Increasing returns and competitive equilibrium—the content and development of Marshall’s theory

Renee Prendergast*

"The root of the difficulty about increasing returns has always been to understand how, where they prevail, equilibrium can exist without the whole supply of the commodity in question becoming concentrated in the hands of one producer." (Robertson, 1930, p. 67)

Introduction

In his Recherches sur les Principes Mathematiques de la Théorie des Richesses first published in 1838, Cournot suggested that the cost of most manufactured articles became proportionately less as production increased. This he attributed to better organisation of the work, discounts on the price of raw material for large purchases and reduction of what was known to producers as general expense (Cournot, 1960, p. 59). Cournot took the view that declining marginal costs were not consistent with competition. Under the hypothesis of unlimited competition, it was necessary to assume that the marginal costs of each firm rose as output increased because, otherwise, firms’ revenues would be insufficient to cover the costs of production. Moreover, in conditions of unlimited competition, if firms had declining marginal costs, nothing would limit the production of the commodity, so no equilibrium would be possible. Cournot concluded by stating that if net income was earned by property or plant, the operation of which involved declining marginal costs, this proved that ‘the effect of monopoly is not wholly extinct, or that competition is not so great but that the variation of the amount produced by each individual producer affects the total production of the article and its price to a perceptible extent’ (ibid., pp. 91–92).

Marshall read Cournot’s Recherches in 1868 and his early work in marginal analysis was largely an extension and development of Cournot’s (Whitaker, 1975, I, pp. 840–845, 117, 121). His intellectual debt to Cournot was acknowledged in the Preface to the first edition of the Principles where he made it clear that he regarded Cournot as the pioneer of marginal analysis (Marshall, 1961, p. x). While the first edition of the Principles contains no direct criticism of Cournot, all subsequent editions contain a somewhat puzzling footnote in which Marshall represents Cournot as having constructed a downward sloping supply curve for the individual firm and having then failed to notice that in such a situation ‘whatever firm first gets a good start will obtain a monopoly of the whole business of its

*Department of Economics, Queen’s University, Belfast. I am greatly indebted to referees of this journal for their helpful comments on earlier drafts.

© 1992 Academic Press Limited
trade in its district' (Marshall, 1961, I, p. 459, n). The issue is clarified somewhat in the Economic Journal of 1898, where Marshall recorded his dissatisfaction with Cournot's treatment of industries conforming to the law of increasing returns in the following terms:

The great Cournot himself misapplied mathematics here. He ignored the conditions which, in real life, prevent the speedy attainment of monopoly by a single manufacturing firm: and the general drift of his argument is practically misleading. His failure contributed to make me hold back my diagrams as to value from formal publication for twenty years (Principles, II, p. 69).

In acknowledging his debt to Cournot in the first edition of the Principles, Marshall made the telling comment that Cournot was to be regarded as the chief influence not on the substance of the Principles but on its form. Credit for having been the chief influences on the substance went to Hegel and to Herbert Spencer. Since a subject-matter influenced by the Hegelian dialectic and Spencer's evolutionary biology and sociology was unlikely to sit comfortably within Cournot's static equilibrium framework, it is not surprising that Marshall's attempted synthesis landed him in considerable difficulties. The puzzle is why he chose to attempt such a synthesis in the first place. One possibility is that, given his mathematical training, Marshall was initially attracted by the apparent rigour of Cournot's analytical framework. Moreover, having visualised economic progress as the outcome of a competitive process, Marshall may have thought he could at least begin to analyse it using Cournot's concept of unlimited competition.

Since the publication of Smith's Wealth of Nations in 1776, classical political economists took it for granted that the division of labour gives rise to improvements in the productive power of labour. They also accepted Smith's proposition that the division of labour was limited by the extent of the market. Taken together, these two propositions allowed them to associate an increase in the output of most manufactured products with a fall in their costs of production. Seen in this way, increasing returns were regarded by Smith and others after him as a normal feature of the competitive economic system.

As Richardson (1975, p. 351) points out, in the Wealth of Nations there was no obvious tension between the theory of economic evolution and the theory of economic equilibrium. In Richardson's view, such tension emerged only with later writers who, in striving for greater analytical rigour, developed the theory of equilibrium in terms of a model of reality which was very different from that implicit in Smith's theory of evolution. An important feature of the new models was that competition came to be seen as price-taking behaviour rather than the active seeking out of new sources of advantage. Likewise, the law of increasing returns came to be seen as a functional relationship between output and cost of production rather than as a result of general economic progress. Once key economic concepts were re-interpreted in this way, it made sense to argue that the conflict between increasing returns and equilibrium was always latent in the classical model and was bound to rear its head as soon as any attempt was made to reformulate classical economics into 'a comprehensive and internally consistent system' (Stigler, 1951, pp. 185–186).

As a result of his attempt to recast classical economics within Cournot's static equilibrium framework, Marshall inevitably had to confront the problem of reconciling increasing returns and competition. It is generally recognised that, in doing so, Marshall developed three different lines of thought. These were: first and most importantly, the concept of external economies; second, the concept of the life cycle of the individual firm and, last, the downward-sloping demand curve for the individual firm. While it is widely accepted that

1 The issue here is that if there are always economies of further subdivision present which an increase in industry output will make viable, then for a time at least that particular stage will be characterised by monopoly rather than competition.
Marshall thought that each of these three elements could have had a role to play, 'the apparent inconsistencies' in his treatment have meant that there remains considerable controversy about the precise role of each of these theories and their relationship with each other (Whitaker, 1990, p. 44).

This article seeks to take forward the debate on Marshall's intentions by examining the evolution of his thought with particular reference to the role of external economies and the life cycle of the firm. The first section briefly considers the merits of the two main interpretations of the reconciliation exercise in the *Principles*. The second section examines the development of Marshall's thought on increasing returns prior to the *Principles*. In the third section, the discussion of increasing returns in the *Principles* is reviewed and interpreted.

1. The interpretations

Marshall's reconciliation of increasing returns and competitive equilibrium has received two very different interpretations. The first interpretation associated, broadly speaking, with Edgeworth (1925), Pigou (1913, 1928), Sraffa (1925) and Viner (1931) assigns the key role to external economies. According to this view, for a given aggregate industry output a 'U' shaped long-run average total cost curve can be drawn for the representative firm. In long-run equilibrium, the firm will be operating at the minimum point of this curve with no unexhausted internal economies. If the aggregate output of the industry increases, this gives rise to external economies. As a result, the shape of the firm's average total cost curve may change and its minimum point (which will have fallen) may correspond to a higher or lower level of output.

The second interpretation, initially associated with Robertson (1930), regards external economies as inessential to the reconciliation exercise and assigns the key role to the life cycle of the firm. Robertson argued that it was possible to conceive of an equilibrium in which the representative firm was working under conditions of decreasing cost with price equal to the average costs of that firm and the industry as a whole obeying the law of increasing returns. If this initial equilibrium was disturbed by an increase in demand, one of the factors in the progress towards the new equilibrium would be the attempt by individual firms to reap the advantages of large-scale production. When the new equilibrium had been reached, the industry would be found to be composed of larger firms producing at lower costs and the representative firm could be said to have grown. Since, on this view of things, the representative firm was assumed to be in equilibrium with price in excess of marginal cost, it was necessary to assume that something was counteracting its inducement to expand. According to Robertson, this is the role that was fulfilled by the life cycle of the firm.

At first sight Robertson's interpretation seems to have the merit of being both plausible and in line with Marshall's intentions. However, as Sraffa (1930) noted, a closer inspection of the argument reveals a serious problem. This is that the life cycle is being given the task of preventing the firm from growing in one position of equilibrium but of permitting it to become larger in another. Since Robertson had explicitly assumed that external economies (i.e. all the means by which a change in the output of the industry can affect the conditions of production of individual firms) were absent, no mechanism was available by means of which an expansion of the industry could produce circumstances favouring large firms (Sraffa, 1930, pp. 91–92).

1 For details of the precise differences between Edgeworth, Pigou and Viner, see Chipman (1965).
This problem was taken up by Shove in his contribution to the Economic Journal symposium. His solution involved the argument that an increase in demand for the output of the industry as a whole would increase the rate at which individual businesses could grow and thus the internal economies available to them at any given stage of their career. The more rapid rate of growth was attributed to the improved chances of obtaining trade for new, young or fortunate business. In a different context, this explanation might have been reasonable enough but it was hardly plausible given that, in the relevant section, Shove was assuming a perfect market and regarded external economies as unnecessary to his argument.¹

A slightly different argument was put forward by Frisch (1950). His solution was based on the assumption that to each alternative quantity produced in the stationary situation there corresponds ‘a typical form of the life-cycles of the firms’, and the life cycle will tend to make for larger firms the larger the aggregate output of the industry (ibid., p. 511). Frisch, however, failed to adduce any solid evidence that Marshall thought that the life cycle of the firm would behave in the prescribed way. The statements he refers to in the Principles simply suggest that, during a period when above normal profits are being made, there will be an increase in the rate of growth of rising firms and a slackening of the rate of growth of declining firms.

As long as one is referring to the period during which demand conditions remain favourable, there is no reason to quarrel with the view that ‘an expansion of demand, by making its environment more favourable, tends to prolong the period of growth and to increase the rate of growth during that period’ (Loasby, 1978). But the question that must be asked is: what grounds have we for believing that the longer life cycle and faster rate of growth will continue to persist when the industry in question has adjusted fully to the increase in demand? Unless there are some such grounds, it cannot be assumed that new firms coming into being will be able to grow to the same extent. Since demand conditions are no longer unusually favourable, it cannot be presumed that the life cycle will continue to be prolonged nor that the speed of growth will continue to be fast unless there has been some change in the conditions of reproduction of the industry. This point, I shall argue, was understood by Marshall which is why ‘external economies’ and not just the life cycle on its own featured in his treatment of the reconciliation exercise in the Principles.

2. The early writings

The Essay on Value²

The Essay on Value, composed around 1870, contains Marshall’s first systematic statement of the partial-equilibrium theory of price determination. Having defined equilibrium as the point of intersection of the demand and supply curves, Marshall proceeded to discuss the conditions under which equilibrium would be stable. In this context, he went on to consider the factors which gave rise to curves of different shapes. These, he suggested, mainly related to the length of the period of time—A, B, C, D—to which the investigation applied. Both periods C and D were sufficiently long for output to change in response to changes in demand. For periods of length C, no new modes of production could be introduced, whereas for D, such changes in the mode of production as were consequent on changes in the amount produced were allowed for (p. 139). These might

¹ See Shove, 1930, section 13 and the final paragraph of section 11.
² Unless otherwise stated, all page references are to ‘The Essay on Value’ in Whitaker, 1975, vol I.
Increasing returns and competitive equilibrium

include ‘systematised division of labour and the application of machinery already invented but which previously to the change it had been inexpedient to employ’ (p. 140).

The supply curve which was upward-sloping in periods C could take a variety of forms in periods D. Marshall’s first example of a downward sloping long-run supply curve was based on the falling price of steel pens between 1820 and 1867. The ‘example’ provided no information on changes in the quantities produced. Furthermore, as Marshall acknowledged, the major source of price reduction was probably inventions which were independent of the amount produced. All of this seems to suggest that, at this stage, the concept of the long-run supply curve was the product of Marshall’s approach to value theory rather than any empirical evidence.¹

Having examined the forms which the supply curve might take in periods of different lengths, Marshall noted that, with a downward-sloping supply curve, multiple equilibria were possible. He suggested that the existence of two points of stable equilibrium corresponded to the case where an article which ‘after being confined to the rich for a long time, suddenly came into ordinary use’ (p. 148). Such a transition might result from a temporary increase in demand which made it profitable to increase production. By the time the disturbance was removed, output might have reached a level such that it would tend to move towards the high output equilibrium rather than the low output one. More often, Marshall thought, the transition was initiated by some individual manufacturer who ‘observing that great economy would result from production on a large scale makes use of extra machinery and a more elaborate scheme of division of labour’ (pp. 150–151).² By selling at a lower price, he attracts customers from other firms as well as new customers who were unable to purchase the commodity at the previously ruling price. Since other producers have not yet increased their output, such a firm can sell at prices above the cost of production and make large profits. Other producers are, however, compelled to follow suit. Competition lowers the price and increases the total amount sold but profits remain large until sufficient new capital has been attracted into the market for the higher equilibrium level of output to be reached.

In attempting to explain how the transition between two points of stable equilibrium might be made, Marshall was forced to confront the problem of possible sources of increasing returns. His theory required that the fall in costs should depend on increases in aggregate output but if the source of such cost reductions was the use of extra machinery and a more elaborate division of labour by individual firms which were increasing their output, there seemed no obvious reason why the same cost reduction could not have been achieved by the displacement of other firms rather than the expansion of the industry. The following passage indicates that Marshall recognised this problem and had begun his search for a solution:

It has been tacitly assumed that an increase in the economy of labour which results from production on a large scale depends on an increase in the total amount produced. If capital moved perfectly freely and there were no practical limit to the proportion of the whole trade connection which a firm can obtain this result might often be brought about through the displacement of small manufactories by one or a few large ones. As it is this cannot in general happen without a social change which is by

¹ The absence of any information on quantities produced seems curious in an example intended to illustrate the downward-sloping supply curve. The example may have been taken from Marx’s Capital, I, p. 433 where it was used in the context of a discussion of the transition of industries through the handicraft and manufacturing stages to the factory stage.

² An argument along similar lines emphasizing entrepreneurial behaviour and a process view of competition is contained in the first edition of the Principles but, from the second edition onwards, the transition from a low-output to a high-output stable equilibrium was regarded as being initiated solely by a temporary increase of demand of sufficiently long duration. See Marshall, 1961, II, pp. 802–803 and I, pp. 805–806.
the very definition of the curves excluded. And the extent of the assumption made is that the smaller the total amount produced be, the more difficult it is to start and to keep employed large manufactories: and that this difficulty remains in general tolerably constant for any given set of social conditions. This difficulty may however occasionally be overcome in an unexpected degree, and the curves rendered inaccurate: and this will happen most frequently in those trades in which the final manufacture is not dependent for some of its stages on subsidiary trades in which increased economy of labour is not readily induced without an increase in the total amount demanded (p. 151).\(^1\)

*The Pure Theory of Domestic Values*\(^2\)

*The Pure Theory of Domestic Values* was begun in 1873 or 1874 and printed for private circulation by Sidgwick following the onset of Marshall’s illness in 1879. In this essay, Marshall directly confronted the question of whether increasing returns in manufacturing require the concentration of production in large establishments. He began by noting a defect in the customary treatment of the advantages of division of labour and production on a large scale. This was that the manner in which such advantages were discussed seemed to imply that ‘the most important of them can only be obtained by the concentration of large masses of workmen in vast establishments’ (p. 195). If this were the case, the introduction of economies into the process of manufacture would not depend directly or mainly on the total output of the particular commodity. Relatively small industries might achieve most of the advantages of production on a large scale if production was concentrated in a small number of firms, while industries of far larger dimensions might reap none of these advantages if they were conducted almost entirely by small masters.

While acknowledging that the disadvantages which small masters are at in competition with larger firms were increasing, Marshall argued (p. 196) that in industries such as the metal trades, the advantages of production on a large scale could be attained almost as effectively by the aggregation of a large number of small masters in a particular district as by the setting up of a few large-scale works. The reasons for this related, first, to possibilities for vertical disintegration and the development of subsidiary industries; second, to the occurrence and dissemination of new ideas.\(^3\)

Marshall noted (p. 196) that with regard to many classes of commodities, it might be possible to divide the production process into several stages in each of which the maximum efficiency of production might be attained by small firms. Where a sufficiently large number of small firms specialised in particular stages of the production process, there would be potential for the development of subsidiary industries adapted to meeting the special needs of these firms (p. 197). Such subsidiary trades might be engaged in the production of special tools and machinery, in the supply of materials and parts required by the small establishments or in the collection and distribution of their output. In some cases a very large demand was necessary to enable these subsidiaries to operate efficiently.

\(^1\) As noted by Whitaker (1990, pp. 42–44), the introduction of subsidiary industries into the equation shifts rather than solves the problem, for the reasons outlined in n. 1, p. 448 above. However, as Whitaker also notes, if the new unit’s market is contestable, it will be forced by potential competition to charge competitive rather than monopoly prices. The suggestion in this passage that capital does not move freely would seem to imply that exit is not costless, in which case we are not entitled to assume that the competitive result will apply. While there is considerable evidence that Marshall regarded potential competition as important, there is no evidence that he believed that potential competition was as effective as actual competition or that he explored the conditions in which this might be so.

\(^2\) Unless otherwise stated, page references are to *The Pure Theory of Domestic Values* in Whitaker, 1975, vol. II.

\(^3\) The benefits of agglomeration enumerated here are not novel. Similar points had been made as early as 1683 in Petty’s *Another Essay in Political Arithmetic Concerning the Growth of the City of London*. See Hull (ed.), 1899, vol. II, pp. 473–475.
Increasing returns and competitive equilibrium

With regard to the development of specialised skills, Marshall argued that where large numbers of men in the same locality are engaged in similar tasks, it would be found that by associating with one another, they educated one another. Furthermore, because of the large numbers engaged in particular trades, there would be scope for the allocation of difficult tasks or supervisory roles to those most fitted for them. The existence of a large pool of skilled labour would also make the expansion of particular businesses easier and lessen the risk that business would be hampered by the death or illness of key employees (pp. 197–198).

Finally, Marshall argued that where only a small number of firms are engaged in a particular industry, only a few people would be in a position to make improvements in the processes of manufacture and invent new machines and new methods. If, however, the total number of people engaged in the industry was large, more people of the intellect and temper necessary to originate new ideas would be available. Each new idea could be improved upon by more minds and would be more likely to result in practical improvements. Marshall noted that the growth of trade newspapers facilitated the dissemination and development of new ideas even where industry was not strongly localised. Here again he emphasised that such publications could only exist where the industry they addressed was sufficiently large (p. 198). On the basis of these arguments, Marshall concluded that an increase in the total production of a commodity could ‘scarcely fail’ to occasion increased economies in its production regardless of whether production was spread among a large number of small capitalists or concentrated in a comparatively small number of large firms (p. 198).

Marshall’s bid to establish that an increase in the aggregate output of a commodity would result in economies of scale, regardless of how this output was partitioned among firms, involved two different arguments. The first, in modern terminology, was that scale economies were not a sufficient condition for monopoly and that monopolisation would not occur if the cost function was not strictly subadditive.1 The second of Marshall’s arguments involved the claim that an increase in aggregate output of a commodity would result in increased economies in its production. While Whitaker sees the elements making up these arguments as Marshall’s first unambiguous statement of the doctrine of external economies, at this stage the concept did not have the generality it later acquired in the Principles. To a substantial extent, ‘external economies’ were regarded as a means of allowing large numbers of small firms to achieve the economies which would otherwise be achievable only by extremely large firms. Furthermore, most of the sources listed by Marshall were dependent on the localisation of industry for their verification. As Sraffa, commenting on similar material in the Economics of Industry, noted ‘the fact that their influence was conditioned by the localisation of industry makes it apparent that they could not be at the root of the tendency towards increasing returns connected exclusively with the increase of production’ (1925).2

The Economics of Industry

The Economics of Industry was begun in 1876 and published in 1879 under the joint names of Marshall and Paley Marshall (Whitaker, 1975, vol. I, pp. 13, 58–63, 67–83). Increasing returns are discussed in a chapter entitled ‘Division of Labour’. This develops the

---

1 A cost function is subadditive if it is cheaper to produce a given output vector in a single firm rather than partition it between two or more firms.
2 All references to Sraffa, 1925 are based on a translation kindly provided by Professor Roncaglia to Eric Rahim, my former colleague at the University of Strathclyde.
3 Unless otherwise stated, all references in this section are to A. Marshall and M. Paley Marshall, 1879.
material already contained in *The Pure Theory* and sets out in a concise form material which was later to occupy several chapters in the *Principles*.

The chapter opens with a discussion of the main advantages of the division of labour. It is emphasised that these advantages cannot be obtained unless the demand is sufficient for the commodity in question to be produced in large quantities (p. 52). The question of whether or not the advantages of the division of labour are dependent on the size of the factories in which the work is done is then examined. Again it is argued that while some of the advantages of the division of labour can be obtained only in very large factories, many of them can be secured by small factories and workshops provided that there are large numbers of them in the same trade. Where a manufacturing process can be divided into several stages, it will become profitable as output grows to devote separate factories to each of the stages. Furthermore, if there are many factories large or small all engaged in the same process, subsidiary industries will grow up to meet their special wants (p. 52). While the growth of subsidiary industries means that small firms need not be excluded from the advantages of the division of labour, the Marshalls suggest that small factories, whatever their numbers, will be at a disadvantage relative to large ones unless they are collected together in the same district (p. 53). The discussion of localisation itself is along the lines of that already contained in the *Pure Theory of Domestic Values*. Towards the end of the discussion, it is noted that while both large and small factories may benefit from the localisation of industry and the assistance of subsidiary trades, these benefits are most important to the small factories (p. 53).

While small firms enjoyed more of the benefits of division of labour than was commonly thought, it was still the case that large firms enjoyed many special advantages. These included greater economies in such matters as buildings, steam engines, auxiliary machinery, maintenance and administration, etc. Large firms could also afford machinery to do work that was done by hand in smaller firms and they had advantages in matters relating to innovation (pp. 53–54). Most important of all, they enjoyed enormous benefits in buying and selling. Their transactions costs did not rise proportionately with the amounts bought and sold. They were able to install special facilities for loading and unloading and to negotiate cheaper rates for freight. They had better access to information and could afford to advertise (p. 54). Finally, the owner of a large firm could economise on his own skill by concentrating on the most important management functions and delegating others to subordinates. This advantage, however, was not as important as it seemed at first sight because the small firm enjoyed advantages of its own in some aspects of the management function.

Comparing the treatment of increasing returns in *The Economics of Industry* with that in the *Pure Theory*, a number of important differences of emphasis emerge. First, in contrast with the earlier effort to play down the importance of large establishments, the Marshalls stress the many advantages of large-scale firms. Second, whereas in the *Pure Theory* 'external economies' were seen mainly as a means by which many small firms could achieve the advantages normally associated with large establishments, in *Economics of Industry* they are regarded as benefiting both large and small firms.

The importance attributed to the size of the factories in which the work is done caused Sraffa (1925, pp. 25–26) to take the view that, in *Economics of Industry*, increasing returns depended mainly on plant size, and external economies were to be found only as embryo and as secondary elements. Sraffa further claimed that, in stressing the importance of the size of individual factories, Marshall was assuming among the causes of the decrease in cost a condition which is incompatible with free competition. According to Sraffa, this
Increasing returns and competitive equilibrium

error was later corrected in the Principles by making external economies the sole cause of increasing returns in a regime of competition. Sraffa’s views are not implausible if we restrict ourselves to the evidence contained in the Economics of Industry. However, when account is taken of the earlier writings, it seems unlikely that Marshall would have fallen into the error of making increasing returns depend mainly on the size of individual factories. The solution to this conundrum seems to be that Marshall believed that in the ‘life cycle’ he had found a new means reconciling competition and increasing returns to the individual firm.

In many industries a large capital can avail itself of great economies that are out of the reach of a small capital; and the large manufacturer can make higher profits than the small manufacturer. These industries would rapidly be concentrated in the hands of a few wealthy firms, if a man whose practical genius has created a large business, could ensure that his successors for several generations should have a like genius. But in the whole course of history we meet but with very few instances of private firms which have been managed with eminent genius for three generations in succession. The sons and grandsons of a successful man of business have seldom that rare combination of ability and assiduity which would enable them to carry on his work. And there are many instances in which a vast inherited business has been quickly destroyed by men who could have managed a small business well (pp. 141–142).

This was Marshall’s first use of the life cycle. Its purpose seems to have been to accommodate the empirical fact that internal economies were important while at the same time ensuring that the availability of cost reductions from the continued expansion of the firm over time would not lead to concentration. It is not clear, however, whether or not Marshall thought that the life cycle allowed internal economies to have a role in explaining the downward-sloping supply curve for the industry or whether it was intended to rule out such a role.

3. The Principles of Economics

Marshall’s Principles of Economics was first published in 1890. A total of eight editions were published during his lifetime, the last of them in 1920. Much of the material on industrial organisation in the Principles was an expanded version of that already contained in the earlier works. Important new elements in the Principles were: the introduction of the terms external and internal economies (first edition); the introduction of the concept of the representative firm (first edition) and the actual term (second edition).

Internal and external economies

The terms internal and external economies were first introduced in the summary chapter of the book on industrial organisation but came into more widespread use from the second edition onwards. The introduction of the term external economies was a formalisation of the position already present in the Economics of Industry according to which the economies of localisation were seen as benefiting both large and small firms rather than as a means by which small firms could obtain the economies which would otherwise be available only to extremely large firms. External economies were defined as those dependent on the general organisation of the trade, on the growth of knowledge and appliances common to the trade and on the development of subsidiary industries. Internal economies were defined as

1 Unless otherwise stated, all references in this section are to Guillebaud (ed.), 1961, Marshall’s Principles of Economics, vol. I or II as indicated.

2 The emphasis on subsidiary industries as a source of external economies is much less marked in the Principles than in the earlier works. According to Marshall, the most important external economies result from the growth of correlated branches of industry which mutually assist one another (Marshall, 1961, I, p. 317).
those dependent on the resources of individual houses of business, on their organisation and on the efficiency of their management. As such, they were seen as intimately bound up with the life of the individual firm, increasing with its rise to the first rank and dwindling with its decay.

The representative firm

The concept of the representative firm was one of Marshall’s major innovations in the Principles and he later described it as a device by which to avoid some of the difficulties and risks attached to the application of the statical method to industries conforming to the law of increasing returns (Marshall, 1898, p. 50). The concept, he claimed, was biological rather than mechanical and its application marked the transition from the mechanical view of the composition of forces to the biological notion of composite organic development (p. 50). An appreciation of the difference between these two concepts could be brought out by considering the notion of equilibrium. In the mechanical concept, demand and supply are considered as two crude forces pressing against one another and tending towards an equilibrium which persists in the absence of exogenous change. In the biological concept, equilibrium is conceived as a balance between the organic forces of life and decay and firms have no equilibrium states which were not subject to internal processes of change. ‘In each case, the forces of life preponderate at first; then those of crystallisation and decay attain to equal terms, and there is balance or equilibrium; afterwards decay predominates’ (Marshall, 1898, p. 43).

The biological concept of equilibrium meant that the firm reached maximum size when the forces making for further growth were exactly balanced by the forces of decay (I, p. 316). While some firms attained greater size than others, it was in general possible to identify a representative producer. Such a firm was defined as one which had a fairly long life and fair success, which was managed with normal ability and which had normal access to the economies, external and internal, which belong to the aggregate volume of production, account being taken of the class of goods produced, the conditions of marketing them and the economic environment generally (I, p. 317). While the representative firm might correspond to some actual firm, since firms were subject to internal processes of change, no particular firm could remain representative for long and only the representative firm could appear in any concept of industry equilibrium.

Marshall emphasised that the representative firm ‘was the particular sort of average firm, at which we need to look in order to see how far the economies, internal and external, of production on a large scale have extended generally in the industry and country in question’ (I, p. 318). He also claimed that the general argument of Book IV of the Principles showed that an increase in the aggregate volume of production of anything generally increased the size, and therefore the internal economies possessed by such a representative firm; and it always increased the external economies to which the firm had access (I, p. 318). This statement has been cited in support of the Robertsonian conjecture that, even if external economies were absent, internal economies and the life cycle could somehow be used to explain a downward-sloping industry supply curve. The strongest evidence that this was not Marshall’s position is contained in his Economic Journal article of 1898. There Marshall wrote:

Now the growth of internal economies is generally more rapid than that of external. The rise and fall of individual firms may be frequent, while a great industry is going through one long oscillation, or even moving steadily forwards; as the leaves of a tree ... grow to maturity, reach equilibrium and decay many times, while the tree itself is steadily growing upwards year by year. For very long
Increasing returns and competitive equilibrium

periods the oscillation of internal economies may almost be neglected: except in so far as they are indirectly dependent on external; for a large industry offers a better field in most (not in all) ways for large individual firms than a small industry offers (Marshall, 1898, p. 50).

While the above suggests that a large industry will tend to be characterised by large individual firms, we are left in no doubt that any increase in the size and internal economies of the representative firm is to be regarded as an indirect consequence of an increase in external economies. This passage, therefore, seems to cast strong doubt on the Robertsonian interpretation of Marshall and seems more in line with Sraffa’s view that, in the Principles, Marshall made external economies the sole cause of increasing returns in a regime of competition.

Why should Marshall have taken this position? The answer is straightforward if we view the representative firm as having exhausted its internal economies. In this case, since firm expansion would bring no reduction in cost, any increase in the size and internal economies of the firm following an expansion of the industry must have been generated indirectly by that expansion. But can all Marshall’s references to the ‘constant’ internal economies of the representative firm be taken to imply that the representative firm has exhausted its internal economies? It has to be said that there was considerable ambiguity on the matter. The following passage from Industry and Trade seems to build on the picture presented in The Pure Theory of Domestic Values of a large number of firms specialising in a particular stage of production and reaping the available economies of scale in their particular set of activities:

Not very long ago the representative firm in most industries and trades was a private partnership; which in the course of one or two generations had attained a goodly reputation, of a personal and individual character. Its plant had become larger and more various, until it commanded all, or nearly all, those economies of production of a large scale, that were inherent in the most advanced methods of production then known for its particular branch of business. Its own (internal) economies were not great; but it took its part in affording a large market for firms in branches of manufacture, which supplied it with made or half-made materials and in developing (external) economies of general organisation, which gradually became common property. Thus each firm, though of moderate size, might reasonably hope to obtain most of the advantages in production, which would be accessible only to vast businesses, if each had been mainly dependent on its own resources (Industry and Trade, pp. 314–315).

Marshall, however, then went on to note that joint stock companies had become universal in the railroad industry and common even in manufacturing industry. In many of these industries there was a tendency towards increasing returns and this meant that a firm which obtained a start on its rivals would be able to undersell them progressively provided its own vigour remained unimpaired and it could obtain all the capital it needed. For such goods, provided transport costs are low, there seemed to be nothing to prevent the concentration of the entire industry output in the hands of a single firm. The reason why this did not generally happen was that no firm ever had a sufficiently long life of unabated energy and power of initiative for the purpose, but this position was probably changed by the advent of the joint stock company.

This seems to suggest that Marshall intended that the representative firm analysis should apply to two quite different cases: first, the case initially outlined in the Pure Theory where the firm’s internal economies were exhausted and, second, the case where the representative firm would have substantial cost reductions open to it should it expand its output. In the first case, the life cycle has no analytical significance but, in the second case, it has the role of ensuring that the representative firm has the normal level of internal economies for the trade concerned. For a given level of industry output, this normal level
is constant but, in this context, constancy does not imply that internal economies are exhausted. Marshall was aware that if increased internal economies were readily available, it might be argued that the decay of human energies was a slow process and insufficient to counter the tendency towards monopoly. It was to deal with this objection that, in the second edition of the Principles, he introduced an argument to the effect that while the production of commodities might be increased quickly and economically, sales could not be so increased (I, p. 287).

Given that Marshall’s representative firm cannot always be regarded as having realised all the internal economies potentially available to it, it is necessary to explore the possibility that in such cases internal economies may have a direct role in explaining the downward-sloping industry supply curve. In approaching this question, we begin by considering a situation where all individual firms have access to further internal economies should they expand. We assume that any such expansion takes place gradually over time, that firms are subject to a cycle of growth and decay, that there are no market imperfections and no external economies. In a stationary state, each firm would reach a certain normal size before decay set in. Suppose now demand increases and that the rate of growth of firms is temporarily increased so that, with a given life cycle, the maximum size of firms in the industry is larger and firms have access to more internal economies. These firms are all subject to the life cycle and will eventually decay. The gains they have made will be lost unless there is some mechanism to insure that new firms can actually grow to the same level. But what is this mechanism to be? Since external economies are assumed to be absent and the industry has long since adjusted to the increase in demand which caused the initial change in growth rate, there are no grounds for assuming that the faster growth rate will have been maintained. The same argument applies to the lengthening of the life cycle. In Marshall’s theory this was determined by human biology rather than the theory of the firm and was thus unlikely to be correlated with the size of aggregate industry output.

If, however, the assumption that external economies are absent is now relaxed, these difficulties disappear. Following the increase in demand, Marshall’s representative firm can now grow for longer and in the process realise greater internal economies. Since aggregate output is greater, external economies will have increased with the consequence that the objective conditions of production facing new firms are different from before. As a result of these changes, even when the industry had fully adjusted to the change in demand and the life cycle returns to normal, the size of the representative firm may be different from what it was before.

While recourse to the texts and to logic seem to suggest that external economies were central to Marshall’s explanation of the downward-sloping supply curve, many commentators remain unconvinced. Their instinct seems to be that the purpose of the life cycle was somehow to give internal economies a role. I shall now argue that this intuition is correct but not for the reasons usually adduced.

To understand the thinking behind Marshall’s treatment of internal economies, it will be useful to start with a brief exposition of the external economies argument along the lines suggested by Chipman (1965, 1970). According to Chipman (1970, p. 349), each entrepreneur is assumed to believe that his firm is operating under constant returns to scale and any departures from this are interpreted as ‘a perturbation of his unit-homogeneous production function’ caused by changes in the output of the industry as a whole. Chipman illustrates the concept by considering Adam Smith’s pin industry. If a particular firm expands, some of the work will be divided and specialities will develop. However, only a substantial expansion of the industry will provide enough openings for a pool of labour to
Increasing returns and competitive equilibrium

develop with a specialised skill but the contribution of a single firm to this process will be so imperceptible that it will be neglected by the entrepreneur. The change in the character of the labour force will be regarded as exogeneous by all firms, even though each firm necessarily contributes to the process.

There is much in the above which would have been acceptable to Marshall but he would, I think, have felt it necessary to supply his individual firms with a motive for making new divisions of labour. As he explained in the Principles:

... the history of the individual firm cannot be made into the history of an industry any more than the history of an individual man can be made into the history of mankind. And yet the history of mankind is the outcome of the history of individuals; and the aggregate production for a general market is the outcome of the motives which induce individual producers to expand or contract production (I, p. 459).

A motive for implementing new divisions of labour would be provided if, during the adjustment period following an increase in demand, firms sought to exploit economies of scale which were potentially available before but which they were unable to exploit because they did not survive long enough. As a result of these attempts by individual firms to gain access to economies of scale, new external economies would be generated in the manner described by Chipman and these, in turn, would allow new firms to grow to a larger size even when supply had fully adjusted to demand. On this scenario, internal economies have a role in explaining the downward-sloping supply curve even though they are not themselves sufficient to account for it.¹

Hegel’s influence

Marshall’s acknowledgement of Hegel’s influence on the substance of the Principles has already been noted. This influence seems to have been especially important in Marshall’s treatment of increasing returns. Marshall’s vision of the role of the individual firm within the industry has strong parallels with Hegel’s treatment of the role of the individual within civil society. On the one hand, the single individual has to accept his world as given and assimilate himself to it (Hegel, 1975, pp. 58 and 83). On the other hand, all substantial change was produced by the activity of individuals as subjects (Hegel, 1952, p. 218). There are resonances of Hegel also in Marshall’s use of the life cycle as well as in his extensive employment of the analogy of the trees of the forest and others drawn from nature:

The tree lives perennially, puts forth shoots, leaves and blossoms, and produces fruit, and thus always starts again from the beginning. The annual plant does not survive its fruition, and although the tree can live for many decades, it too eventually dies. The reawakening of nature is merely the repetition of one and the same process; it is a tedious chronicle in which the same cycle recurs again and again. There is nothing new under the sun (Hegel, 1975, pp. 60–61).

In the natural world, development was confined to individual things whereas the species themselves were enduring. It was otherwise with spiritual forms; in this case, change occurred not just on the surface but within the concept. The concept itself was modified so that new possibilities were created. As a result, development consisted in genuine progress and not, as in the natural world, a continually recurring cycle (Hegel, 1975, p. 128).

¹ The key point here is that, in a growing market, the search by individual firms for new sources of advantage has the effect of creating a change in the conditions of production for all firms in the industry. Marshall generally took such changes to be irreversible, that is, they involved improvements in knowledge and organisation which would not be lost even if there was a decline in industry output. It is important to note that if the search by individual firms for competitive advantage is to have an impact on the industry as a whole in the required way, either their independent searches must take them in the same direction or, as seems more likely, successful innovations by one firm are emulated by others.
It is clear that Hegel's natural world in which the survival of the species consists of the repetition of one and the same mode of existence has much in common with Marshall's stationary state (II, pp. 382–383 and I, p. 367). What is less obvious is that, for Marshall, changes in external economies have roughly the same function as changes within the concept had for Hegel. With the same concept (conditions of reproduction), the rise and fall of individual entities produced no overall change whereas with a change within the concept (new conditions of reproduction) the rise and fall of individual entities resulted in genuine progress.

Hegel's account of how the modification of the concept or creation of new possibilities is accomplished by 'the great individuals of world history' is also illuminating. Such individuals 'do not find their aims and vocation in the calm and regular system of the present . . . they draw their inspiration from that hidden spirit whose hour is near but which still lies beneath the surface . . . they are the far sighted ones [they] have recognised the concept, the next universal to emerge' (Hegel, 1975, p. 83). Others flock to their standard, for it is they who express what the age requires. With some minor modifications, we have here an apt description of the entrepreneurial and competitive processes which Marshall saw as giving rise to economic progress in the chapters on industrial organisation in the Principles.

Summary and conclusions

By the time he published the first edition of his Principles, Marshall had formulated an ingenious theoretical solution to the problem of reconciling increasing returns and competition within the framework devised by Cournot. The solution involved the introduction of the concept of external economies which were to be viewed as the sole cause of increasing returns in a regime of competition. Interpreted as a perturbation of a firm's unit-homogeneous production function caused by changes in the output of the industry as a whole, external economies are a device of considerable power and elegance and a credit to Marshall's theoretical imagination. As Sraffa long ago pointed out, the real difficulties with the concept begin when we try to imagine possible sources which are consistent with a static equilibrium framework. Marshall himself seems to have been aware of this problem and, as shown above, in his efforts to overcome it he was led to employ a complex combination of external economy and life cycle arguments.

While Marshall's thinking on the problem is more ingenious and subtle than is commonly realised, his own attempts to reconcile declining costs and competition seem to have led him to recognise that real world sources of declining costs involved qualitative change through time and not timeless quantitative change of a reversible type such as were required by static equilibrium theory. In his Economic Journal article of 1898, Marshall stressed that the chief difficulties of economic science in his own time were occasioned by the rapid alteration of the conditions of work and life. These alterations involved changes in the character of economic and social forces and not simply changes in their magnitude. Consequently, mechanical analogies, whether static or dynamic in character, were inappropriate and it was necessary to adopt an evolutionary or organic approach (pp. 42–43). Static hypotheses could be used as temporary auxiliaries to dynamic—or rather biological—concepts; but the central idea of economics, even when its foundations alone were under discussion, must be that of living force and movement (II, p. 58). The static theory of equilibrium was 'barely even an introduction to the study of the progress and development of industries which show a tendency to increasing return' (p. 461). We were,
Increasing returns and competitive equilibrium

as he put it here, verging on the high theme of economic progress and such problems were imperfectly presented when they were treated as problems of static equilibrium and not organic growth (ibid.).

If the requirements of Marshall’s theory of economic progress and his equilibrium theory of value were in conflict, it might have been expected that later writers would abandon equilibrium theory and develop a theory of value whose requirements were not in conflict with the reality it purported to describe. With a few notable exceptions, the trend has been in the opposite direction. Equilibrium theory was further refined and, because they were incompatible with its requirements, increasing returns continued to be ignored. There are indications, however, that change is in the air. Evolutionary theories are beginning to flourish and ‘there are signs that the subject will return to its Marshallian affinities to biology’ (Hahn, 1991, p. 48). It is worth reminding those who would take the evolutionary approach that, as Sraffa (1930) suggested and as, I think, Marshall himself would have agreed, the construction of adequate foundations for it will require something more than an equilibrium theory of value coupled with more frequent recourse to the concept of external economies.

Bibliography

Pigou, A. C. 1927. The laws of diminishing and increasing cost, Economic Journal, vol. 37, June, pp. 188–197