gamma = \underset{}{0.607} - \underset{9.40}{0.206Y} - \underset{3.25}{0.225W/P^{ma}_{-m}}, \quad R^2 = 0.947, DW = 1.862. \tag{5}$$

From the econometric standpoint, in the USA, two years is the most convenient lag for the second explanatory variable; in Italy, it is three years. In the USA, the values of γ are between 1.01 and 0.90, while, in Italy, they oscillate between 0.84 and 0.96 (see Figs. 2 and 3 in Section 8). According to the results of Guarini and Tassinari, the value of β is more than 135%, but there is a negative constant in that estimate, whereas α equals 11% and has almost nil significance ($t = 0.209$). The γ values are close to unity, so comply with 'Romer's puzzle', since β is the neoclassical counterpart of γ and thus α is close to zero.

It is worth noting that, in these equations, the signs of the explanatory variables are different: in (4), Y is negative and W/P^{ma}_{-m} is positive, whereas, in (5), the opposite is the case. However, in the productivity equations I published some years ago, the variables are the same but the signs are both positive. The reason is that the variable explained is different, as are the links between the variables; this will be reconsidered in Section 8.

6. Exponent γ and dynamic substitution

It is worth reflecting particularly on the second of the two variables that influence γ, i.e. the ratio of wages to machinery prices.

If one accepts that, in modern industrial capitalism, the ratio of wages to machinery prices systematically tends to increase, as a result of the upward pressure exercised by wages or, as occurred in the last century, as a result of their relative downward rigidity compared with machinery prices (as well as those of most other goods), it follows that technical progress tends prevalently to save labour. In fact, faced with increasing labour costs, managers tend to replace a number of workers with machines. They decide to do so when the value of the machinery is less, even slightly, than labour costs, suitably discounted, over the course of a given number of years, for the workers who can be 'saved', in real or potential terms. Also, since this is an almost uninterrupted process in modern capitalism, the producers of capital goods tend to produce systematically labour-saving machinery: only the degree of the saving effect varies. Thus, as a result, every year, the value of the machines that replace workers can only be a multiple of the annual costs of the wages of the workers 'saved'. Given that the durable capital goods that replace labour, in real or potential terms, constitute most of the durable capital goods produced annually, it follows that the fixed capital over time tends to

increase more rapidly than does labour:

$$K(1 + r_K)^n > L(1 + r_L)^n,$$

where r_K and r_L are the average rates of variation of K and L, and n is the number of years. Production also increases more rapidly than does employment, since the ratio Y/L – which represents average labour productivity – increases spectacularly with respect to the labour-saving trend:

$$Y(1 + r_Y)^n > L(1 + r_L)^n.$$

The two ratios

$$\frac{Y(1 + r_Y)^n}{L(1 + r_L)^n} \quad \text{and} \quad \frac{K(1 + r_K)^n}{L(1 + r_L)^n}$$

increase at the same speed if $r_Y = r_K$; if, instead, the two rates are different, then the equality between the two ratios can occur, by giving the second ratio an exponent different from 1: $\gamma = 1$ if $r_Y = r_K$, $\gamma > 1$ of $r_Y > r_K$ and $\gamma < 1$ if $r_Y < r_K$. At this point, a warning must be given. In the above reasoning, the question of the measure of aggregate capital reappears. However, the new interpretation no longer has the requisites and constraints that led the neoclassical analysis into a dead-end. The new interpretation does not aim to explain distributive shares, having taken aggregate capital as given, but tends to identify the role of labour-saving technological progress in the process of accumulation and growth. If one adopts this interpretation, then there is no objection to the use of aggregate capital measured in constant prices. In this problem – as in many other problems in economics – logical rigour must not apply to the exact measurement of the quantities considered, but to the sequence of interactions between the different quantities. Only if the adoption of one measure rather than another can upset the interpretation of that sequence does the measure's absolute precision become vitally important. However, this does not seem to be the case in this problem, since I do not pose it in terms of the marginalist theory of distribution.

Thus, the values of γ have nothing to do with the elasticity of substitution, marginal productivity of factors, or distributive shares: they are values that depend on the rate of increase of the three quantities considered. Therefore, when Y/L increases in almost the same proportion as K/L, γ nears unity: this is the rule and not the exception. In the logical sequence examined here, the increase in labour productivity depends on the increase in capital per worker; an increase in turn stimulated by the increase in output and by an increase in the ratio of wages to machinery prices. In modern capitalism, such an increase is the rule, whereas a decrease is the exception. However, three points must be highlighted.

First, the impulse that comes from the increase in the ratio W/P^{ma} is the sum of a series of other impulses, among which that deriving from technical progress applied to the same production of machinery, which leads to a decrease in their prices compared with wages (cf. Pasinetti, 1959). However, in certain periods, the acceleration of the wage increase depends on the intensification of social conflicts; in these

periods – such as that of the restructuring of manufacturing industry in Italy in the early 1980s – managers introduce labour-saving machinery mainly as a reaction to these conflicts rather than to save increased labour costs.

Secondly, if it is true that technology changes almost without interruption in all economic activities, including those that produce durable capital goods, it is impossible to distinguish investments directed merely at increasing productive capacity and those aimed at increasing labour productivity. However, it is possible to distinguish the investments on the basis of the prevalence of the object sought and it is possible to claim that the relative weight of investments directed prevalently at increasing labour productivity tends to rise hand in hand with the rate of increase of the ratio W/P^{ma}.

Thirdly, the variations of income and of the ratio of wages to machinery prices have an important, but not exclusive, role in the increase of labour productivity. This is because technological progress takes place not only as a result of the two above-mentioned impulses, but also because of innovations that can be classed as autonomous. Furthermore, technological progress also originates new goods and saves not only labour but also 'capital': this can occur autonomously or as a result of stimuli from the economic system, such as those that originate from an expansion in output (which stimulates the introduction of new machinery) and from an increase, compared with wages, of machinery prices and the means used to produce them (the ratio W/P^{ma} decreases), or from an increase in prices of energy sources that constitute the complementary goods of this machinery. It remains true, in any case, that the pre-eminent phenomenon is that of labour-saving technological progress motivated by production increase and by the systematic increase in the ratio of wages to machinery prices.

In the relationship $Y/L = (K/L)^{\gamma}$ the variations of the ratio of fixed real capital to labour should be looked at the same time as an index of the changes in the technological and organizational characteristics of productive activity, and in the personal abilities of the workers and, thus, of their degree of education and skills. Naturally, an aggregate approach is wholly unsuited to analysing such aspects. The fundamental concept is that technological capital and human capital are strictly complementary factors, so that idea of isolating the specific contributions to production and productivity increases of the two factors is an attractive idea but, all in all, an illusion – as is the attempt to attribute to education and to other factors the 'residual factor' not explained by the hypothetical shifts of the production function. (This does not prevent one having to make every effort to evaluate the role and the suitability, even by means of international comparisons, of public or private resources destined for education, professional training and research.)

7. The CES function – the 'residual' factor: a spurious problem

As already noted, apart from cross-section tests concerning the economy of sectors of a given country in a given year, it is possible to make cross-section tests of economies of sectors of different countries in a given year. The only example of this

type of test that I know of is by Arrow et al. (1961), who, in an important paper, propose the well-known constant elasticity of substitution (CES) function. To construct the relative isoquants for each industrial sector, the authors identify two points that correspond to the pair of quantities labour and capital – much 'labour' and little 'capital' in the country where labour is relatively abundant and cheap and capital is scarce, and little 'labour' and much 'capital' in the other country. The two countries are Japan and the USA in the period 1953–1954, when the ratio of wage rates was about 1/8 – presumably, the ratio of average individual incomes was not dissimilar. Today, if we are to trust World Bank data, the ratio is more than 9/8: the Japanese economy has really come far! Despite the great diversity in the dimensions of the markets regarding the various industries, the authors maintain that they can mark both points on the same isoquant, because of two assumptions – constant returns and neutral technological progress – that would allow the isoquants' curves to be considered perfectly parallel. Now, to suppose that the combination of the two factors depends solely on the ratio of relative prices, excluding any influence on the extent of the two markets, means entering a fantasy world that has no relation to reality, since we know from Adam Smith onwards that the increase in the extent of the market can only bring about changes and improvements in technological and organizational methods. Therefore, the line linking the Japanese and US points cannot be interpreted as an isoquant: it does not even represent a possible evolutionary trajectory, because, when factor prices in Japan neared those in the US, technology and organizational operations became so different as to provoke a considerable shift from the US point of 1953–1954. These considerations imply that the isoquant that links the point of the country with lower wages and higher labour intensity with the respective point of the other country – having placed capital on the abscissa and labour on the ordinates – is not a reversible curve, as a static curve must necessarily be; it is irreversible. Only by hypothesizing a cataclysm could one imagine a return from the lowest to the highest point on the curve, but a catacylsm is a possibility that cannot be envisaged in a theoretical model.

It is worth pointing out that, conceptually, the CES function represents a generalization of the Cobb–Douglas function: this assumes an elasticity of substitution of the two factors that is always equal to unity, whereas elasticity being equal to unity represents only a particular case of the CES function. For the interpretation of the distributive share, the Cobb-Douglas function also represents a particular case of the CES function, which is, therefore, not well suited to the marginalist interpretation of distribution, although it is not in contrast to it. The CES function was used for dynamic analysis, by means of a similar expedient to that used by Solow for the dynamic use of the Cobb–Douglas function; i.e. with the introduction of an exogenous time trend that explains the part of the increase of output held not to be explained by the increase in capital per employed person. That is what Solow did in his 1957 paper: the trend element would therefore explain the 'residual factor' – as has been defined subsequently.

In commenting on Table 1 above, I have emphasized the manipulations undertaken to arrive at reasonable estimates for the two exponents: on the basis of an act of faith in the marginalist theory of distribution, the constraint $\alpha + \beta = 1$ was introduced,

sometimes implicitly, and the value of β was even introduced exogenously. The result is that neither the values of α and β nor the so-called 'residual factor' can be considered theoretically significant.

It may help to note the remarks made by Neild in summarizing the debate that took place many years ago at a Paris conference on the economics of education (OECD, 1964, p. 304):

> When one applies this [Cobb–Douglas] function to ex post series, the quality of the adjustment is nowhere near excellent: the analysis of the evolution of the quantity of capital per labour unit shows, in fact, that in the long term output per worker increases in proportion – and sometimes more than in proportion – to the increased capital per worker. Therefore, if one tries to adapt a Cobb–Douglas function, one is obliged to attribute most of the movement observed not to the function as such but to a residual factor, which can account for four-fifths of the change observed, if not more.

If these considerations and those I propose in my analysis are correct, then one must conclude that the much-debated question of the 'residual factor' represents a relevant example of a spurious problem. The 'residual factor' was attributed to technological progress or to improved human capital. The great statistical weight of the 'residual factor' is given as evidence of the great importance of technological progress (or of the higher education level in the growth process). I maintain that the proportion is even greater than that indicated by the 'residual factor': not 70%, 80% or 85% of the productivity increase, but – if one conceives of the growth process as a unitarian process, as one must – 100%. Even productive operations that continue with unchanged technology – and, period by period, this phenomenon in some firms is undoubtedly real – are carried out by firms whose commercial and financial organisations cannot but change irreversibly over time. The expansion of the market necessarily leads to an increasing division of social labour, and this increase necessarily leads to technological or organisational improvements, or a combination of both.

My interpretation is consistent with the view taken by Jorgenson in a paper in the 1987 edition of *The Palgrave Dictionary*; an outlook that indicates the consciousness, in one of the most brilliant and creative defenders of traditional theory, that there is something very unsatisfactory in this theory. Jorgenson writes: 'An important area for future research is the implementation of dynamic models of technology. These models are based on substitution possibilities among outputs and inputs at different points of time' (Vol. 3, p. 1007). He shows that he has glimpsed the importance of the process that I have defined as dynamic substitution, but he does not appear to have realized that what he hopes for is not the continuation of an old line of research; it is, in fact, a break.

8. Exogenous and endogenous innovations

To conceive of technological progress as a phenomenon that can be represented by shifts of the isoquants implies excluding the consideration of dynamic substitution,

which involves changes in the production methods that modify the actual contents of both K and L. The impulses considered above are precisely those that set dynamic substitution in motion. Technological progress, however, does not depend solely on such impulses: there are also innovations that derive essentially from scientific progress, which can be, and is, pushed and accelerated by investments in research. From this point of view, even these innovations can be traced to economic impulses. Here, however, the link between innovations and economic impulses is not strong and, in any case, is less strong than that between the impulses considered above and innovations: for this reason, I think it is preferable to define the former as 'exogenous' and the others as 'endogenous'. From the scientific point of view, exogenous innovations are often more important than endogenous ones; in the very long term, this also applies from the economic point of view. However, for the continuity of the growth process, and at least in the short and medium terms, endogenous innovations are not less but are even more important than the other ones, so that it is not misleading to concentrate one's attention – as is done here – on endogenous innovations.

The analysis worked out here, which takes as terms of reference the variations of γ, leads one to negate any utility in the distinction – indissolubly tied to the notion of static substitution – between virtual movements along the production function and shifts of that function.

At this point, it is worth considering with greater attention the relationship between the variations of exponent γ and of productivity. As we have seen, these two variations are subject to the same impulses that make productivity vary, and yet the signs are different: in the equation for the US, income has a minus sign, whereas, in that for Italy, the ratio of wages to machinery prices is negative; vice versa, in the productivity equation all the explanatory variables were of a positive sign.

Starting from the relationship $\log (Y/L) = \gamma \log (K/L)$, γ can be seen as the ratio of $\log Y/L$ to $\log K/L$, or, graphically, as in Fig. 1.

It appears evident that γ increases when $\log Y/L$ tends to increase more rapidly than $\log K/L$, and decreases in the opposite case. γ decreases when both labour productivity (Y/L) and the degree of mechanization (K/L) increase, if the increase of the former is slower than the increase of the latter; indeed, in this case, the angle that determines γ turns downward. In general, when the variables that determine γ have different signs, its trend will increase or decrease according to whether the push coming from the variable with the positive sign is stronger than that from the one with the negative sign, and vice versa: not only the direction counts, but also the relative speed of the two variables. The fact is that investments allow growth of both productive capacity and productivity, but in different proportions. In the case of the US from 1899–1920, γ tends to decrease. One may assume that, in that period, the component that leads to an increase in productive capacity prevailed in investments, so that K/L increases more than Y/L, whereas, in Italy from 1970–1991, γ generally tends to increase. In this period, and especially from 1976–1980 and 1986–1989, the prevailing component was that leading to a productivity increase – let us recall that a radical restructuring of Italian industry took place during the central part of that period. The trend in the two countries

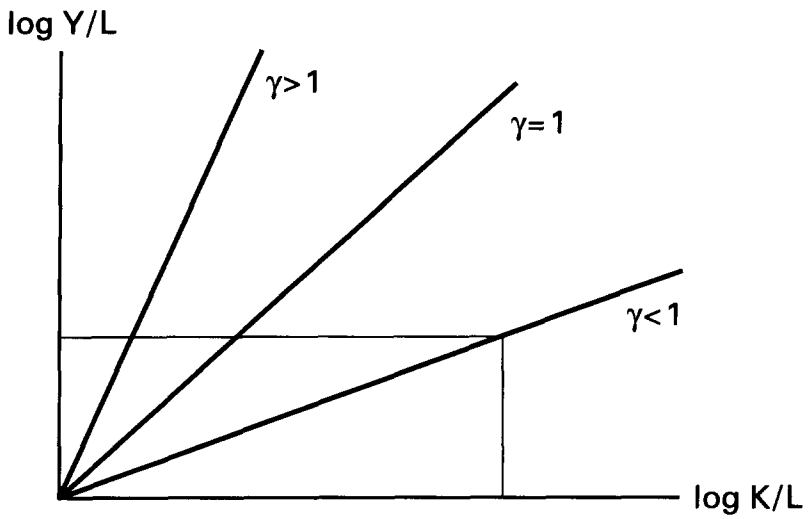

Fig. 1.

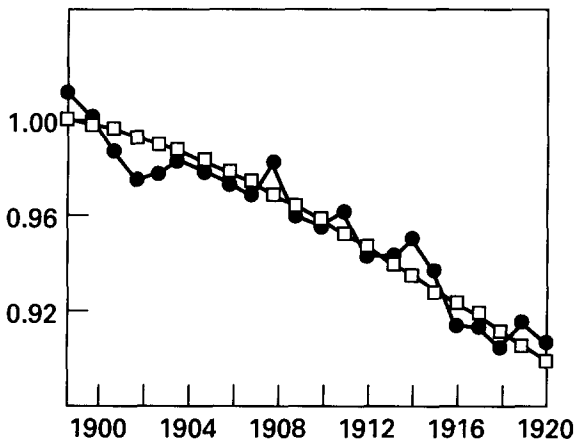

Fig. 2. USA: Eq. (4). ●, real values; □, calculated values.

and in the two periods is shown in Figs. 2 and 3 (the data are those of the equations presented in Section 5).

The above analysis shows that – in contrast to the neoclassical interpretation of the Cobb–Douglas function – the exponent can be stable for limited periods and only by chance. Then, it is no wonder that, in the Cobb–Douglas counterparts, the "estimates of the parameters are not stable over time" (Walters, 1983, p. 27). It is worth pointing out that, in three of the four cases that are most 'favourable' to the Cobb–Douglas function (see Section 3, Table 1), the period observed covers the first two decades of this century: in the USA during this period productivity was increasing

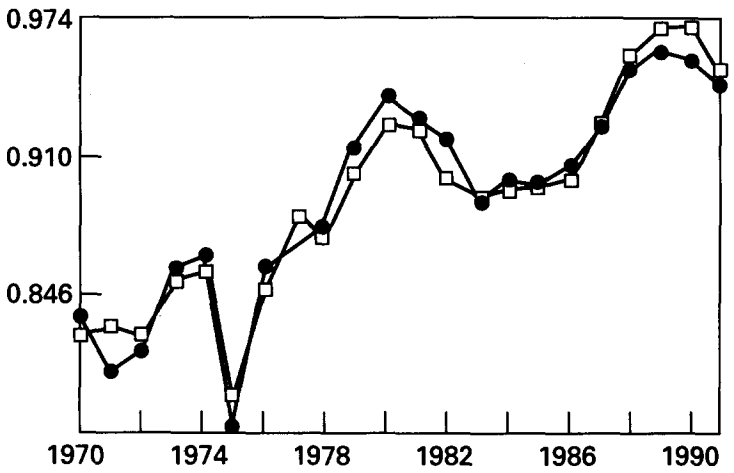

Fig. 3. Italy: Eq. (5). ●, real values; □, calculated values.

at a relatively low rate and K was increasing more rapidly than Y – a rather exceptional occurrence.

This way of posing the problem has the advantage of overcoming the dichotomy, which is fictitious, between capital accumulation and technological change. Indeed, the variations in the degree of mechanization as formulated here incorporate technological change and the variations in the angle of γ, express, at the same time, both the quantity and type of capital, which necessarily incorporates technological progress and, in particular, labour-saving technological progress.

9. Growth models and income distribution

Solow's model, which takes account of technical progress by means of a trend factor, is at the beginning of a series of growth models that are flanked by Harrod–Domar-type growth models. The main difference between these two types of models, apart from their formal specifications, lies in the fact that the former claims to simultaneously explain, even only as a first approximation, the main features of income distribution as well as of growth.

If we abandon the marginalist interpretation of the aggregate production function and consider the meaning of the relationship $\log(Y/L) = \gamma \log(K/L)$, then we have the advantage of freeing ourselves from the constraints and logical contradictions already noted; and also – I think – of breaking the series of growth models founded on the aggregate production function that, by their nature, hinder a proper analysis of technological progress. We therefore find outselves with a growth model that has a very limited explanatory value, but is amenable to generalization in terms of multi-sector models.[3]

[3] For further related literature, the reader is referred to Arcelli (1962, 1967), Castellino (1992), CEE (1961), Simon and Levy (1963), Solow (1956) and Sylos Labini (1989).

10. Concluding remarks

Two concluding remarks and a wish now follow. First, the theoretical criticisms of the concept of aggregate capital are to be considered jointly with the criticisms concerning the empirical results. Now, if we reflect on the strength of the theoretical criticism and on the flimsiness of the empirical testing of the time series, one might wonder how the Cobb–Douglas function showed such vitality – a vitality confirmed by the recent new wave of models that, in one way or another, are connected with that function. My amazement is even greater, because certain criticisms – such as those of Mendershausen and Phelps Brown – are quite old. The explanation is not difficult: the Cobb–Douglas production function represents an essential component of marginalist construction, especially the marginalist theory of distribution. The fundamental problem is the way substitution is to be conceived: in traditional theory, only static substitution is considered, which assumes given technology, whereas the decisive phenomenon for interpreting the growth process of modern capitalism is that of dynamic substitution. A switch from one technology to another, both known, can certainly take place when relative factor prices vary. However, dynamic substitution is much more important than static substitution, being the essential element of the growth process. Thus becomes clear if we refer to a very long period – mechanization, then automation and, lastly, in certain areas, robotization of productive operations, are simply different phases in a single process – which expresses long-term dynamic substitution and relates essentially to the growth in the production of commodities. These phenomena must be linked to the increasing relative scarcity of labour, which must naturally be conceived in dynamic terms and which is translated into an almost uninterrupted increase in real wages.

Secondly, a multi-sector dynamic analysis presents serious difficulties from both the theoretical and empirical points of view, especially if we accept the idea that every effort must be made to link the two aspects. An analysis of this kind is therefore very difficult but, at this point, it is the only one that promises important results.

In a monograph published many years ago, I presented a three-sector dynamic model and examined the consequences on the price system, production and employment of an innovation introduced into one of the sectors, by considering alternative market forms. In that model, I considered technological progress as exogenous. In successive papers, I tried to distinguish technological progress generated by endogenous impulses, i.e. by impulses that originate within the economic system, from the exogenous technological process, with the warning that the former is that which goes on continuously and deserves greatest attention from the economist, even if innovations that occur discontinuously are generally based on more important scientific inventions: the impulses that make exponent γ vary (discussed above) are precisely those that promote endogenous technical progress and, as a consequence, generate the systematic increase in labour productivity. For the three-sector model and for the equations concerning exponent γ and productivity, I consider my analysis to be preliminary.

Pasinetti followed a different, but logically not dissimilar, route to mine and, in his analysis founded on the concept of the vertically integrated sectors, examined

three interconnected structural dynamics: those of prices, of production and of employment (Pasinetti (1981) summarizes and develops previously published analyses). He maintains that production dynamics depend on those of demand, which according to Engel's law are not, and cannot be, proportional. I do not see a contrast between Pasinetti's and my analyses, although there are at least two differences. He considers all technological progress as exogenous and ascribes a decisive role to demand and to the scale of needs put forward by Engel's law. I think that such a law applies when average individual income increases, starting from very low levels. In these conditions, demand flows prevalently to goods that satisfy primary needs; however, above a certain income level, Engel's law – understood as a somehow predetermined sequence that starts with goods that satisfy primary needs and moves towards those that satisfy ever less pressing needs – has a decreasing interpretative value, as discontinuous innovations that lead to the production of completely new goods (such as television) and the pressure exercised by firms on consumer preferences by means of advertising become ever more important. However, in such conditions, the analysis becomes, at the same time, more indeterminate and more complicated. However, if Engel's law is invoked only to justify the hypothesis of the non-proportional growth of demand, I have no objection, as I have no objection to the assumption that technological progress is autonomous if this is introduced only for ease of analysis and does not imply the concept that technical progress depends on a scientific development that proceeds independently of the economy. For me, the same dynamic substitution must be seen as a process stimulated by variations of economic quantities; essentially, wages and machinery prices. Furthermore, I must emphasize that the concept of dynamic substitution proposed by Pasinetti (who prefers to speak of change of techniques, or of changes in technical methods over time) coincides with the concept that I propose, as can be seen by Pasinetti's paper published in 1977 and cited in the References. It seems evident to me that the two ways of facing the problem of non-proportional growth (Pasinetti's and my own) converge. The convergence seems clear if we compare our models with the multi-sector models that have been developed in recent years and in which the characteristic of non-proportionality in growth and in price variations does not appear, which is an essential characteristic of multi-sector models where innovations are introduced.

Finally, I hope that young economists will devote fewer efforts to static models and to the logically similar proportional-growth models, and instead try to develop the type of dynamic approach that is here proposed.

References

Arcelli, M., 1962, La funzione della produzione strumento per la programmazione, ISCO, Rome.

Arcelli, M., 1967, Variazioni qualitative dei fattori e progresso tecnico (Giuffré, Milan).

Arrow, K. J., H. B., Chenery, B. S. Minhas and R. M. Solow, 1961, Capital-labor substitution and economic efficiency, Review of Economics and Statistics, 43, no. 3, 231–246.

Castellino, O., 1992, Una breve introduzione al modello di crescita endogena, Economia politica – Rivista di teoria e di analisi IX, no. 3, 387–404. (I was induced to reconsider the Cobb–Douglas function after reading a preliminary draft of this paper in 1991.)

CEE, 1961, Metodi di previsione dello sviluppo economico a lungo termine, Informazioni statistiche, Paris.
Douglas, P. H., 1934, The theory of wages (Macmillan, New York).
Douglas, P. H., 1948, Are there laws of production?, American Economic Review.XXXVIII, no. 1, 1–41.
Douglas, P. H., 1967, Comments on the Cobb–Douglas production function, in: M. Brown, ed., The theory and empirical analysis of production, (National Bureau of Economic Research, New York).
Ferguson, C. E., 1969, The neoclassical theory of production and distribution (Cambridge University Press, Cambridge).
Ferguson, C. E., 1972, The current state of capital theory: A tale of two paradigms, Southern Economic Journal, Vol. XXXIX, no. 1, 160–176.
Fisher, F. M., 1971, The existence of aggregate production function: Reply to Mrs. Robinson, Econometrica IXL, no. 2, 405–406.
Guarini, R. and F. Tassinari, 1990, Statistica economica – Problemi e metodi di analisi (Il Mulino, Bologna).
Harcourt, G. C., 1972, Some Cambridge controversies in the theory of capital (Cambridge University Press, Cambridge).
Jorgenson, D. W., 1987, Production functions, in: The Palgrave Dictionary (Macmillan, London).
Kendrick, J. W., 1961, Productive trends in the United States (National Bureau of Economic Research, Princeton University Press, Princeton, NJ).
Mendershausen, H., 1938, On the significance of Professor Douglas's production function, Econometrica 6, no. 2, 143–153.
Mendershausen, H., 1939, A correction, Econometrica 7, no. 4, 362.
OECD, 1964, Le facteur résiduel et le progres économique, Paris.
Pasinetti, L. L., 1959, On concepts and measures of changes in productivity, Review of Economics and Statistics, Vol. XLI, no. 3, 270–286.
Pasinetti, L. L., 1977, On non-substitution in production models, Cambridge Journal of Economics, 1, no. 4, 389–394.
Pasinetti, L. L., 1981, Structural change and economic growth, A theoretical essay on the dynamics of the wealth of nations, (Cambridge University Press, Cambridge).
Patterson, K. D. and K. Schott, eds., 1979, The measurement of capital (Macmillan, London).
Phelps Brown, E. H., 1957, The meaning of the fitted Cobb–Douglas function, Quarterly Journal of Economics 71, no. 4, 546–560.
Robinson, J., 1953–54, The production function and the theory of capital, Review of Economic Studies 21, no. 1, 81–106.
Romer, P. M., 1987, Crazy explanations for the productivity slowdown, Macroeconomic Annual (National Bureau of Economic Research, MIT Press, Cambridge MA), 163–210.
Roncaglia, A., 1978, Sraffa and the theory of prices (John Wiley, London, New York).
Simon, H. A., 1979, On parsimonious explanations of production relations, The Scandinavian Journal of Economics 81, no. 4, 459–474.
Simon, H. A. and F. K. Levy, 1963, A note on the Cobb–Douglas function, Review of Economic Studies 30, no. 1, 93–94.
Solow, R. M., 1956, A contribution to the theory of economic growth, Quarterly Journal of Economics 70, no. 1, 65–94.
Solow, R. M., 1957, Technical change and the aggregate production function, Review of Economics and Statistics 39, no. 2, 312–330.
Sraffa, P., 1960, Production of commodities by means of commodities (Cambridge University Press, Cambridge).
Sylos Labini, P., 1984, The forces of economic growth and decline (MIT Press, Cambridge MA).
Sylos Labini, P., 1989, Nuove tecnologie e disoccupazione (Laterza, Roma-Bari).
Sylos Labini, P., 1993, Economic growth and business cycles – Prices and the process of cyclical development (Edgar Elgar, Aldershot).
Walters, A. A., 1983, Production and cost functions: An econometric survey, Econometrica 31, no. 1, 1–66.