Abstract

This paper shows that the Cournot-Marshall dilemma on the incompatibility of increasing returns and competition may be overcome under an integrated theory of the firm that takes into account not only returns to scale but also knowledge and transaction cost considerations. Knowledge and transaction-based considerations help to understand why, contrary to predictions, the tendency toward an increased dimension and monopoly may not operate in spite of the presence of significant economies of scale. If productive and transactional knowledge are not full and free, the development of capabilities and transactional relations may boost the presence of a greater number of producers and may impede the tendency towards monopoly. This suggests that the boundaries of the firm cannot be attributed solely to one single cause, but are instead the result of the interplay between knowledge-, scale- and transaction-based considerations.

Keywords: Cournot’s dilemma, transaction costs, capabilities, economies of scale
EAEPE research area: E - Theory of the firm
Thanks to the advantages of falling costs, whoever starts out first in the market will run everybody else out of the ... business. Does that mean that ... monopolies are inevitable? If scale economies are so important, how do small firms manage to exist at all? (Warsh, 2006, p. 46).

Introduction

This paper addresses the Cournot-Marshall dilemma on the incompatibility of increasing returns and competition. Building on interesting insights provided by the recent literature on this issue, the paper shows that this longstanding conundrum may be overcome under an integrated theory of the firm that takes into account not only returns to scale but also knowledge and transaction cost considerations.

Increasing returns to scale and competition are thought to be incompatible because, under increasing returns, the growth of a firm implies a rise in its efficiency which may lead it to dominate the entire market, driving all others out of a given sector. Increasing returns to scale occur when a rise in the scale of production leads a reduction in input requirements per unit of output. Given the input prices, this reduction brings about economies of scale that consist in a decrease in total average cost as the dimension of scale increases. If, for one individual firm in particular, average total cost decreases with the increasing size, the tendency towards monopoly conditions appears inevitable.

The following section discusses the analytical context in which the Cournot-Marshall dilemma originated. The subsequent and last section presents a possible solution of this dilemma, examining how knowledge and transaction-based considerations help to understand why, contrary to predictions, the tendency toward an increased dimension and monopoly may not operate in spite of the presence of significant economies of scale. If productive and transactional knowledge are not full and free, the development of capabilities and transactional relations may boost the presence of a greater number of producers and may impede the tendency towards

---

1 In the past few years the Cournot-Marshall dilemma has been discussed in some interesting articles published by The Cambridge Journal of Economics (see, Prendergast, 1992; Hart, 1996; and Niman, 2004). Recent contributions on this longstanding conundrum are also in Marchionatti (2006, p. 617ff.) and Morroni (2006, pp. 253-6).
monopoly. This suggests that the boundaries of the firm cannot be attributed solely to one single cause, but are instead the result of the interplay between knowledge-, scale- and transaction-based considerations; at the same time, there is an unavoidable subjective element concerning the way in which the firm’s management is able and willing to grasp opportunities for growth. In this respect, analysis of the connections between the three above-stated aspects of organisational coordination entails a new perspective that offers a rich agenda for future theoretical and empirical research.

1. The reconciliation problem

Alfred Marshall remarked that “some, among whom Cournot himself is to be counted,” have considered internal economies “apparently without noticing that their premises lead inevitably to the conclusion that, whatever firm first gets a good start will obtain a monopoly of the whole business of its trade …”. However, Marshall observed that this conclusion was challenged by empirical evidence showing that in numerous sectors of activity competition was thriving in spite of the presence of significant internal economies of scale.\(^2\)

Marshall attempted to reconcile increasing returns with competition on the basis of the life cycle of the representative firm and of a definition of competition that is far from the concept of perfect competition. His definition takes into account the existence of “particular markets” characterised by a decreasing demand curve for individual firms.

The representative firm is a “normal” firm in the sense that it is “managed with normal ability”, and it “has normal access to the economies, external and internal, which belong to that aggregate volume of production (Marshall, 1890, p. 265). Marshall maintains that in order to understand why increasing returns to scale are unlikely to result in monopolisation of an industry, it is necessary to consider that very few firms have a long-continued life of active progress because the degree of entrepreneurial

---

\(^2\) Marshall (1890, pp. 378, 380 fn.1); cf. Cournot (1838, p. 96ff.). Marshall’s parenthetic remark, “apparently without noticing”, has been judged puzzling because Cournot was, of course, perfectly aware that increasing returns imply the tendency towards monopoly. Marshall’s words may be attributed, on the one hand, to failure to consider the contradiction between the logical consequences of the theory and empirical evidence showing that increasing returns are widespread and, on the other, to the absence of investigation, on the part of Cournot, into the existence of forces that may counteract the tendency towards monopolisation. On this interpretation, see Hart (1996, p. 355, fn. 3).
inventiveness and energy is not constant over time. Accordingly, the tendency to monopoly is not inevitable because of the onset of diminishing returns in the life cycle of the firm (op. cit., p. 379). In arguing that the life of a firm rarely lasts very long, Marshall uses the celebrated metaphor of the “three in the forest”. Moreover, in the reconciliation exercise he abandoned the definition of perfect competition to embrace a definition based on freedom of entry. “Each firm – he writes - is likely to be confined more or less to its own particular market; and, ... any hasty increase in its production is likely to lower the demand price in that market out of all proportion to the increased economies that it will gain” (ibid.).

As convincingly argued by Roberto Marchionatti, Marshall’s attempt to solve Cournot’s dilemma “prevented the theory of competition from having a precise setting and became unacceptable for the newer generation of economists” (Marchionatti, 2006, p. 619). For this reason, some of Marshall’s followers and contemporaries tried to circumvent such a dilemma by referring to the role of external economies that derive from the development of trade and the growth of knowledge and equipment common to all firms operating in a sector of activity (ibid.). For instance, Pigou maintained that provided some external economies are common to all producers, “the presence of increasing returns in respect of all together is compatible with the presence of diminishing returns in respect of the special work of each severally. And this is sufficient to permit a stable equilibrium”. But although the concept of external economies was introduced by Marshall himself, it was used as a device to overcome Cournot’s dilemma mainly by Pigou and other followers of Marshall, whilst it played only a very minor role in Marshall’s own reconciliation exercise.

3 Neil Hart observes that “[f]rom the 6th edition of Principles …, Marshall added a significant qualification to the “tree in the forest” analogy by acknowledging the growing importance of joint-stock companies” without noticing, however, that “the emergence of joint-stock companies substantially altered his general proposition.” “The forced recognition that the emerging joint-stock companies could escape the inevitable decay imposed by his life cycle analogy severely challenged the relevance of Marshall’s solution to the reconciliation problem” (Hart, 1996, pp. 361, 366, passim).


5 Hart (1996, pp. 353-4). A similar idea is expressed by Marchionatti (2006, p. 619), who notices that “Marshall did not introduce external economies mainly as a device to overcome the dilemma on increasing returns and competition, but some Marshall scholars (for example Stigler 1941) disagree with this view.” For a discussion on a different line of interpretation which instead assigns a key role to external economies in Marshall’s analysis, see Prendergast (1992, p. 449).
In the 1920s, the developments of partial equilibrium analysis by Pigou, Edgeworth and others became the object of strong criticism and an intense debate known as the “cost controversy”. In particular, the use of the concept of external economies - as a device to overcome Cournot’s dilemma and to reconcile increasing returns with competitive equilibrium - was criticised by several economists. First the consideration of economies external to the firms but internal to the industry appears as an expedient of “scarce practical relevance” that does not limit the development of internal economies of scale and therefore the tendency toward monopoly (Young, 1913). Secondly, the inability to incorporate the time irreversibility associated with external (and internal) increasing returns challenged the viability of the static analysis of partial equilibrium. Thirdly, external economies appear inconsistent with the partial equilibrium framework, which is based on small changes in quantities produced, because they “are not likely to be called forth by small increases in production”.

To conclude this section, it can be argued that within the static analysis of perfect competition, the dilemma whereby the incompatibility between increasing returns and competition contrasts with the empirical evidence is an unsolvable impasse, because in this analytical context the presence of internal increasing returns inevitably leads to monopoly. As rightly noticed by Arrow (1979, p. 156) Cournot’s “dilemma has been thoroughly discussed; it has not been thoroughly resolved”. I will argue that the solution to this dilemma may be found if an integrated perspective of the boundaries of the firm within a dynamic analytical framework is considered.

---

6 It should be noted that Marshall did not proceed to the analytical demonstration of short and long-run equilibrium. The demonstration was carried out for the first time by Edgeworth and Pigou on the basis of the law of variable proportions, considering the firm simply as a production function. This new approach abandons the detailed study of the relationship between organisation and efficiency, which was characteristic not only of the classical economists but also of Marshall’s analysis itself. See, for instance, Moss (1984), Loasby (1988). On the “cost controversy” see references and discussion in Morroni (1998a, 1998b); Marchionatti (2001, p. 43ff., 2003, p. 49ff.); Raffaelli (2003, pp. 113-4).

7 Sraffa (1926, p. 540). As Prendergast (1992, p. 460) suggests, “Marshall himself seems to have been aware of this problem”.
2. A possible solution to the dilemma within an integrated perspective

Internal economies of scale play an important role in many sectors of activities.\(^8\) They are mainly due to the balancing of the various productive capacities of different production elements,\(^9\) the presence of fixed costs due to set-up processes, increasing division of labour, the three-dimensionality of space, reduction of heat losses and economies in holding reserves. Furthermore, knowledge characteristics and the presence of significant transaction costs may involve internal economies of scale and therefore may favour an increased dimension of individual firms.\(^10\) Growth of the firm may entail some economies which are due to the indivisibility of some unused resources or to some particular competence or market position that create a differential advantage in expanding the size of the firm.\(^11\)

If we leave out from the analysis the action of knowledge or transaction aspects and focus solely on the way internal economies of scale shape the growth of the firm, a precise evaluation of the role of internal economies may be hampered. The reason is that the effects of such economies on organisational boundaries may be positively or negatively influenced precisely by the above-mentioned excluded aspects, namely knowledge or transaction aspects. In contrast, if an integrated perspective on organisational boundaries is adopted, thereby including the knowledge or transaction effects, then a better understanding of the role of the economies of scale may be gained. At the same time, the integrated perspective may also provide a solution to the Cournot-Marshall dilemma, offering an explanation which reconciles internal economies of scale

---

\(^8\) On this see Pratten (1988) and Martin (2003), who provide very useful surveys on empirical findings concerning economies of scale in various industries.

\(^9\) Economies of productive capacity balancing derive from the possibility that an increase in the whole process could bring about more efficient utilisation of the productive capacities of some inputs. Where there are indivisible and complementary inputs, bottlenecks in productive capacity may arise. These bottlenecks often lead to idle times and waste of the productive capacities of some devices or machines. Therefore, a high volume of production is often necessary to render their different productive capacities compatible.

\(^10\) On the analysis of these sources of economies of scale, see Morroni (2006, pp. 166-177).

\(^11\) Economies of this kind are available only during the growth of the firm and therefore disappear when the expansion has been completed (Penrose, 1959, pp. 99-101).
with competition and clarifies why internal economies do not necessarily lead to a monopoly.\(^\text{12}\)

If productive and transactional knowledge are not full and free, the development of capabilities and of transactional relations may boost the presence of a greater number of producers and may impede the tendency towards monopoly that is held to spring from significant internal economies of scale. Exploitation of potential economies of scale is by no means instantaneous, but it is instead the fruit of an evolutionary path moulded by a learning process through which techniques are developed and adopted. The crucial point is that during the time span in which this growth process takes place, the possibility of increasing the size of the firm in order to reap economies of scale cannot be taken for granted, but depends on various factors, in particular:

a. The growth of the firm requires the ability on the part of the firm to develop adequate capabilities and to arrange transactional relations.

b. Heterogeneity in customer preferences may entail the presence of diverse types of firms.

c. Innovative activity and learning processes may need a variety of technical and organisational structures.

d. Some internal weaknesses and external counteracting forces may arise, thus hampering the growth of the firm and contrasting the tendency towards monopoly inherent in the presence of significant economies of scale.

Let us examine each of these four factors.

a) The development of capabilities

The process of growth of the firm takes time because the firm has to develop the necessary capabilities both as a means to solve “problems of scaling up the processes of production” (Chandler, 1992, p. 84) and to create a demand for its commodities. For each given scale dimension achieved and technique that can be chosen by the production unit, there are different stages of the development of abilities that facilitate the use of specific machines and equipment. Moreover, the growth of the firm requires expansion of its market share within a given market or creation of a new demand, thereby enlarging the market extension. Often the opportunity to exploit potential

\(^{12}\) Note that the solution presented here revives, in a different analytical context, some of Marshall’s insights on the firm’s particular market, and also builds on the relationship between organisation and knowledge that could not find place within the Marshallian *vulgata* elaborated by Pigou and others, which forms the basis of the partial equilibrium analysis illustrated in all microenomics textbooks.
economies of scale is boosted by the success of a specific product, which is linked to the capacity to create a competitive advantage by exploiting technological opportunities in complementary commodities, matching potential demand and changes in consumer tastes and habits.\textsuperscript{13} When information asymmetries are present, the development of an integrated set of dynamic capabilities through learning becomes an important basis for competitive advantage and therefore “constitutes the foundation for continuing growth” (Chandler, 2003: chapter 1, pp. 7, 15). Dynamic capabilities consist in the firm’s ability to integrate, build and reconfigure internal and external knowledge in order to address rapidly changing environments.\textsuperscript{14} The formation of new capabilities is made possible by developing or tracking down new abilities and skills which are generated by the ability to utilise outside knowledge (absorptive capacity) and by intensive outside-inside interaction. Coad and Rao (2007) and Holzl and Friesenbichler (2007) provide evidence that learning processes – linked with investments in innovative activity and in product diversification - are of great importance for high-growth firms.

\textbf{b) Heterogeneity in customer preferences}

Heterogeneity in customer preferences - concerning output qualities, learning processes and sales assistance - may favour the presence of diverse types of firms within the same sector of activity.\textsuperscript{15} The coexistence of several firms of contrasting sizes arises from the specific benefits inherent in the small scale of production, which allows a rather different relationship between manufacturers and customers in terms of contractual advantages, flexibility and learning opportunities.

The product of a large mass-production firm is very different from the analogous product supplied by a craftsman or by large-scale industrial flexible production. Thus in spite of the significant economies of scale that characterise mass production, so highly prized is individual crafting that in many activities handicrafts have never been completely supplanted by cheaper industrial production, having instead survived alongside it. Analogously, just as traditional artisan production was not ousted by the rise of mass production in the twentieth century, it seems likely that mass production

\textsuperscript{13} See Rosenberg (1969, pp. 111-2); Bianchi (1998, pp. 9-11).

\textsuperscript{14} Teece, Pisano and Shuen (1997, p. 204); Loasby (1998, p. 176). Knowledge-based considerations have recently been highlighted by Niman (2004, pp. 282-3) as a possible solution of Cournot’s dilemma.

\textsuperscript{15} On the role of heterogeneity of demand and its effect on the industrial structure, see Bonaccorsi and Giuri (2003, pp. 59, 75ff.).
itself will not be doomed as a result of flexible industrial production (on a large or small scale): rather, the different forms meet quite different needs.

c) **Innovative activity**

Innovative activity and learning processes may need heterogeneous abilities and a high degree of variety in technical and organisational structures. Innovative activities are favoured by the presence within the market of a myriad of small organisations that can generate diverse options. “When uncertainties are very high, as is inevitably the case with new technologies”, the market place encourages exploration “along a wide variety of alternative paths”. Even if new business initiatives grounded on basic research are generally carried out by large firms, it should not be overlooked that, in some new technologies, start-up firms have a prominent role in the development of innovations. To understand why this is so, it should be noted that creative and innovative activities can be favoured by the absence of rigid, authoritarian and hierarchical relationships, a circumstance that enhances diversity of options and tolerates variety. “In exploring unknown territory to the goal of better technologies”, “multiple sources of decision making” formed by numerous small business organisations are essential. This explains why innovations are often associated with multiple organisations or new entries to a field.

Innovative activity broadens variety. Innovations are produced because firms deliberately seek a way of differentiating themselves from rivals and adapting to a changing external environment. Hence, variety derives from the purposeful ability to introduce a genuine new idea. Purposefulness plays a crucial role in the selection processes that take place in a social context. Variety and diversity constitute a general condition both for the growth of knowledge and for profit.

There is thus a two-way relationship between innovative activity and heterogeneous abilities. Innovative activity may create asymmetric information and

---


heterogeneous abilities possessed by individuals. On the other hand, heterogeneous abilities explain why individuals may have a different propensity or ability to innovate.

In some cases, new technology allows a high degree of flexibility in large-scale production, while in other cases it encourages the economic potential of small firms or production units. This results in the presence of a variety of technical and organisational structures in different industries. Moreover, the “locality of learning” and the “opaqueness” of the environment imply the persistence of different organisational forms and size among firms, and the coexistence of diverse hybrid arrangements operating in the same sector of activity (Coriat and Dosi, 1998, p. 112; and Ménard, 2004, pp. 369-70).

d) Internal weaknesses and external counteracting forces

The exploitation of potential economies of scale through continuous expansion of the organisational boundaries may encounter setbacks because of internal weaknesses and external difficulties.

As far as internal weaknesses are concerned, firms may be affected by misperception and errors. Unclear allocation of rights and responsibilities tends to make decision-making mechanisms inefficient. Furthermore, as the firm grows, top management may lose touch with events affecting day-to-day decisions. This can cause problems in vertical communication within organisations, leading to decreased efficiency and efficacy, and resulting at times in quid pro quos between managers and subordinates. In addition, difficulty in establishing rewards and focusing incentives may pave the way to shirking, high coordination, monitoring and control costs, influence activities, collusion, subverted inspections, parochial interests and unimplemented decisions. If increasing coordination and control difficulties are encountered due to mounting organisational complexity, then it may prove necessary for a rising share of the labour force to be dedicated to administration, engendering growing burdens of bureaucracy as a firm grows larger.  

Finally, the firm’s growth process is generally characterised by changes that typically involve transformations in the organisational set-up and the power structure, but this may meet with resistance on the part of some members of the firm who are unwilling to accept or strongly oppose any modification in the existing equilibria. For instance, the owner of the physical

---

assets or the top managers may have no interest in the growth of the firm if this would imply the development of capabilities beyond their control and hence an alteration of power relationships.

The management may make erroneous decisions because of various forms of cognitive inertia and myopia. Ignoring some variables and applying routinised operating procedures greatly simplifies decision-making, but cannot avert the risk that the routines adopted could turn out to be inefficient because their domain of validity may have been overstepped. Erroneous decisions may be due to the tendency to adopt routinised operating procedures that have proved efficacious in the past, extending them “beyond their original domain” in which they had been tested.\textsuperscript{21} Temporal myopia implies the tendency to privilege the short run and to disregard the distant future. Spatial myopia consists in ignoring changes in environmental conditions. Failure myopia involves neglecting the possibility of failure and overestimating the possibility of success. (Levinthal and March, 1993, p. 95ff.).

With regard to external counteracting forces, full exploitation of economies of scale of a individual firm may be impeded if the increase in the market share of an individual firm is slowed down by the competition of incumbent firms or new entrants, who successfully develop new capabilities and are adept in arranging transactional relations. During the process of the firm’s expansion, competitors may develop different capabilities and exploit potential economies of scale as well. Incumbent competitors may establish transactional relations with costumers and such relations may become rather stable over time because of acquaintance relationships, the existence of distribution networks, fidelity to a certain trade mark, consumption habits, or lack of information regarding the value for price of new products.

\textit{The coexistence of different sizes and organisational structures}

The upshot of this competitive process is that the management’s ability to seize opportunities for growth and to thwart countervailing forces is characterised by an unavoidable subjective dimension which moulds the firm’s evolutionary path and yields a large variety of outcomes.\textsuperscript{22} This evolutionary path takes shape by means of


\textsuperscript{22}On the logical impossibility of determining the \textit{optimum size} of the firm, see Georgescu-Roegen (1964, p. 296); Morroni (1992, pp. 141-2); Penrose (1995, p. xii); Foss (2002, p. 153).
adjustments that occur day by day and which are cumulative, partially irreversible and specific to one business organisation. The firm’s revealed performance, size and boundaries must be regarded as “time- and path-dependent” (non-path-determined) phenomena constantly influenced not only by the various basic conditions but also by the specific manner in which the different stakeholders’ interests are weighed in decisions. Different behaviour and strategies are determinants of firms’ differential in revealed performances. For these reasons, distinct organisational structures can live side by side. Indeed, in industrial countries, firms exhibit a very wide range of possible property and financial structures, hierarchical set-ups, incentive and control structures, size and market power, arrangement of production processes, degree of vertical integration and organisational features. This generates a huge variety of sizes of firm even in the same sector of activity. Size ranges from a very small business to very large companies with hundreds of thousands of employees that give rise to a turnover constituting a significant percentage of their country’s Gross National Product (GNP). The enormous distinctions in the size of firms yield remarkable disparities in market power.

The 1990s saw the emergence of contrasting tendencies with regard to integration and concentration processes. High levels of merger and acquisition activity coexisted throughout the decade with a trend in the opposite direction towards contracting out. The growth of giant firms has been based on horizontal expansion and diversification, and increasing concentration has been balanced by a tendency toward outsourcing and by new entries.

An example

To conclude this paper let us consider the example of aircraft production to demonstrate that economies of scale considerations are important but not sufficient to explain the size of the firm and the market structure. In the aircraft industry both engine suppliers and airframe producers enjoy significant economies of scale in manufacturing, R&D and marketing activities, originating from indivisibility and

---

23 Basic conditions concern: the characteristics of information and knowledge, the characteristics of techniques and equipment, individual motivations and aims, individual abilities, the degree of uncertainty, structural change, institutional and market conditions. On these conditions, which influence the behaviour and performance of the firms, see Morroni (2006: chapter 1).

24 On this see Holmström and Roberts (1998, pp. 73, 80, 83-9).

25 This example is drawn from Morroni (2006, pp. 253-4).
complementarity effects at the human capital and technical infrastructure level. Yet in spite of the fact that both sectors share the same conditions for economies of scale, they exhibit very different market structures. The production of commercial jet engines is characterised by “intense competition, a low level of concentration and strong instability of market shares”, while airframe production has witnessed “a tremendous process of concentration” (Bonaccorsi and Giuri, 2003, pp. 52, 74). Therefore, the effect of economies of scale on market structure appears to be unimportant in the jet engine sector, but significant in the airframe sector.

Why is it that the presence of internal economies of scale has not led to a strong concentration or a monopoly in jet engine production? On the mere basis of scale considerations we cannot answer this question. In other words, by focusing exclusively on the phenomenon of economies of scale, it is impossible to pinpoint factors in this sector that prevent the occurrence of a shakeout pattern in which a large incumbent would be in a position to exploit economies of scale and thus to dominate the industry. To account for this divergent pattern between engine suppliers and aircraft manufacturers, the analysis must be broadened in order to include aspects linked to the development of technological knowledge and to transactions that take place within vertical relations among producers. The low concentration level in the production of jet engines is explained in terms of the multiple-sourcing strategy adopted by large aircraft manufacturers, who have an interest in entering into transactional relations with many engine suppliers so as to exploit different technical capabilities. In this case, supplier heterogeneity enhances the learning processes that strengthen the competitive advantage of the aircraft manufacturer, because each customer relation is a source of specific learning. Aircraft manufacturers “do not just demand products, but rather co-develop them” with suppliers, by looking for new solutions that integrate engine and airframe (Bonaccorsi and Giuri, 2003, pp. 79, 84).

Hence, this example indicates that taking into account not only internal economies of scale but also knowledge- and transaction-based considerations may offer a solution to the Cournot-Marshall dilemma regarding the incompatibility between competition and internal economies of scale.
References


---

26 Pages quoted in the text always refer to the original edition unless the reprint or the translation date is indicated in the references between square brackets.


Martin, S. (2003), “Globalization and the natural limits of competition”, preliminary version, March, Department of Economics, Krannert School of Management, Purdue University, Purdue, Indiana


