THE RESEARCH INSTITUTE OF INDUSTRIAL ECONOMICS

Working Paper No. 556, 2001

Merger Control and Enterprise Competitiveness - Empirical Analysis and Policy Recommendations

by Johan Stennek and Frank Verboven

IUI, The Research Institute of Industrial Economics
P.O. Box 5501
SE-114 85 Stockholm
Sweden

Merger Control and Enterprise Competitiveness

- Empirical Analysis and Policy Recommendations -

3 May 2001

Johan Stennek

Research Institute for Industrial Economics, Stockholm and CEPR

Frank Verboven

Universiteit Antwerpen and CEPR

Contents:

1. EXECUTIVE SUMMARY

This report studies the importance of efficiency gains from horizontal mergers. A general theme throughout this report is that efficiency gains, and their pass-on to consumers, may vary substantially from merger to merger. For this reason it seems appropriate to reconsider current practice in European merger control, which does not allow the merging parties to appeal to an efficiency defence. Our report provides a detailed examination of two main parts of an efficiency analysis. The first chapter considers the presence of efficiencies from mergers, with a focus on economies of scale. The second chapter consider the pass-on of efficiencies to consumers in the form of lower prices.

1.1 Economies of scale

Cost savings, or other efficiency gains, are necessary for horizontal mergers to benefit consumers and the economy as a whole. Cost is also important way to increase enterprise competitiveness. One of the most important sources of cost savings is economies of scale.

The aim of chapter 2 is to investigate to what extent mergers are likely to generate cost savings through better exploitation of scale economies. This information is an important input to the discussion about the pros and cons of introducing an efficiency defence in the Merger Regulation. For this purpose we review the theoretical and empirical literature on scale economies. Although this literature contains information that is useful for the appropriate design of merger control, many issues are unsolved. A second purpose of this review is therefore to carefully point out where the existing literature is still insufficient.

We begin our analysis with an intuitive discussion of the sources of scale economies (section 2.2.1). A production process is said to have scale economies if the unit cost of production is lower when more is produced. There are several sources of scale economies, such as mechanisation and specialisation achieved through division of labour. Without constructing a full typology, we compile a “checklist” of the most common sources of scale economies.

We also draw a sharp distinction between technologically determined economies of scale and overheads such as advertising and R&D. Due to state of economic knowledge, we are forced to primarily focus this overview on the first type of scale economies. However, we also argue that this state of knowledge is very unsatisfactory. Advertising and R&D are important in most European industries with a high concentration.

Next, we define the term scale economies more rigorously (section 2.2.2). This section is an important input for discussing the relationship between mergers and production costs, and also for understanding the empirical literature. In the economic literature there exists two different definitions of scale economies, one based on the so-called production function, the other based on the so-called cost function. We argue that for the purpose of analysing efficiency gains from mergers, it is vital to use the concept based on the cost function. The reason is that the exploitation of some scale economies (e.g. mechanisation) requires changing the factor proportions. Such gains are only captured by the concept based on the cost function. The primary complications in defining scale economies arises when firms produce more than one product. In this case, there may exist overall diseconomies of scale even if there are (product specific) economies of scale for all individual products. The reason is that there may exist so-called diseconomies of scope. Another complication is that the minimum efficient scale of a firm depends on the mix in witch the firm produces the different products.

We then discuss the theoretical relation between mergers and scale economies (section 2.2.3). That is, if production is characterised by scale economies, under what conditions can they be exploited by means of merger? In part, this question was analysed by Farrell and Shapiro (1990).[1] However, a limitation in their analysis is that they focus on the case when the capital stock in each plant is fixed—the short run. Drawing on the theory of sub-additivity, we extend their analysis to include the effect of mergers on cost in the long run—when firms can adjust their capital stock freely.

In the single product case, the relation is rather straightforward. Economies of scale (in the relevant range) are sufficient, but not necessary, for a merger to reduce cost. In the multiple-product case, the primary finding is that that the relation between scale economies and mergers is rather complex. Even if there are unexploited scale economies, mergers may well raise rather than reduce cost. This implies that even if firms can demonstrate the existence of scale economies (in the relevant range) to competition authorities, allowing a merger does not necessarily lead to cost savings. A positive finding is that one can establish different sets of sufficient conditions for mergers to reduce cost. These conditions are based on the notion of sub-additivity of the cost function. For example, if there are product specific economies of scale for all products, and there are economies of scope, then a merger reduces cost. In effect, these theoretical results show that in evaluating the effect of a merger on cost, competition authorities need to take into account both product specific economies of scale and potential economies of scope. However, in order to make these conditions more precise additional analysis, similar to that undertaken by Farrell and Shapiro for the short-run case, is needed.

Using the concept of sub-additivity, we are also able to say if a full merger is the best way to restructure production in order to produce the given output as cheaply as possible. This is important since there are often several alternatives, short of a full merger, that may reduce cost without imposing all the anti-competitive effects of a horizontal merger. Antitrust authorities are often required to make such a broad evaluation of mergers. Only cost savings that are merger specific should qualify as an efficiency defence.

Farrell's and Shapiro's short-run analysis and the long-run analysis undertaken in this report are complements. The first analysis considers the extreme case when the capital stock in every plant is fixed. The second analysis considers the other extreme case when firms freely can adjust their capital. In the long run, cost savings are larger than in the short run. Hence, the latter form of analysis is more likely to produce a positive picture of the effects of the merger. As a consequence, antitrust authorities may reach

[1] Their analysis is described in Röller, Stennek and Verboven (1999).

different decisions, depending on whether they consider the long run or the short run effects. To complicate the picture further, it is likely that capital adjustments require different amounts of time in different industries. Therefore, if the antitrust authority evaluates the effects of the merger within the first two-year period, long-run analysis may be relevant for some industries and short-run analysis may be relevant for other industries.[2] In sum:

> **Conclusion:** *Theoretical analysis can establish the conditions under which mergers generate cost savings, related to scale economies, in the long run. Such conditions may also include a test for merger-specificity. To make these conditions fully operational for antitrust authorities, more research is needed however.*

In section 2.3, we review the empirical literature on economies of scale. We both provide a brief overview of empirical estimates of scale economies in several industries, and discuss the methodologies that are used. However, since many of these studies are rather old, and since scale economies in many industries are likely to change over time, as a result of technological development, we do not attempt to provide a detailed account of the measurement of scale economies for all industries. Rather, the main purpose is to extract the general conclusions, from this literature, that are relevant for the discussion of an efficiency defence in merger control.

The primary finding is that the exploitation of scale economies is a source of substantial cost savings in some, but not all, industries. Economies of scale are often present at low volumes of output, and exhausted at larger volumes. In our view, the empirical results have an important implication for merger policy. The variability of scale economies across industries and different output volumes support the idea that cost savings should be evaluated on a case-by-case basis, i.e. an efficiency defence. Moreover, the empirical literature shows that concentration in many industries has gone further than motivated by scale economies alone. Thus, in evaluating mergers it is important that only scale economies in the relevant range should be used as a defence for a horizontal merger. In sum:

> **Conclusion:** *The variability of scale economies across industries supports the idea that cost savings should be evaluated on a case-by-case basis. Only scale economies in the relevant range should be used as an efficiency defence.*

In order for antitrust authorities to take cost savings related to scale economies into account, they need some methodology to assess the importance of these economies. Therefore, we also provide a short introduction to the empirical methodologies that have been used to estimate economies of scale. Essentially, three different methods have been used to study the relationship between size and unit cost. The econometric approach uses statistical techniques to analyse cost-output data from some collection of firms. The mathematical programming approach is similar in that it also analyses cost-output data, but using linear programming. The engineering approach relies on interviews with engineers who design new production units.

[2] Unfortunately, the empirical literature does not report any measures of how long time that is required in order to achieve scale economies in different industries. Another important issue that is unexplored in the literature is the possible differences in time required for achieving scale economies by means of merger and by means of internal growth.

In our view, all of these methodologies are useful for antitrust authorities, albeit in different ways. The already existing studies based on the econometric and mathematical programming approaches have produced useful “background knowledge.” Whenever a case has to be decided concerning some industry, these studies can give a first hint if scale economies are important in that particular industry. What is the order of magnitude of potential cost savings? Is it reasonable to carry out a more detailed study?

However, for several reasons the econometric approach is probably not suitable for the antitrust authorities to carry out such more detailed investigation. The main reason is data limitation. In order to estimate scale economies, cost and output data also from firms not participating in the merger is necessary. Another reason is that the econometric approach does not identify the sources of scale economies, which is important in order for assessing merger specificity. Estimates of scale economies do not indicate the time it takes for a merger to achieve cost savings related to scale economies. In contrast, the engineering approach should lend itself for such more detailed studies in particular cases. In particular, the engineering approach may supply an “incremental analysis,” that is focus on the factors that will be affected by the merger. In sum:

> **Conclusion:** *Antitrust authorities can use the following two step procedure. First, use existing econometric evidence on scale economies, to indicate if the order of magnitude of potential cost savings motivate a more detailed study, in particular merger cases. Second, if detailed investigation is warranted, studies based on the engineering approach can be commissioned.*

1.2 Pass-on of cost savings into consumer prices

The second chapter analyses the pass-on of cost savings on consumer prices. Price effects are important for two related but distinct reasons. First, assessing price effects is essential for evaluating changes in consumer welfare, which is often the first or main objective of merger control. Second, understanding the price effects is essential to evaluate changes in the merging firms’ competitive position. Indeed, if a merger leads to a reduction in price, the merging firms typically realize an increase in their market share. In contrast, if a merger leads to a price increase, the merging firms are likely to experience a shrinking market share. While the merging firms’ profits may still increase in this case, this would be caused by increased market power and thus not by an increased competitiveness.

The effect of a merger on price may be decomposed into two separate components. First, there is the price increase stemming from an increase in (unilateral or collusive) market power, holding the merging firms’ costs constant. Most theories of oligopoly predict such a price increase. Second, there is a possible price reduction stemming from the cost savings due to the merger. Cost savings may be passed on to consumers if it is the marginal cost that is reduced. As we explain in detail in the report, also some conditions on demand must be satisfied, in order for cost savings to be passed on to consumers.

The total price effect from the merger is negative if pass-on of marginal cost savings is sufficiently high to outweigh the market power effects.[3] Only in the extreme case of no pass-on a merger will necessarily increase price no matter how large the marginal cost savings are. More generally, the greater the pass-on, the more likely a merger is going to decrease price in the presence of cost savings. For example, if 50 percent of the cost savings are passed on to consumers, then a merger decreases price if the realized cost savings are at least twice the percentage amount of the price effects from increased market power. If cost savings are fully passed on to consumers, then a merger decreases price if the realized cost savings are at least the same percentage amount as the price effects from increased market power.

These examples illustrate the importance of understanding the degree of pass-on to evaluate the full price effects from mergers. Since pass on is not a constant, but varies between markets and over time, several questions must be addressed. What are the relevant determinants of pass-on? Do technological conditions affect the degree of pass-on? What is the role played by competition? How important is the distinction between pass-on of firm-specific and industry-wide cost savings? What is the empirical evidence regarding pass-on? We take up these questions in this chapter with the purpose of obtaining a better understanding into the role of efficiencies in mergers.

In a first section we review the theoretical determinants of pass-on of cost savings. We begin by reviewing pass-on of industry wide cost savings, i.e. cost savings realized by all firms in the industry. Under perfect competition, the price elasticity of demand and the shape of the marginal cost curve are the main determinants. As consumers become more price inelastic, and as the supply curve becomes more elastic (little capacity constraints) pass-on of industry wide cost savings will be more complete. If supply is perfectly elastic (no capacity constraints), pass-on of industry-wide cost savings is complete under perfect competition. The picture is different under monopoly or oligopoly. In this case, pass-on of industry wide cost savings may be incomplete even if supply is perfectly elastic. The reason is that firms have market power and charge markups which depend on the price elasticity of consumer demand. In the “typical” case, consumers become more price sensitive as the price increases. If this is the case, then pass-on will be incomplete. This is especially true in the monopoly case. Nevertheless, it is worth stressing that even in the monopoly case (or the dominant firm case) firms pass on at least part of the cost savings onto consumers. As the number of firms increases, pass-on of industry wide cost savings generally becomes more complete.

Next we emphasize the crucial importance of distinguishing between industry-wide and firm-specific pass-on. An understanding of firm-specific pass-on is especially relevant in the context of mergers, since it are typically only the merging firms who benefit from cost reductions. Nevertheless, both policy makers and the merging parties often confuse firm-specific pass-on with industry-wide pass-on. Our analysis shows that firm-specific pass-on is generally less than industry wide pass-on. For example, when firms are identical, firm-specific pass-on is equal to industry-wide pass-on divided by the number of firms in the industry. Hence the more firms there

[3] Assume that the merger tends to increase price by dP/P percent as a result of the first effect. Then, for the net effect to be non-positive, it is necessary that $(dc/c)(dP/dc)(c/p)\leq -dP/P$, where (dc/c) is the percentage cost reduction, and $\beta=(dP/dc)(c/p)$ is the pass-on elasticity. The pass-on elasticity indicates by how many percent the price would be reduced if cost is reduced by one percent.

are, the lower is the degree of firm-specific pass-on. When firms are not identical, no general results on firm-specific pass-on are available. Nevertheless, it is clear that the market share of the firm who realizes the cost reduction matters. On the one hand, a high market share of the cost-reducing firm makes the cost saving close to an industry-wide cost saving and thus makes pass-on more complete. On the other hand, a high market share also means a lot of market power, which may provide an incentive to pass on incompletely. The results of these forces is that pass-on in an industry is typically the largest for firms with intermediate market share, i.e. not the very small firms but also not the dominant firms.

> **Conclusion**: *For merger analysis it is important to concentrate on pass-on of firm-specific rather than industry-wide cost savings. The extent of pass-on depends on several factors, including the presence of capacity constraints, the price elasticity of demand, the degree of market power, and the market share of the merging firms.*

In the second section of this chapter we review the empirical literature on pass-on. We start by reviewing the rich empirical literature on tax incidence, intermediate goods price transmission and exchange rate pass-through. The empirical results vary sometimes substantially from sector to sector. Yet it is still possible to make some empirical generalizations. It seems fair to say that the literature on the effects of excise taxes and intermediate goods prices finds that pass-on is close to 100 percent, at least when one considers a sufficiently large time horizon (10 weeks or more). The literature on exchange rate pass-through tends to find incomplete pass-on, of an order of magnitude of 60-70 percent. Part of this literature also relates the extent of exchange rate pass-though to the market share of the exporting sector. Finally, the scarce literature on firm-specific pass-on finds a relatively low degree of pass-on, especially when the market share of the firms is small. The estimates are in the range of 10-20 percent.

This evidence makes clear the empirical importance of considering firm-specific rather than industry-wide pass-on. While industry-wide pass-on is more or less complete, firm-specific pass-on may be substantially smaller. At the same time, however, it is not sensible to draw strong general conclusions about the extent of firm-specific pass-on. The extent of firm-specific pass-on is likely to vary substantially from merger to merger, and it seems unreasonable to aim for a general presumption on the extent of firm-specific pass-on. In particular, there is a central role for market share in explaining firm-specific pass-on. Low firm-specific pass-on may be due to low market shares; yet in those cases the market power effects from mergers may also be low.

> **Conclusion:** *The empirical evidence is consistent with theoretical findings. Whereas pass-on of industry-wide cost savings is often found to be complete, pass-on of firm-specific cost savings is typically smaller. Nevertheless, the extent of firm-specific pass-on varies substantially from case to case, supporting the idea that cost savings, and their effects, should be evaluated on a case-by-case basis.*

Our empirical review does not only stress the importance of assessing pass-on on a case-by-case basis; it has also introduced several possible methodologies for assessing pass-on of cost savings, both reduced form and structural approaches. We hope our analysis makes it clear that the methodology lends itself quite well for a fast implementation, as is required in merger cases. This is especially true for the reduced form analysis, yet also structural approaches could be feasible in some cases. The major potential constraint to assess pass-on in merger cases is thus not necessarily in the application of sophisticated technical analysis. Rather, it is important to make sure that the data required for the analysis are collected from the parties in an efficient way.

> **Conclusion:** *There exist various practical methods for empirically evaluating pass-on of cost savings from mergers. This suggests the feasibility of a case-by-case approach to studying pass. The critical task is to collect data efficiently.*

We illustrate the feasibility of the approach with an extended example in Chapter 4.

2. ECONOMIES OF SCALE

2.1 The Importance of Scale Economies

For a horizontal merger to benefit consumers and the economy as a whole it must reduce the firms' costs or generate some other form of efficiency gain. Cost savings are also one of many important determinants of enterprise competitiveness. One of the most important sources of cost savings is economies of scale.

The aim of this chapter is to investigate to what extent mergers are likely to generate cost savings through better exploitation of scale economies. This information is an important input to the discussion about the pros and cons of introducing an efficiency defence in the Merger Regulation. For this purpose we review the theoretical and empirical literature on scale economies. As will be demonstrated below, this literature contains information that is useful for the appropriate design of merger control. As always, however, there are more questions than answers. A second purpose of this review is therefore to carefully point out where the existing literature is still insufficient.

The first part is theoretical (section 2.2). We begin our analysis with an intuitive discussion of the sources of scale economies (section 2.2.1). Without constructing a full typology, we compile a "checklist" of the most common sources of scale economies, such as mechanisation and specialisation achieved through division of labour. We also draw a sharp distinction between technologically determined economies of scale and overheads such as advertising and R&D. Next, we define the term scale economies more rigorously (section 2.2.2). This section is an important input for discussing the relationship between mergers and production costs, and also for presenting the empirical literature. The primary complication in defining scale economies arises when firms produce more than one product.

We then discuss the theoretical relation between mergers and scale economies (section 2.2.3). That is, if production is characterised by scale economies, under what conditions can they be exploited by means of merger? Drawing on the theory of sub-additivity, we analyse the effect of mergers on cost in the long run—when firms can adjust their capital stock freely. Our main conclusion is that theoretical analysis can establish the conditions under which mergers generate cost savings, related to scale economies, in the long run. Such conditions may also include a test for merger-specificity. To make these conditions operational for antitrust authorities, more research is needed, however.

In section 2.3, we review the empirical literature on economies of scale. We both provide a brief overview of empirical estimates of scale economies in several industries, and discuss the methodologies that are used. However, since many of these studies are rather old, and since scale economies in many industries are likely to change over time, as a result of technological development, we do not attempt to provide a detailed account of the measurement of scale economies for all industries. Rather, the main purpose is to extract the general conclusions, from this literature, that are relevant for the discussion of an efficiency defence in merger control. Our main finding is that the variability of scale economies across industries supports the idea that cost savings should be evaluated on a case-by-case basis. We also argue that the methods used in the economic literature should be useful also in antitrust cases.

The summary and conclusions from the theoretical part is found in section 2.2.4, and for the empirical part in section 2.3.5.

2.2 Theory of Scale Economies and Mergers

2.2.1 Sources of Scale Economies

A production process has scale economies if the unit cost of production is lower when more is produced. Scale economies may be product-specific, that is associated with the volume of a single product produced and sold; plant-specific, that is associated with the total output (of possibly many different products) of an entire plant; and firm-specific (or multiple-plant) economies, associated with a firm's operations of several plants.

Economies of scale may also be attributed to the different activities performed by the firm, such as production, purchasing, marketing and R&D. For reasons to be explained below, it is convenient to treat economies of scale arising from advertising and R&D as different from other "production economies." Therefore production economies are discussed in section 2.2.1.1 and advertising and R&D are discussed in section 2.2.1.2.

2.2.1.1 Production

Economies of Scale. There are many different reasons for why unit costs may decrease with volume. We list and explain some of the most commonly mentioned examples. The presentation is not intended to be complete, and the list should not be taken as a typology of scale economies.

1. *Indivisibilities*. In some industries some *factors of production* are indivisible. The most well known example is the critical mass in nuclear power generation. At the critical mass, some amount of electricity is generated. Producing less than this amount of electricity means that the cost of the indivisible input is spread over fewer units of output, implying a higher cost per unit of electricity. There are also many examples of indivisible *tasks*. Starting up a production run requires that machines are fine-tuned for the process. If this set up cost is spread over many units, average cost is reduced. Some management and marketing tasks have to be done independent of the scale of operation. Again, the more units produced, the lower is the cost per unit.

2. *Specialization.* Division of labour and specialisation is an important source of scale economies. A production process may require 100 different tasks. If the firm has 5 employees each must perform 20 tasks on average. If the firm has 1000 employees, no individual will be required to perform more than one, and all may become more skilled at what they are doing—a specialist. Specialisation may be both "horizontal" into different production tasks and "vertical" into production and management tasks. Specialisation may be an important part of multiple-plant economies. That is, if production of different products is allocated to different

plants, specialisation can be driven further than if all products are produced within one plant.

3. *Mechanisation.* Producing at a larger scale may make it meaningful to let a machine do some tasks. However, at a low volume the machine may perhaps only be used part of the time, and not pay its cost (indivisibility). The larger the volumes the more specialised the tools can be used. Also parts can be transferred automatically between the different processing stages.

4. *Massed Reserves*. A manufacturer may need to keep one machine in reserve so that a breakdown does not interrupt production. Assume for example that a delivery failure of at most one percent is acceptable. Assume also that a machine produces ten units if it works. Then, having just one machine does not enable the firm to write any contracts that satisfy the failure threshold. However, with two machines it can contract ten units.[4]

5. *Increased Dimensions*. In many industries, the output of a processing unit tends to be roughly proportional to the volume of the unit, while the amount of materials and construction work is linked to the surface area of the unit's reaction chambers, storage tanks, connecting pipes, and so on. Since the area of a sphere varies as the two-thirds power of volume, there are considerable scale economies to be gained.

6. *Learning by doing*. Economies of scale have an important dynamic dimension. There is considerable empirical support for the so-called learning curve, which shows that unit cost falls with cumulative output.

Economies of Scope. Traditional economic analysis has concentrated on single-product firms. But, in reality, most businesses produce many products, and many antitrust issues involve only these enterprises. In recent years, economists and policymakers dealing with antitrust issues have increasingly recognised the need for a theory that can be used to evaluate the efficiency of market structures in industries dominated by a few firms operating in a diverse range of markets. For such firms, conventional concepts such as economies of scale do not adequately capture the complexity of the situation.[5] The new literature has introduced as a complement to the old concept of scale economies, the new concept of "economies of scope," which measures the cost advantages to firms of providing a large number of diversified products as against specialising in the production of a single output.

1. *Joint production.* Marshallian joint production exists if one or more factors of production are public input. That is, once acquired for use in producing one good, they are costlessly available for use in the production of others (up to some limit). For example, crude oil is a public input to producing petrol and asphalt. Similarly, certain tasks may be public tasks. For example, if the production of two outputs requires the same input, then the acquisition of these inputs may be cheaper if performed within the same firm. The reason for this may be both due to increased monopsony power but also real cost savings, e.g. related to improved quality control.

2. *Indivisibilities*. Indivisible input or tasks may not only result in economies of scale, but also in scope economies. Assembly line production with human labour

[4] See Mulligan (1983).
[5] Bailey and Friedlander (1982).

is often economical for single-product runs of large scale. Combinations of robots and computer-controlled machine tools bring the benefits of large "scale" to the production of even small batches, also when demand for every particular product variety is small. Here it is the robot's ability to switch form one task to another that counts. The robot production line has the flexibility to handle a variety of separate products, and makes it possible to take advantage of the economies associated with automation.

3. *Networking.* In the aviation industry, the cost of providing a seat is much lower on big planes than on small planes. This provides an incentive for airlines to organise the provision of services by offering a networking arrangement involving one-stop service. Individuals with a variety of destinations are boarded on a single larger plane. Planes from a variety of origins then stop at a central airport at approximately the same time. The passengers to each of the destinations are then combined, again making it possible to use larger planes. If one firm is more efficient in co-ordinating the scheduling of services to a variety of city-pair markets than several small firms, there are economies of scope.[6]

4. *Knowledge-sharing.* Multiple-product firms may be more cost efficient than combinations of single-product firms when several industries require similar know-how but there are transaction difficulties affecting the transfer that know-how (how can one sell information without disclosing it?).[7]

Economies of Vertical Integration. There are many reasons for why vertical integration or disintegration may affect the costs of production. For an overview of the economic literature on vertical integration, see Perry (1989). The aim of the present survey is to discuss efficiency gains associated with horizontal mergers. For such mergers, the degree of vertical integration remains constant. For this reason we will not discuss the economies and diseconomies associated with vertical integration. The reason for this limitation is that vertical mergers and horizontal mergers have very different effects on both competition and costs. This limitation obviously means that the conclusions drawn from this report are only valid for horizontal mergers.

Diseconomies of Scale. While economies of scale may be common in many industries, especially at low levels of volume, there are also good reasons to believe that for large firms the opposite can be true: a reduction of the scale of a large firm may lower unit cost. In that case there exists diseconomies of scale. What are the sources of such diseconomies?

1. *Management is a fixed factor*. Any firm must have some individual who assumes executive authority and responsibility. This individual has a limited time and cognitive capacity. At the same time, a large plant or firm is more difficult to manage than a small plant or firm. Thus, the larger the firm becomes, the more tasks must be delegated to lower level management, implying more loss of control.

2. *Transportation costs*. If a firm produces more it must sell more implying that it may be necessary to reach out to more distant customers. Thus, including transportation costs, there may be dis-economies to plant size. The importance of transportation costs as a check to scale economies depends on the technological

[6] See Bailey and Friedlander (1982).
[7] See Teece (1980).

level in transportation. The more efficient transportation is, the more profitable it is to exploit economies of scale.

3. *Worker alienation.* There exists psychological evidence that shows that workers become less satisfied with their jobs in larger plants.[8]

Since the sources of economies of scale become less important at a large scale, the sources of diseconomies of scale become especially relevant at large levels of production.[9]

Several of the above points reveal that a useful understanding of economies and diseconomies of scale and scope requires research going behind the neo-classical production function theory. Is knowledge sharing a valid argument for a conglomerate merger? To analyse that issue requires an analysis of the relative efficiency with which knowledge can be transferred between organisations and within organisations.

Unfortunately, there does not exist any research showing which effects that are most important in different industries. The reason for this shortcoming is that most economic research, both theoretical and empirical, views the technology as given. There are few studies trying to go behind the production function and explain its properties. (An exception is given by the theoretical studies deriving decreasing returns to scale as a result of loss of control.) Similarly, no research has been undertaken allowing us to identify which types of mergers that generate the different types of scale economies.

2.2.1.2 Advertising and R&D

In recent times, the economic literature has started to draw a sharp distinction between technologically determined economies of scale and overheads such as advertising and R&D. The above discussion of scale economies concerns the first type: technologically determined economies of scale. The important point about these scale economies, for example economies due to mechanisation, is that they are given by the firms' technological knowledge, and that they thus are fixed at any point in time. Merger policy may affect the extent to which such scale economies can be exploited by the firms. Merger policy does not, however, affect the existence of such scale economies.

There will also be economies of scale in advertising if some minimum amount of public exposure is needed before a sales campaign makes any impression at all on the public. The cost of this minimum amount of advertising is a fixed cost—the same, no matter how much the firm sells. Like any fixed cost, it will create a region of economies of scale in which the fixed advertising cost is spread over more and more units of output. Similarly, there may be a minimum amount of R&D that is needed before production can start. Again, spreading this fixed cost over a larger volume, reduces unit cost.[10]

A merger may eliminate the duplication of certain R&D (and advertising) activities and, the production of the different products may benefit from knowledge sharing.

---

[8] See Scherer and Ross (1990).

[9] See Teece (1980) for a discussion of less easily quantifiable sources of economies of scope, for example managerial expertise, a good financial rating and a sales staff.

[10] Another issue is that there may exist scale economies internally in these activities. There is some evidence that the average cost of advertising falls as its volume increases (Martin, 1993).

The crucial question is: should such elimination of duplicated R&D activities be counted as cost saving, and to be used as an efficiency defence?

The complicating factor is that advertising and R&D outlays are not fixed to the firms. Rather, they are choice variables to the firms. The firms actively use advertising and R&D outlays as competitive weapons. Their levels are determined in response to for example the size of the market and the intensity of price competition.[11] Thus, a merger will not only allow knowledge sharing. It will also affect the firms' incentives to undertake R&D activities and to produce new knowledge. Potentially, a merger may induce an anti-competitive reduction in the amount of R&D.

Unfortunately, however, the economic literature has only to a little degree studied the effect of mergers in industries where advertising and R&D are important. As a result it is not clear how one should treat the endogenous scale economies that are an inalienable aspect of such industries. On the one hand, one may argue that an efficiency defence should take into account the fact that a merger may allow advertising and R&D outlays to be spread over larger volumes. The argument is clear: lower unit cost is good for welfare. On the other hand, by doing so, merger policy will affect the firms' incentives to undertake R&D and to advertise their products. On this point, the arguments are not clear. We do not know exactly how competition policy will affect the firms' incentives for advertising and R&D, and we do not understand the welfare consequences of changing the firms' incentives in these respects.[12]

However, this state of the art is very unfortunate since it is the industries that are characterised by high advertising or R&D intensities that are highly concentrated in the EU. Thus, it is in these industries that horizontal mergers are more likely to raise competitive concerns. Among the 20 most concentrated (3-digit level) industries in the EU (industries where the five largest firms' aggregate market share is over 33.3%, e.g. optical instruments, computers and tobacco) 16 are characterised by either high advertising or R&D intensity. Among the 25 least concentrated (3-digit level) industries in the EU (industries where the five largest firms' aggregate market share is lower than 10%, e.g. wooden furniture and clothing) none is characterised by high advertising or R&D intensity.[13]

Nevertheless, due to the lack of economic knowledge, the following discussion must be primarily confined to the case of exogenous scale economies (in the same way as

[11] See Sutton (1991, 1998). To illustrate the crucial difference between endogenous and exogenous (sunk) costs, consider the different effects that these costs have on industry structure. The equilibrium level of concentration declines with the ratio of market size to technologically determined setup costs and rises with the toughness of price competition. Intuitively, the larger the market is, the more firms can cover the (exogenously given) setup costs, and thus the more firms are induced to enter. This relationship brakes down for endogenous setup costs. By incurring greater advertising or R&D costs a firm can raise the perceived quality of its products thereby enhancing the demand for these products. Competition between the firms involves an escalation of outlays by firms and so lead to higher sunk costs being incurred. Moreover, the larger the market, and thus the larger the potential profits are, the greater the incentive the firms have to channel demand towards their products. Thus, the larger the market, the higher are sunk costs. As a result, it is not possible to support more firms, the larger the market is.

[12] In the case of some forms of advertising there are additional problems. In particular, it is not clear how one should evaluate the effects of non-informative image building advertising on consumer welfare.

[13] See Davies and Lyons (1996, ch. 4).

any discussion of the anti-competitive effects of merger must be confined to price competition and leave out R&D competition).

2.2.2 Definition of Scale Economies[14]

A single-product technology is characterised by increasing returns to scale if unit cost is lower the larger is production. There are constant returns to scale if unit cost is unaffected by volume, and there are decreasing returns if unit cost is increased as production is increased.

This section presents the more rigorous definitions of the notion of economies of scale that are used in the economic literature. Rigour and formalism is necessary because the simple and intuitive idea of scale economies turns out to be rather complex in the context of multi-product firms. The more precise definition of returns to scale is also a necessary basis for discussing the relationship between returns to scale and mergers in section 2.2.3, and a necessary basis for the empirical studies reviewed in section 2.3.

To arrive at a useful definition of scale economies, that includes the various possible sources of scale economies discussed in section 2.2.1, it is convenient to start from the concept of the production technology, which is central to the economic theory of the firm. The production technology may be represented either by its production function or by its cost function (cf. duality theory). The production function describes which input-output combinations are technically feasible to the firm. The cost function describes the minimum cost for producing any possible output level (given the technology and factor prices).[15] There are several reasons for focusing on the cost function rather than the production function in defining economies of scale. The first reason is that it is the shape of the cost function that plays a key role in determining firm and industry structure. Second, in exploring the potential sources of increasing returns (section 2.2.1) it was recognised that the lower unit cost achieved at larger size is associated with organising production in a way that is different from how production can be organised in small plants or firms. An inevitable aspect of scale economies it therefore that factor proportions in large plants must be different from factor proportions in small plants. For example, increased mechanisation implies a higher capital/labour ratio. However, as will be discussed below, the definitions of scale economies based on the production function require factor proportions to be held constant. Such a restriction would eliminate much of the economies of large scale.[16] A third reason for focusing on the cost function is also motivated by empirical considerations.[17]

---

[14] This section is based on Panzar (1989).

[15] In the theoretical literature it is often emphasised that the production and cost functions represent a given state of technological knowledge. However, at least the first firm to organise production at some scale will necessarily have to adapt the technology to that scale. It may be difficult both theoretically and empirically to discriminate between innovations that lower the cost function from the adaptations necessary to achieve the lower unit cost at a higher scale. For a more detailed discussion of these problems, see Gold (1981).

[16] For a more detailed discussion of this point, see Gold (1981).

[17] Estimating a production function ignores the endogeneity of the input factors. A cost function gives a theory on how these input factors are determined, with the resulting "elegant instruments", being the input prices (being exogenously determined on a perfectly competitive input market).

2.2.2.1 Single-Product Firms

The production function. In order to define scale economies precisely we need to introduce some formal notation. Consider a firm producing a single output. The quantity of this output is denoted *q*. The firm uses *n* different inputs, for example different types of labour and different types of machinery. The quantity used of input *i* is denoted x_i, and the complete consumption of resources is described by the vector of inputs $x = (x_1,..., x_n)$. The technology is described by the so-called production function. We write

$$q = Q(x).$$

The production function should be interpreted as saying: if the firm uses input vector *x*, then it can produce at most *Q(x)*. Likewise, if the firm wants to produce *q*, then it needs to consume an input vector *x* such that *q = Q(x)*.

The cost function. If a firm wants to produce quantity *q*, there are two conditions that must be fulfilled in order for the firm to be cost-efficient. First, it should not use more inputs than is necessary to produce *q*. This amount of necessary inputs is given by the production function. Second, the firm should use different inputs in a proportion that balances their relative contribution to production to their relative prices. Let w_i denote the price of input x_i, and $w = (w_1,..., w_n)$ the vector of all *n* input prices. The cost-efficient quantity of input *i* depends on how much the firm is producing (that is *q*) and on the prevailing factor prices (that is *w*). Normally, the higher *q* is, or the lower the price w_i is, the more the firm will use of the input x_i. We write $X_i(q,w)$ to indicate the cost-efficient quantity of input *i* and that it depends on quantity and factor prices. Similarly *X(q,w)* denotes the whole vector of cost-efficient inputs. The total cost of a cost-efficient firm is given by the sum of expenditures for all efficiently chosen input, that is

$$c = C(q,w) = \sum_{i=1}^{n} w_i X_i(q,w).$$

where *c* is the production cost, and *C* is the cost function.

Economies of scale. In order to define scale economies, we first need to define two more basic notions, namely average cost and marginal cost. Average cost (or unit cost) is simply the cost per unit produced, that is

$$AC(q) = \frac{C(q)}{q}.$$

Marginal cost, which we denote by *MC(q)*, is the cost of the last unit. More formally, marginal cost is given by the derivative of the cost function. That is:

$$MC(q) = \frac{dC(q)}{dq}.$$

The definition of scale economies that we use in this report is based on the cost function. The intuitive idea that we want to capture is that there is increasing returns to scale if unit cost falls when output is increased. Similarly, there are decreasing returns to scale if unit cost is increasing when output is increased, and there are constant returns to scale if a change in output does not affect unit cost. Thus, economies of scale are determined by how unit cost is affected by a change in production. More formally, we say that there are increasing returns to scale if

$$\frac{dAC(q)}{dq} < 0,$$

and that there are decreasing returns to scale if the inequality is reversed. Note that this requirement is equivalent to marginal cost being lower than average cost, i.e.

$$MC(q) < AC(q).$$

The reason is simple. If the cost of producing an additional unit is lower than the average cost of previous units, then producing the additional unit lowers average cost.[18] If the first unit produced costs EUR 10 and the marginal cost of producing the a second unit is EUR 5, then average cost is reduced from EUR 10 to EUR 7.5 if production is increased from one to two units. Actually, the *degree* of scale economies is defined as the ratio between average and marginal costs, i.e.:

$$S(q,w) = \frac{AC(q,w)}{MC(q,w)}.$$

Returns to scale are said to be increasing, constant or decreasing as *S* is greater than, equal or less than one. Note that *S* depends on quantity (and input prices). In particular, it is common that average cost is falling at low output levels, then constant, and finally increasing at high output levels. In that case, *S* would be larger than one at small output levels, and smaller than one at high output levels.

Economies of scale II. To complicate matters, there is a second definition of scale economies that is often used in the literature. This definition is based on the production function. Given a proportional increase in all input levels, does output increase more or less than proportionally? If output increases more (less), there are increasing (decreasing) returns to scale. For example, if output more than doubles when all inputs are doubled, there are scale economies. Formally, we define the elasticity of scale:

$$E(q,x) = \sum_{i=1}^{n} \frac{dQ}{Q} \Big/ \frac{dx_i}{x_i}.$$

If $E(q,x) > 1$ production increases more than input, and there are increasing returns to scale. To complicate matters even further, the production function based measure $E(q,x)$ need not coincide with the cost function based measure $S(q,w)$. The reason is that it may not be cost efficient to increase all inputs by the same proportion to achieve the increase in production. Thus, even if per unit cost does not fall when output is increased by expanding all inputs proportionally, it may decrease when inputs are chosen in a cost-minimising manner. For example, it may be that doubling both capital and labour only leads to a doubling of output (i.e. constant returns to

[18] Formally this can be shown by taking the derivative of the average cost function, that is

$$\frac{d\left[\frac{C(q,w)}{q}\right]}{dq} = \frac{MC}{q} - \frac{C}{q^2} = \frac{MC - AC}{q}.$$

scale, as defined from the production function). However, it may be that the larger output level can also be produced using a more capital intensive technique. If the more capital intensive technique is less costly, cost is not doubled as output is doubled (i.e. increasing returns to scale, as defined from the cost function).[19]

Minimum efficient scale. We will sometimes assume that average cost is first decreasing and then increasing as the volume increases so that the average cost curve is "U-shaped." This corresponds to assuming that for low output levels there are increasing returns to scale and for large output levels there are decreasing returns to scale. Other times we will assume that the average cost curve is "L-shaped," so that there are only increasing and then constant returns to scale. The lowest quantity where average cost is minimised is called the minimum efficient scale (MES). The MES is an important determinant for industry and firm structure.

A potentially complicating factor is that the measurement of scale economies depends on input prices. Hence, even if the technological knowledge is the same in two countries, returns to scale and the minimum efficient scale may well differ between the two countries, if input prices differ. As a consequence, one can not take empirical estimates of MES from one country as a standard in another country without investigating the importance of factor price differentials.

2.2.2.2 Multiple-Product Firms

Most firms of reasonable size produce more than one product, and mergers normally create a firm that produces an even wider range of products. For this reason it is essential to define what is meant by economies of scale also for multiple-product firms. Unfortunately, however, the simplicity of the definition from the single firm case is lost. The reason is that in the multiple-product context, a firm's cost is not only determined by the scale of each single output, but also by the composition (or mix) of output. The concept of scale economies can be generalised to a multiple-product setting by considering two related measures, namely product-specific economies of scale and so-called ray economies of scale. However, to fully capture the effect of changes in the composition of output on cost, we also need to introduce the concepts of economies of scope and of so-called transray convexity.

To keep the formalism at a minimum we consider the case of a two-product technology. Let q_1 denote the quantity produced of good one and q_2 the quantity produced of the second good. The cost of producing the two products is given by $C(q_1, q_2)$ for any output combination $q = (q_1, q_2)$. Naturally, the cost also depends on factor prices *w*, but for simplicity we do not indicate this dependence explicitly. The cost of producing an additional unit of output (i.e. marginal cost) is denoted by $MC_1(q_1, q_2)$ and $MC_2(q_1, q_2)$ for products one and two respectively. It is customary to define the degree of scale economies as:

$$S(q_1, q_2) = \frac{C(q_1, q_2)}{q_1 MC_1(q_1, q_2) + q_2 MC_2(q_1, q_2)}.$$

[19] If one assumes that all inputs are chosen in a cost-minimising manner, that is according to *x=X(q,w)* then the two measures coinside, that is

$$E(q, X(q,w)) = S(q,w).$$

This definition is a natural extension of the definition in the single product case.[20] In the case of multi-product firms, however, the notion of scale economies is more complex than in the single-product case. In particular, overall scale economies $S(q,w)$ is constituted by three different parts, namely the product specific economies of scale for products one and two respectively, and the economies of scope. We will discuss product specific scale economies and economies of scope in turn.

Although the discussion in this report is confined to the case of two products, all definitions and results are valid also for the general case of *m* types of output. The discussion here is completely focused on the cost function. However, it is possible to define scale economies in the multiple-product case directly from the technology.

Product Specific Economies of Scale. The incremental cost of a single product, say q_1, is defined as the extra cost of producing q_1 given that q_2 is already produced, i.e. as $C_1(q_1,q_2) = C(q_1,q_2) - C(0,q_2)$. The average incremental cost of product one is defined as $AC_1(q_1,q_2) = C_1(q_1,q_2)/q_1$. The degree of scale economies specific to product one is defined as

$$S_1(q_1,q_2) = \frac{AC_1(q_1,q_2)}{MC_1(q_1,q_2)}.$$

Product specific returns to scale are said to be increasing (decreasing) if S_1 is larger (smaller) than one.

There exists a relationship between the overall degree of scale economies *S* and the degree of product specific scale economies S_i. This relation can be described by the following equation:

$$S(q_1,q_2) = \frac{[a_1 S_1(q_1,q_2) + a_2 S_2(q_1,q_2)]}{b} \qquad (1)$$

where

$$a_1 = \frac{MC_1}{MC_1 + MC_2}, a_2 = \frac{MC_2}{MC_1 + MC_2}$$

and

$$b = \frac{C_1(q_1,q_2) + C_2(q_1,q_2)}{C(q_1,q_2)}$$

Hence, the numerator in equation (1) is the weighted average of the product specific returns to scale. Thus, the more product specific economies of scope there are (large S_1 and S_2), the more overall scale economies there are (large S). The denominator is related to the so-called economies of scope.

Economies of Scope. In addition to the more familiar cost savings deriving from the scale of operations, cost savings may also result from the production of several

[20] In the single product case, the degree of scale economies can be expressed as:

$$S(q,w) = \frac{AC(q,w)}{MC(q,w)} = \frac{C(q,w)}{qMC(q,w)}.$$

different outputs in one firm rather than each being produced in its own specialised firm. A striking example of scope economies is that it is cheaper to produce different forms of oil derivatives in one firm than in several specialised firms. We say that there exists economies of scope if the cost of producing q_1 and q_2 is smaller in the case production of both goods is carried out in a single firm, rather than in two separate firms.[21] Formally, there are economies of scope if

$$C(q_1, q_2) < C(q_1, 0) + C(0, q_2).$$

The *degree* of economies of scope measures the percentage increase in cost that would result from dividing the production of q_1 and q_2 into two separate firms, rather than producing both goods in a single firm. Thus, the degree of economies of scope is defined as

$$SC(q_1, q_2) = \frac{C(q_1, 0) + C(0, q_2) - C(q_1, q_2)}{C(q_1, q_2)}.$$

For example, if the cost of production is fifty per cent larger in case the two products are produced separately, then SC = 0.5 (that is 50 per cent). If there are economies of scope, then $SC > 0$. If there are diseconomies of scope, then $SC < 0$. One can show that since all products have positive incremental cost, then $SC(q_1, q_2) < 1$.[22]

Using the definition of scope economies, overall scale economies (that is, equation (1)) can be written as

$$S(q_1, q_2) = \frac{a_1 S_1(q_1, q_2) + a_2 S_2(q_1, q_2)}{1 - SC(q_1, q_2)}. \qquad (1').$$

Equation (1') shows that the overall scale economies are determined by three factors. The product specific scale economies of products one and two, and the economies of scope. If the product specific economies of scale are strong, also overall scale economies tend to, but need not, be strong. Actually, even if there are product specific economies of scale, there may be overall diseconomies of scale. The reason is that there may be diseconomies of scope, and that the diseconomies of scope may dominate the product specific economies of scale. For example, if $S_1(q_1, q_2) = S_2(q_1, q_2) = 2$, there are overall diseconomies of scale if $SC(q_1, q_2) < -1$.[23]

Another concept to describe how cost depends on the composition of output is transray convexity. Roughly speaking, the cost function is transray convex if as a firm changes the composition of output while holding fixed the level of some aggregate

---

[22] To see this, first note that note that $SC(q_1, q_2) = [C(q_1, q_2) - C_1(q_1, q_2) - C_2(q_1, q_2)]/C(q_1, q_2)$. Using the definition of incremental cost, the degree of scale economies can be written

$$SC(q_1, q_2) = 1 - \frac{C_1(q_1, q_2) + C_2(q_1, q_2)}{C(q_1, q_2)}.$$

Since the last term is positive when incremental costs are positive, it follows that $SC(q_1, q_2) < 1$.

[23] One may derive conditions (weak cost complementarities) in terms of the properties of the multi-product cost function that can be used for empirically inferring the presence of economies of scope.

measure of output, costs will be lower for diverse rather than specialised output mixes. For example, transray convexity would imply that

$$\frac{1}{2}C(q_1,0)+\frac{1}{2}C(0,q_2)>C\left(\frac{q_1}{2},\frac{q_2}{2}\right).$$

Minimum Efficient Scale. In the multi-product case, also the idea of minimum efficient scale (MES) becomes more complex. To discuss MES, one defines the so-called ray-average cost, by

$$RAC(q_1,q_2)=\frac{C(q_1,q_2)}{b_1q_1+b_2q_2}.$$

The weights b_1 and b_2 are two completely arbitrary numbers. Their only role is to combine the two quantities into a single measure of size.

Although this construct is artificial, it surprisingly makes it possible to formally relate the slope of the *RAC* curve to the degree of scale economies. In particular, one may consider how *RAC* is affected by increased production, assuming that the production of the two products is increased by the same proportion (e.g. ten per cent). One may show that, if ray-average cost is reduced (increased) as production is increased in a proportional way, there are increasing (decreasing) returns to scale.

Using the notion of ray-average cost, one may talk about “U-shaped” average cost also in the case of multiple products. The ray-average cost is U-shaped if ray-average cost is first decreasing, then constant, and eventually increasing. The only complication in the multiple-product case is that the size of the output bundle at which scale economies are exhausted will tend to vary with the composition of the bundle. For example, it may be that the MES is small if the firm is producing relatively much of product one in comparison to product two, while MES is large if the firm is producing relatively much of product two. Thus, instead of a single point of minimum efficient scale, there will be “locus” of such points.

2.2.2.3 Multiple-Plant Firms

In the discussion above we did not distinguish between the production technology of a firm and the production technology of a single plant. However, since firms can and often do run more than one plant, the distinction is important.

Some scale economies can be achieved already in a single plant. Examples may be some economies related to specialisation and mechanisation. When the scale economies at the plant level are exhausted, for example due to worker alienation, it is better to divide production between two plants than to produce all output in a single plant. Another reason for dividing production between two plants is that having two plants located in different geographical markets can lower transportation costs. The fact that plant economies are exhausted does not imply, however, that firm size is optimal. The reason is that there may exist economies of scale associated with running more than one plant in each firm. For example purchasing of inputs may be better organised by one centralised department than to have one small department in each plant. However, also multi-plant economies are often exhausted above some level. The cost per unit will rise when the firm operates so many plants that corporate management starts to lose track of local operations.

The relation between the plant level cost function and the firm level cost function can be described in the following way. Assume that a firm has only one plant, and denote the cost function for this firm by $C(q,1)$ where 1 indicates that the firm has only one plant. This cost function (which may also be called plant level cost function) indicates the lowest cost of producing q units of output in a single plant. Similarly $C(q,2)$ indicates the lowest cost of producing q units in a two-plant firm. This cost function presupposes that the production is allocated in a cost-minimising way between the two plants. It also presumes that the firm has chosen to organise the purchasing function (joint or separate) in the most cost efficient way. The cost function at the firm level is given by

$$C^{Firm}(q) = \min_n \{C(q,1), C(q,2), ..., C(q,n), ...\}$$

Thus, the firm level cost function is built on the assumption that the firm, for any output level q, will chose the number of plants that produces q at the lowest possible cost.

The discussion in the previous sections defines what we mean by economies of scale for a given cost function. Actually, that discussion is general enough to be interpreted both in terms of the technology of a single plant and in terms of the technology of a firm. In the first case the cost function *C(q)* in the above section should be understood as *C(q,1)* and in the second case it should be understood as $C^{Firm}(q)$. Thus, there is no reason to have a separate discussion of measures of scale economies at the firm level.

In some empirical studies of economies of scale associated with $C^{Firm}(q)$ it is found that there is one region of low output in which unit cost is falling, and that there is one region of high output where unit cost is increasing. There is also a substantial range of intermediary output levels where unit cost is at the minimum. The reason for this shape may be multi-plant operations. The scale economies at low output levels may have to do with scale economies at the plant level. Then there is the intermediate region where the firm operates a number of firms at minimum cost. Finally there is a region where multi-plant diseconomies are important.

In the context of mergers one should distinguish between plant level scale economies and firm level economies. Most economies of scale at the plant level can only be exploited in the long run, when the use of capital can be adjusted. In contrast, important multi-plant economies of scale may be exploited even in the short-run. First, the elimination of administrative duplication may not require long lags. Second, there are also often product-specific economies of multiplant operation. If a firm produces different products in a single plant production time will inevitably be lost as assembly lines are switched from one product to another. By operating more than one plant, a firm can specialise the production of high-volume products in single plants. This will reduce down time due to shifting production, resulting in run-length economies. Run-length economies are important in many industries, including fabric weaving and finishing, shoe making, bottle blowing, and bearing manufacturing.[24] However, it is also important to point out that such short-run gains related to economies of multiplant operation can be expected to be insufficient for a merger to reduce price. The elimination of duplicated administrative functions affects fixed

[24] Scherer et al (1975, p 51).

rather than variable costs. The specialisation of production in different plants is an instance of so-called rationalisation, which is insufficient to reduce price.[25]

2.2.3 Scale Economies and Mergers

In this section we discuss the relationship between, on the one hand scale economies, and on the other hand the effect of merger on cost. If production is characterised by scale economies, under what conditions can they be exploited by means of merger? In part, this question was analysed by Farrell and Shapiro (1990).[26] However, a limitation in their analysis is that they focus on the case when the capital stock in each plant is fixed—the short run. Drawing on the theory of sub-additivity, we analyse the effect of mergers on cost in the long run—when firms can adjust their capital stock freely.

Farrell's and Shapiro's short-run analysis and the long-run analysis undertaken in this report are complements. The first analysis considers the extreme case when the capital stock in every plant is fixed. The second analysis considers the other extreme case when firms freely can adjust their capital. In the long run, cost savings are larger than in the short run. Hence, the latter form of analysis is more likely to produce a positive picture of the effects of the merger. As a consequence, antitrust authorities may reach different decisions regarding a merger, depending on whether they consider the long run or the short run effects. To complicate the picture further, it is likely that capital adjustments require different amounts of time in different industries. If the antitrust authority must evaluate the effects of the merger within the first two-year period, long-run analysis may be relevant for some industries and short-run analysis may be relevant for other industries.

A strength with the analysis is that we are able to analyse not only the effect of merger on cost. Using the concept of sub-additivity, we are also able to say if a full merger is the best way to restructure production in order to produce the given output as cheaply as possible (section 2.2.3.3). This is important since there are often several alternatives, short of a full merger, that may reduce cost without imposing all the anti-competitive effects of a horizontal merger. Antitrust authorities are often required to make such a broad evaluation of mergers. The reason is the requirement that cost savings should be merger specific.

We should also mention that all cost comparisons are made at a given quantity of output. In reality, a merger will affect output for several reasons. The reduction of competition will tend to reduce output. The reduction of cost, on the other hand, will tend to increase production. The net effect may be negative or positive depending on the circumstances. It is customary, however, to decompose the change in cost into two parts. The first part is the change in cost due to increased efficiency. This effect is must be measured at some given level production. Following Williamson (1968) the normal procedure is to measure this effect at the level of production chosen by the merged entity. The second part is the change in cost due to the change in quantity. (The second part is normally subtracted from the change in consumers' surplus to generate a measure of the change in dead weight loss resulting from the merger.)

---

[25] Farrell and Shapiro (1990).

[26] Their analysis is described in Röller, Stennek and Verboven (1999).

However, there are also some limitations in this analysis. Any adjustment costs associated with merger are not included in the analysis.

2.2.3.1 Single-Product Firms

In the single product case, economies of scale (in the relevant range) implies that a merger reduces cost. To see this, consider an industry where there are economies of scale up to production of at least q units. In other words:

$$\frac{C(q)}{q} < \frac{C(tq)}{tq} \quad \text{for all } t < 1.$$

That is, the unit cost of producing q is lower than the unit cost of producing only $tq < q$. Consider a merger between two firms, one producing tq and the other producing $(1-t)q$. Thus, they together produce q units. Before the merger the total cost of production is $C(tq) + C((1-t)q)$ and after the merger the total cost of production is $C(q)$. In this case, a merger reduces cost, that is[27]

$$C(q) < C(tq) + C((1-t)q).$$

The reason is simple. Since the unit cost of production is lower at q than at any level below q, it must be less costly to produce $tq + (1-t)q = q$ units in one firm than in two separate firms. Thus, in the single product case there exists a simple and important relation between scale economies and merger:

> **Observation 2.1:** Economies of scale (in the relevant range) are sufficient for a merger to reduce cost in the long run, in the single product case.

However, we should emphasise that this analysis suffers from several limitations. First, the analysis does not have anything to say about what happens to cost in the short run, before the input of capital can be changed. Second, the analysis abstracts from any adjustment costs.

The qualifier *in the relevant range* is important. In many industries the technology is characterised by increasing returns to scale at low production levels, then a segment of approximately constant returns to scale, and eventually, there are diseconomies of scale. In short, the unit cost function is U-shaped. In order for a merger to reduce cost, it is only important that there are scale economies up to the level of production that firms will carry out. Above that level, possible diseconomies of scale will not affect the merger.

Although scale economies are sufficient for a merger to lower cost, it is important to note that scale economies are not necessary for a merger to reduce cost. Here, we will consider two reasons for this to be so. If the two firms have access to *different*

[27] To see this rewrite the inequality as

$$\frac{C(q)}{q}\left[tq + (1-t)q\right] < \frac{C(tq)}{tq}tq + \frac{C((1-t)q)}{(1-t)q}(1-t)q\,.$$

Hence:

$$0 < \left[\frac{C(tq)}{tq} - \frac{C(q)}{q}\right]tq + \left[\frac{C((1-t)q)}{(1-t)q} - \frac{C(q)}{q}\right](1-t)q\,.$$

It is easy to see that economies of scale implies that this inequality is fulfilled, since both []-parenthesis are positive.

technologies (i.e. different knowledge) a merger may reduce cost also in the absence of economies of scale. It may be that both firms have constant returns to scale but that one firm has a superior technology with a lower marginal cost. Then, transferring knowledge to the other plant, or transferring production to the superior plant, lowers cost. It may also be that the two firms have different, but complementary, patents so that a merger expands the production technology to both plants.

A more surprising result is that even if the two firms have access to the *same technology*, a merger may reduce cost in the presence of diseconomies of scale. Assume that the firms have the cost function $C(q) = a + cq$ for $0 < q < y$ and $C(q) = a + b + cq$ for $y < q$, where $a > b > 0$. That is, there is a fixed cost (a) associated with production up to y, and then there is an additional fixed cost ($b < a$) associated with production beyond that level. For example, consider a shipping market. In order to ship the quantity $0 < q < y$ it is necessary to build a firm including various administrative functions and to invest in one ship (i.e. the fixed cost a is equal to the cost of administration plus the cost of one ship). In order to ship the quantity $0 < q < 2y$ it is necessary to invest in a second ship (i.e. the fixed cost b is equal to the cost of one ship). However, the administration may not need to be duplicated. Then, average cost is increasing as production is expanded from a level slightly lower than y to a level slightly higher than y (a second ship must be bought). Nevertheless, it is cheaper to produce in one firm, than to keep two independent firms since $b < a$ (running only one firm avoids the duplication of administrative costs). Thus:

> **Observation 2.2:** Economies of scale are not necessary for a merger to reduce cost, even if the firms have access to the same technology.

To sum up the case of single output technologies: economies of scale are sufficient, but not necessary, for a horizontal merger to reduce cost in the long run.

2.2.3.2 Multi-Product Firms

In the multi product case, economies of scale are not sufficient for a merger to reduce cost. One reason for this is that there may be overall scale economies at the same time as there are diseconomies of scope. In that case, a merger between two formerly separate product lines may increase costs. This fact is demonstrated in a theoretical example (Example 1 in Appendix 2.2.3.4). Thus, despite possible econometric evidence showing the existence of scale economies, a merger would nevertheless increase cost.

More surprising is that economies of scale and economies of scope together do not suffice to guarantee that a merger will reduce cost. This fact is demonstrated in a theoretical example (Example 2 in Appendix 2.2.3.4). The gist of the example is the following. Consider a farmer producing proteins and textile fibres. The farmer can use three processes. The farmer can raise sheep, producing both proteins and fibres (Marshallian joint production). Alternatively, the farmer can grow beans to produce proteins or flax to produce textile fibres. However, since the cheep will destroy the crop, the farmer must fence in the smaller of these operations which is costly. The presence of a (small) fixed cost entails economies of scale. The economies of scope comes from the Marshallian joint production in raising sheep. It is more expensive to produce one unit of fibre and one unit of protein by growing flax and beans. At the same time there is an element of diseconomies of scope due to the negative externality between raising sheep and growing crop. However, in the example the first effect dominates the latter. To minimise production cost in this case there should be one firm raising sheep to produce proteins and fibres. Then there should be one firm growing either beans or flax depending on whether there is need to produce additional protein or additional fibres. In this way Marshallian joint production is exploited and fencing is avoided. Thus:

> **Observation 2.3:** Economies of scale and economies of scope (in the relevant range), together, do not suffice to guarantee that a merger will reduce cost.

Thus, in the multi-product case, there is no logical connection between economies of scale and cost savings from mergers:

> **Observation 2.4:** Economies of scale are neither necessary nor sufficient for a merger to reduce cost, in the case of multi-product firms.

In view of this result, it is clear that a stronger set of sufficient conditions is required to guarantee that mergers reduce costs. Actually, when pushing this analysis further, it is reasonable to generalize the discussion, and to replace the concept of scale economies with the concept of subadditivity. This discussion is found in the next Section 2.2.3.3.

2.2.3.3 Merger Specificity

The above discussion was exclusively focusing on the question if a merger reduces cost. From a policy point of view, however, that is not the relevant question. The reason is that in most countries with an efficiency defence, there is a so-called merger specificity clause. According to this requirement, cost reductions can not save a horizontal merger unless the merger is the least anti-competitive means of achieving the cost savings. Thus, if there exists some alternative restructuring of production, short of a full merger, that achieves the same cost cuts (or more), the merger should not be allowed. In Example 3 below we show that it may well be that a merger reduces cost as compared to the initial production structure, but that there may exist another restructuring that lower cost even more. In the example there are two product varieties and two firms that both produce both varieties. A merger would reduce cost due to product specific economies of scale. However, due to diseconomies of scope, cost would be reduced even further if the two firms specialise in one of the varieties each. This alternative restructuring may also be preferred from a competition point of view, if consumers consider the two varieties as substitutable.

In order to accommodate this complication, the rest of the discussion is not focusing on the question if a merger reduces cost. Rather, we ask the question if a merger *minimises* the cost of producing a given output (equal to the aggregate output of the two firms before the merger). The central concept in this discussion is so-called strict subadditivity. To define this concept, consider the case of two products. Assume that we want to produce the quantity q_1 of product one and q_2 of product two. We may divide production in any way we want into k separate firms. Let q_1^2 denote the quantity of product one produced by firm two, and so on. Then,

$$q_1^1 + q_1^2 + ... + q_1^k = q_1 ,$$

$$q_2^1 + q_2^2 + ... + q_2^k = q_2 .$$

A cost function $C(q_1, q_2)$ is said to be *strictly subadditive* in the production of (q_1, q_2) if

$$C(q_1^1, q_2^1) + C(q_1^2, q_2^2) + ... + C(q_1^k, q_2^k) > C(q_1, q_2),$$

independent of how production is divided between the k firms, and independent of how many separate firm (i.e. k) there are.

> **Observation 2.5:** If the cost function is subadditive (in the relevant range), a merger reduces cost. Moreover, there does not exist any less concentrated alternative to the merger that reduces cost even more.

Hence, if the merging firms can show that their cost function is subadditive, they have established that, from the point of view of productive efficiency, a full merger is the best way of organising the production of (q_1, q_2). Expressed in a different way, subadditivity but not scale economies is the relevant concept to use in judging if a full merger is the best way of achieving cost efficiency. Naturally, productive efficiency must still be traded off against possible anti-competitive effects.

Still, however, subadditivity and scale economies are related, and the economic literature contains several sets of sufficient conditions (for example in terms of scale economies) for the cost function to be subadditive. These conditions may be of practical use to merger control, even though the sufficient conditions are more restrictive than subadditivity itself. The reason is that it may often be easier to find econometric evidence on for example scale economies than econometric evidence on subadditivity. The first condition is also intuitively appealing.

> **Observation 2.6:** If (i) there are product specific economies of scale for all products, and (ii) there are (weak) economies of scope, then the cost function is subadditive.[28]

Actually one can show that if there are product specific economies of scale for a particular product, then industry cost minimisation requires that production of this good is consolidated in a single firm. Economies of scope implies that all production lines should be allocated within a single firm.

Additional sets of sufficient conditions for subadditivity, for example in terms of transray convexity, can be found in the literature.[29]

2.2.3.4 Appendix: Theoretical Examples

The following examples are chosen because of their simplicity. They are not intended to portray real world merger decisions.

Example 1.

The simplest possible example to illustrate that economies of scale are not sufficient for a merger to reduce cost is the following. Assume that firms' costs are given by

[28] This condition was established by Baumol, Panzar and Willig (1982).
[29] See Baumol, Panzar and Willig (1982, ch.7) and Sharkey (1982, ch. 4).

$$C(q_1,q_2)=\begin{cases}0 & q_1=q_2=0\\ f+cq_1 & q_1>0, q_2=0\\ f+cq_2 & q_1=0, q_2>0\\ f+g+c(q_1+q_2) & q_1,q_2>0\end{cases}.$$

In this case, there are overall economies of scale since $S(q_1,q_2)=1+(f+g)/c(q_1+q_2)$ is larger than one. There are also product specific economies of scale since $S_1(q_1,q_2)=1+g/cq_1$ is larger than one (and similarly for product two). However, there are diseconomies of scope if $g>f$ since then $SC(q_1,q_2)=(f-g)/[f+g+c(q_1+q_2)]<0$. In this case, a merger increases cost if the two firms before the merger produced one of the two products each. In particular,

$C(q_1,q_2)-C(q_1,0)-C(0,q_2)=[f+g+c(q_1+q_2)]-[f+cq_1]-[f+cq_2]=g-f>0$.

Thus, despite possible econometric evidence showing the existence of scale economies, a merger would nevertheless increase cost.

Example 2.

To demonstrate the second point (that economies of scale and economies of scope together do not suffice to guarantee that a merger will reduce cost) consider the following example. Raising sheep costs EUR 10 per animal and yields one unit of protein and one unit of fibre. Growing beans or flax costs EUR 6 per unit of protein or fibre obtained. Fencing costs EUR 1 per unit of the smallest activity. The setup cost, *e*, is small. To minimise the cost of producing proteins, *p*, and fibres, *f*, the farmer should raise sheep, *s*. In particular he should raise $s=\min\{p,f\}$ sheep. The farmer must also grow crop, either flax or beans. In particular he must grow the quantity $c=\max\{p,f\}-\min\{p,f\}$. Thus, the cost function is

$$C(p,f)=10s+6c+\min\{s,c\}+e,$$

where the third term is the fencing cost. Due to the fixed cost, *e*, the cost function exhibits increasing returns to scale everywhere. If the same amount of fibres and proteins are produced in separate firms, the cost is is

$$C(p,0)+C(0,f)=6(p+f)+2e=6(\max\{p,f\}+\min\{p,f\})+2e.$$

The cost function exhibits economies of scope everywhere since

$$C(p,0)+C(0,f)-C(p,f)=$$

$$=2\min\{p,f\}-\min\{\min\{p,f\},\max\{p,f\}-\min\{p,f\}\}+e\geq\min\{p,f\}+e\geq 0$$

Assume now that the market demands more protein than fibres, so that $p>f$. Consider dividing production of (p,f) between two firms with output levels (f,f) and $(p-f,0)$. This division results in total costs of

$$C(f,f)+C(p-f,0)=[10f+e]+[6(p-f)+e].$$

This division of production into two firms lowers cost since

$$C(f,f)+C(p-f,0)-C(p,f)=$$

$$= [4f + 6p + 2e] - [10f + 6(p - f) + \min\{f, p - f\} + e] < 0,$$

when e is small. (Remember that the fixed cost is small.) A merger between the two firms would increase cost due to the need for fencing.

Example 3.

Consider again the cost function in Example 1 in Section 2.2.3.2. Assume that in the outset, both firms produce both products. Let q_1^2 denote firm two's production of good one, and so on. The firms propose a merger. Such a merger will reduce cost:

$$C(q_1^1 + q_1^2, q_2^1 + q_2^2) - C(q_1^1, q_2^1) - C(q_1^2, q_2^2) =$$
$$= [f + g + c(q_1^1 + q_1^2 + q_2^1 + q_2^2)] - [f + g + c(q_1^1 + q_2^1)] - [f + g + c(q_1^2 + q_2^2)] =$$
$$= -(f + g) < 0.$$

However, if the two firms instead of a merger agree to specialise in one of the products each, then the costs of production would be reduced even further (this follows immediately from Example 1).

2.2.4 Summary and Conclusions

We begin our analysis with an intuitive discussion of the sources of scale economies. A production process is said to have scale economies if the unit cost of production is lower when more is produced. There are several sources of scale economies, such as mechanisation and specialisation achieved through division of labour. Without constructing a full typology, we compile a "checklist" of the most common sources of scale economies.

We also draw a sharp distinction between technologically determined economies of scale and overheads such as advertising and R&D. Due to state of economic knowledge, we are forced to primarily focus this overview on the first type of scale economies. However, we also argue that this state of knowledge is very unsatisfactory. Advertising and R&D are important features of most concentrated European industries.

Next, we define the term scale economies more rigorously. This section is an important input for discussing the relationship between mergers and production costs, and also for presenting the empirical literature. In the economic literature there exists two different definitions of scale economies, one based on the so-called production function, the other based on the so-called cost function. We argue that for the purpose of analysing efficiency gains from mergers, it is vital to use the concept based on the cost function. The reason is that the exploitation of some scale economies (e.g. mechanisation) requires changing the factor proportions. Such gains are only captured by the concept based on the cost function. The primary complication in defining scale economies arises when firms produce more than one product. In this case, there may exist overall diseconomies of scale even if there are (product specific) economies of scale for all individual products. The reason is that there may exist so-called diseconomies of scope. Another complication is that the minimum efficient scale of a firm depends on the mix in witch the firm produces the different products.

We then discuss the theoretical relation between mergers and scale economies. That is, if production is characterised by scale economies, under what conditions can they be exploited by means of merger? In part, this question was analysed by Farrell and Shapiro (1990).[30] However, a limitation in their analysis is that they focus on the case when the capital stock in each plant is fixed—the short run. Drawing on the theory of sub-additivity, we analyse the effect of mergers on cost in the long run—when firms can adjust their capital stock freely.

In the single product case, the relation is rather straightforward. Economies of scale (in the relevant range) are sufficient, but not necessary, for a merger to reduce cost.[31] In the multiple-product case, the primary finding is that that the relation between scale economies and mergers is rather complex. Even if there are unexploited scale economies, mergers may well raise rather than reduce cost. A positive finding is that one can establish different sets of sufficient conditions for mergers to reduce cost. These conditions are based on the notion of sub-additivitity of the cost function. For example, if there are product specific economies of scale for all products, and there are economies of scope, then a merger reduces cost. In effect, these theoretical results show that in evaluating the effect of a merger on cost, competition authorities need to take into account both product specific economies of scale and potential economies of scope. However, in order to make these conditions more precise additional analysis, similar to that undertaken by Farrell and Shapiro for the short-run case, is needed.

A strength with the analysis is that we are able to analyse not only the effect of merger on cost. Using the concept of sub-additivity, we are also able to say if a full merger is the best way to restructure production in order to produce the given output as cheaply as possible. This is important since there are often several alternatives, short of a full merger, that may reduce cost without imposing all the anti-competitive effects of a horizontal merger. Antitrust authorities are often required to make such a broad evaluation of mergers. The reason is the requirement that cost savings should be merger specific.

Farrell's and Shapiro's short-run analysis and the long-run analysis undertaken in this report are complements. The first analysis considers the extreme case when the capital stock in every plant is fixed. The second analysis considers the other extreme case when firms freely can adjust their capital. In the long run, cost savings are larger than in the short run. Hence, the latter form of analysis is more likely to produce a positive picture of the effects of the merger. As a consequence, antitrust authorities may reach different decisions regarding a merger, depending on whether they consider the long run or the short run effects. To complicate the picture further, it is likely that capital adjustments require different amounts of time in different industries. Therefore, if the antitrust authority evaluates the effects of the merger within the first two-year period,

[30] Their analysis is described in Röller, Stennek and Verboven (1999).

[31] If the two firms have access to *different technologies* (i.e. different knowledge) a merger may reduce cost also in the absence of economies of scale. It may be that both firms have constant returns to scale but that one firm has a superior technology with a lower marginal cost. Then, transferring knowledge to the other plant, or transferring production to the superior plant, lowers cost. It may also be that the two firms have different, but complementary, patents so that a merger expands the production technology to both plants.

long-run analysis may be relevant for some industries and short-run analysis may be relevant for other industries.[32]

In sum:

> **Conclusion:** *Theoretical analysis can establish the conditions under which mergers generate cost savings, related to scale economies, in the long run. Such conditions may also include a test for merger-specificity. To make these conditions fully operational for antitrust authorities, more research is needed.*

2.3 Empirical Estimates of Scale Economies

The question that we are interested in answering is to what extent production costs can be reduced as a result of horizontal mergers. Unfortunately, however, there is very little direct empirical evidence on the effect of merger on cost. The purpose of this section is to review the empirical studies of scale economies in various industries, and to investigate if such evidence can be used to indirectly shed light on the likely efficiency gains from mergers.

There are several approaches for measuring cost-scale relationships. One approach has been to analyse profitability as a function of firm size. The attractiveness of this approach is that data on profits at the firm level are relatively easy to collect. However, we will not focus on this approach since profits are affected by so many other circumstances than just scale economies. In particular, since profits are sensitive to market power, and since firm size may well be related to market power, this approach does not seem well suited in the present context an efficiency defence in merger control, which is all about the relative strength of market power and cost efficiency. A second approach is the so-called survivor test.[33] Firm or plant sizes that survive and gain larger market shares over time are assumed to be efficient, and those that lose market shares over time are considered to be too large or too small. Again, one may suspect that selection operates on profitability (or possibly relative profitability) and hence that it is difficult to separate scale economies from market power using this approach.

In this review we focus on direct estimates of the cost-scale relationship. There are three approaches within this class, namely econometric cost function (or production function) analysis, the mathematical programming approach (DEA), and the engineering approach.

Engineering approach. Bain[34] who covered twenty different industries undertook the pioneering study. This approach is built on interviews (or questioners) with engineers whose profession it is to plan and design new production units and plants.

Econometric approach. Economists have used econometric techniques to estimate cost or production functions. The estimated functions may be investigated concerning their scale economy properties. In this category we include traditional least squares econometric production models and stochastic frontiers methods.

---

[32] Unfortunately, the empirical literature does not report any measures of how long time that is required in order to achieve scale economies in different industries. Another important issue that is unexplored in the literature is the possible differences in time required for achieving scale economies by means of merger and by means of internal growth.

[33] See Stigler (1958).

[34] See Bain (1956).

Mathematical programming approach. This approach is similar to the econometric approach in that the aim is to estimate a frontier cost or production function, using input and output data collected from different firms. However, the so-called data envelopment analysis (DEA) builds on mathematical programming rather than standard statistical techniques.

For a much more detailed (technical) discussion and comparison of the econometric and mathematical programming approaches, see Coelli, Rao and Battese (1998). A methodological discussion and an extensive survey of the engineering approach can be found in Pratten (1988).

2.3.1 Econometric Approach: Methodology

This section will review the empirical results from cost-function based estimates of scale economies. The discussion of the empirical results is sorted by industry. We start with some methodological issues.[35]

Cost vs. Production Function. One issue is whether to directly estimate a production function, or to estimate a cost function. As we have already argued in section 2.2.2, in the present context the cost-function approach is preferred. It is the cost function that determines industry structure. From an econometric perspective, estimation of a cost function may be preferable since it accounts for the endogeneity of inputs. Another reason is that cost functions are more easily generalised to multiple output situations.[36]

Varying Vintage. To estimate a cost function one has to gather data on the cost of production from existing plants operating at varying scales of output. Such data usually relate to plants built at different points in time. The plant and equipment is of varying vintages and the latest plant and equipment may incorporate knowledge which was not available when the earlier units where built. As a result the estimates will usually not reflect the technology of a particular point in time. Expressed in other words, cost differences between different firms may not only reflect varying scale but also differences in vintages.

Aggregation: The Homogeneity Problem. The early empirical literature on scale economies has been severely criticised on several accounts. According to Gold (1981),[37] the single most important reason for the inadequacies of the early research was the tendency to carry out the analysis at too high levels of aggregation. Often analysis was carried out at four, three or event two digit levels of industrial classification. At this level of aggregation firms do not produce homogenous goods, and comparisons of costs at different "scales" may reflect differences in products rather than economies of scale. This aggregation prevents conformance with the requirements not only of theory, but of practical evaluations by industrial executives and by government officials as well.

---

[35] A recent discussion of the econometric issues involved in empirical analysis of production can be found in Coelli, Rao and Battese (1998).

[36] For more on these issue see Coelli, Rao and Battese (1998). They also discuss the choice of functional forms. The first study to use the cost function, with factor prices as arguments, in the study of scale economies appears to be Nerlove (1963).

[37] Gold (1981).

Sources of Scale Economies. Researchers have lacked an interest to identify the sources of scale economies.[38] The dominant concern has been to mechanically describe the relation between firm or plant size and unit cost. This is unfortunate in the present context. Scale economies may often be achieved in different ways, some of which may be less damaging to competition than mergers. In some cases a joint venture may achieve the gains. Thus, for antitrust authorities to evaluate different alternatives, it should be valuable to have an understanding of the sources of scale economies.

Aggregation: The Multiple-Product Problem. Since most large firms produce a large variety of products, estimation of their cost functions involves a large number of parameters. Due to data limitations, researchers must aggregate the output data in some way. There are two methods that have been used to do so. The typical method in the early literature was to construct a single scalar measure of output by only considering firm revenues. Unfortunately, this approach implicitly imposes the restriction that the so-called transformation rate between different products is constant. In other words, it is assumed that the cost function can be written as $C(q_1, q_2) = \tilde{C}(a_1 q_1 + a_2 q_2)$ where a_1 and a_2 are two constants. As a result, it is impossible to distinguish between different products' product-specific economies of scale. It is also impossible to distinguish between economies of scale and economies of scope.

There is one case in which it is possible to infer properties of C from estimates of $\tilde{C}$. If all firms at all times (included in the sample) produce the same output mix, the economies of scale calculated from $\tilde{C}$ will coincide with the degree of multiple-product economies of scale *at the observed product mix*. However, it is in general not valid to extrapolate this measure to other product mixes.

The more modern method is the so-called *hedonic cost function* approach.[39] Assume that we want to estimate the cost function for airlines. Every destination-origin pair is a separate market. Thus, in an economy with 100 cities, there are 9 900 different types of output. Thus the cost function $C(q_1, ..., q_{9900})$ is likely to contain a very large number of parameters. However, it is likely that passenger miles and the number of cities served by the firm determine the cost for airlines. One may then estimate a hedonic cost function $\tilde{C}(m, c)$ where *m* is passenger miles and *c* is the number of cities served. In this case 9 900 variables have been turned into only two. After $\tilde{C}$ has been estimated one may indirectly compute measures of scale economies for *C*. However, this final step is not trivial, and the reasoning behind it is an important contribution in itself.[40]

Long vs. Short-Run Functions. Another empirical problem is that the available data often allows a short-run but not long-run cost function to be estimated. For example, if a cost function is to be estimated using monthly data from a single firm, it would be unrealistic to assume that the firms capital inputs were adjusted to the cost-minimising level associated with each month's output level. Unfortunately, the short-

---

[38] Gold (1981).
[39] This approach was pioneered by Spady and Friedlaender (1978).
[40] See Caves, Christensen and Tretheway (1984).

run cost function does in general not carry the information needed to analyse economies of scale.[41]

2.3.2 Econometric Approach: Empirical Estimates

Regulated or formerly regulated industries have been much more studied than many other industries. The primary reasons for this is that the data have been much more available due to the regulations. In a sense this is fortunate since antitrust issues are especially important in these industries after deregulation. Here we review the development in the trucking and electric power industries that have generated much of the methodological development. However, there has also been econometric studies of the economies of scale of several other industries, including railroads,[42] autos,[43] telecom,[44] and others.

Trucking

In empirically analysing the costs of a multiple-product firm, the basic problem is one of aggregation. Since most firms typically produce a wide range of diverse products, it is generally impossible to incorporate all outputs in the analysis, because of the limited number of observations relative to the potential number of parameters that would have to be estimated. For example, trucking output is very heterogeneous. Not only do trucks haul different commodities with different handling requirements, but they also haul them in different shipment sizes, in different loads, and for different lengths of haul. Drawing on theoretical work on production theory, Spady and Friedlaender (1978) attempted to take the composition of output into account by introducing the operating characteristics of the firm explicitly into the cost function. They introduced a so-called hedonic output function that attempted to measure effective output as a function of physical output (q) and the characteristics of that output ($x_1,...,x_m$) where y was measured in tonmiles and the x's represented variables such as shipment size, average load, average length of haul, and so on. Thus, instead of estimating a traditional cost function as $c = C(q,w)$ they estimated a cost function as $c = \tilde{C}(H(q,x),w)$ where $H(q,x)$ represents hedonically adjusted output and w represents a vector of factor prices.

They found that the inclusion of operating characteristics made a significant difference. When costs are estimated using physical output (q) alone, marked economies of scale where found. However, when operating characteristics (the x's) were included, the measured scale economies disappeared. Since larger firms tend to utilize longer hauls and larger shipment sizes, they tend to have lower costs. However, if smaller firms could have the same operating characteristics as large firms, they could presumably achieve comparable advantages. (The small firms in the sample were kept from achieving the operating characteristics or the large firms because of limited regulatory operating rights.)

The study by Spady and Friedlaender (1978) has two important limitations. Due to the specification of the hedonic cost function the operating characteristics (or output

[41] See Braeutigam and Daughety (1983).

[42] See Jara-Diaz and Whinston (1981), and Jara-Diaz and Cortes (1996).

[43] See Friedlaender, Whinston, and Wang (1983).

[44] See Fuss and Waverman (1981), and Fuss (1983) for a survey. See also Foreman et al. (1999).

composition) is not allowed to affect the substitutability between different factors of production. However, Friedlaender and Spady (1981) solved this problem by specifying the cost function as $c = \widetilde{C}(q, x, w)$.

The second limitation is that the aggregate measures of output not only vary with the scale of output but also with its composition. Thus, measures of scale economies based on their approach are necessarily ambiguous since they may reflect the effects of both scale and scope upon costs. Later contributions have resolved also this problem. Thus the hedonic cost function approach can be adapted to distinguish between different types of economies of largeness (see the study by Caves et al., 1984).

There is also a study of the trucking industry attempting to distinguish between economies of scale and scope. This study incorporates desegregate output, but also variables that reflect the configuration and utilisation of the network over which the firm operates. Wang Chiang (1981)[45] shows that firms enjoy marked economies of networking. Networks having many connected links and networks permitting direct routing leads to significant saving in costs. These network effects explain much of the economies of scale and scope that Wang Chiang observed.

Electric Power

Nerlove (1963) studies economies of scale in electricity generation, using data from 1955.[46] The study is based on modern duality theory. The estimated cost function includes factor prices as arguments. The basic estimation equation was a Cobb-Douglas log-linear specification of the cost function:

$$\ln C = \kappa + (1/\sigma)\ln q + (1/\sigma)\sum \alpha_i [\ln p_i].$$

Here σ is the (single product) degree of scale economies. The p_i's refer to the prices of labour, fuel and capital. The sum $\sum \alpha_i$ is constrained to equal unity. The main finding is that there exists important economies of scale in electricity generation. In particular the model yielded an estimate of $\sigma \approx 1.4$. That would mean that if the firm's cost would exceed its revenues by forty percent if its output were priced at marginal cost.

The functional form used restricts the degree of economies of scale to be the same for all output levels. To check the importance of this restriction, Nerlove divided the sample into five output categories. Doing so revealed that the economies of scale are exhausted at large plant sizes. In the largest group there where diseconomies of scale.

Nerlove's study has had a methodological influential in two ways. First, it established the cost function, with its factor price arguments, as the proper framework in which to study economies of scale. Second, it showed that the degree of scale economies declines with output, thereby demonstrating the need for more flexible functional forms than the one used by Nerlove himself.

Christensen and Greene (1976) employ more flexible functional forms, allowing the degree of scale economies to vary with the level of output. In particular, they employed the translog cost function:

---

45 This study is usefully summarized in Bailey and Friedlaender (1982).
46 This study is usefully summarized in Panzar (1989).

$$\ln c = \kappa + \alpha \ln q + \beta [\ln q]^2 + \sum a_i \ln p_i + \sum \sum b_{ij} \ln p_i \ln p_j + \sum \gamma_i \ln q \ln p_i \ .$$

In order for this function to represent a proper cost structure some restrictions on the parameters of the model must be imposed. Using data from 1970, they found that the cost function is U-shaped, with a large segment that is approximately flat. Approximately 45 percent of the output was produced in the minimum cost segment.

A problem with the above studies is that they do not take the multiproduct nature of the studied industry into account. Many electric utilities sell natural gas as well. Another complication is that firms producing electricity at peak times incur much higher costs than firms generating electricity off peak. If the peak/off peak mix varies across firms, estimates of the single output cost function may be biased. Empirical work on multiproduct issues did not begin until the developments reported in the section above had taken place.

Mayo (1984) extends the analysis to incorporate the multi-product nature of operations. In particular, he allows for two types of output, namely electricity and gas. Using data from 1979, he found that diseconomies of scope set in at small output levels. Chappell and Wilder (1986) used 1981 data for the same sample but left out utilities that generate more than 10 percent of their electricity using nuclear power plants. In contrast to the earlier results, their estimates indicate product specific economies of scale and economies of scope that are not exhausted even at very large levels of production. This has triggered a debate whether or not one should include very different techniques (nuclear vs. other) in the same study (see Mayo, 1986; Gold, 1981; Panzar, 1989).

2.3.3 Mathematical Programming Approach

In the previous sections we have implicitly presumed that all firms are fully efficient. In this section we relax this assumption and describe methods that may be used to estimate so-called production frontier functions and measure the efficiencies of firms relative to these estimated frontiers. The two principal methods that have been used are data envelopment analysis (DEA) and stochastic frontiers. We focus here on the first type. Since the methodology is based on mathematical programming rather than standard econometric analysis it is more different from the methods discussed above.

For simplicity, consider an industry that produces one output *q* using one input *x*. Assume that we have a data on the inputs used and outputs produced for *n* firms. The data points are: $(q_1, x_1), \ldots, (q_n, x_n)$..

The essential assumption behind DEA analysis is that the production possibility set is convex. That is, if two input-output combination (q_1, x_1) and (q_2, x_2) are technically feasible, then also any input-output combination (q, x), where $q = \lambda q_1 + (1-\lambda) q_2$ and $x = \lambda x_1 + (1-\lambda) x_2$ with $0 \leq \lambda \leq 1$, is technically feasible. Due to the convexity assumption, the production frontier can be approximated as the piece-wise linear function that connects the most efficient firms. Next, we can measure the firms (technical) efficiency and their scale efficiency. The degree of technical efficiency is defined in the following way. Firm one consumed x_1 to produce q_1. If it would have been possible to produce the same quantity q_1 using only $x_1^t < x_1$, then the technical efficiency of the firm is equal to x_1^t / x_1. Note that only if $x_1^t = x_1$ the technical efficiency is 100 per cent. Otherwise it is lower. Note also that the cost of production could have been lowered by $\left(x_1 - x_1^t\right)/x_1$ percent. Thus the degree of technical efficiency is directly corresponding to the cost-savings potential. If the efficiency is 90 percent, cost could be lowered by 10 percent. The degree of scale inefficiency is defined in a similar way. If it would have been possible to produce the same quantity q_1 using only $x_1^s < x_1^t$ at the most efficient scale, then the firm's scale efficiency is equal to x_1^s / x_1^t . This measure of scale inefficiency is very convenient since it is expressed in terms of how much cost can be reduced if production is carried out at the most efficient scale. One may also note that total efficiency is the product of the technical efficiency and scale efficiency, that is $x_1^s / x_1 = \left(x_1^s / x_1^t\right)\left(x_1^t / x_1\right)$.

We can illustrate the application is DEA analysis with a study of Australian universities (Coelli, 1996).[47] In 1996 the University of New England (UNE) formed a committee to to look at the performance of UNE relative to 36 other universities in Australia. Of particular concern was some evidence that suggested that the UNE's administration appeared to be larger than some comparable universities. The study concerned the calender year 1994. The specification of the outputs of a university administration was a challenging task. But it was decided that the output would be proxied by two measures: the number of students and the number of staff. Some of the results on scale efficiency are reported in Table 1.

Table 1: DEA Results for the Australian Universities Study.

University	Scale efficiency	Type
Australian Catholic University	0.94	Increasing returns
Australian National University	1.00	-
…	….	…
University of New England	0.99	Increasing returns
…	…	…
University of Western Sydney	0.99	Decreasing returns
Mean	0.94	-

Source: Coelli et al. (1998).

In total the study found that 14 of 36 universities are operating on a too low scale with increasing returns to scale, and that 13 of the universities were operating on a too large scale with decreasing returns to scale. The mean scale efficiency was 0.94 indicating that if universities had been operated at optimal scale they would have been able to save on average six percent (not weighted) of their cost.

[47] This study is usefully summarized in Coelli et al (1998).

2.3.4 Engineering Approach

The pioneering study was Bain (1956) who covered twenty different industries. This approach is built on interviews (or questioners) with engineers whose profession it is to plan and design new production units and plants. The engineering approach has also been followed by Pratten (1971), Scherer et al. (1975), Weiss (1975) and more recently, by Gasmi and Laffont (1999). Most of the engineering studies are relatively old, and there are reasons to believe that the estimates for many industries may have changed since the times of the studies. Thus, in our view, there is not much point in presenting a detailed account of the studies at the level of particular industries. Moreover, such an extensive survey of the results can be found in Pratten (1988, Section 5). Instead we point at the more general implications of these studies for merger control. However, to give some feeling for these studies, we do provide some results from one of them below.

The central concept in the studies based on the engineering approach is the minimum efficient scale (MES). In theory, the MES is the level of output at which unit cost is at its minimum. In practice, the MES is measured as the minimum level of output (or other measure of size) that must be reached in order to have only a “slight” cost disadvantage compared to the minimum unit cost. The definition of the minimum efficient scale differs between different studies (Pratten, 1988). Another source of difference between different studies is that some studies include all types of costs, while other studies only focus on production costs and exclude selling and distribution costs (e.g. Sherer, 1975). Still others go even further and exclude development costs.

The engineering studies produce two important measures. First, in order to estimate how much that is gained by exploiting scale economies, the studies measure the increase in cost of operating below the minimum efficient scale. For example, the cost may be ten percent higher if a plant is operated at half or one third of the MES. Second, in order to measure the implications of scale economies for (the cost efficient) industry structure, the studies measure the size of the minimum efficient scale plant or firm relative to the size of the market varies. For example, if the minimum efficient scale of a firm is five percent, then the market can support twenty independent firms and still allow production at minimum cost.

The main result of the engineering studies is that the extent of economies of scale varies across industries. The increase in cost of operating below the minimum efficient scale varies, and the size of the minimum efficient scale plant or firm relative to the size of the market varies. These results have two important implications for merger policy.

First, the differences between industries support the idea that cost savings related to scale economies should be evaluated on a case-by-case or at least industry-by-industry basis. A general presumptions approach, with thresholds common for all industries, is likely to produce important decision-making errors. At the same time the engineering studies show that it should be possible for antitrust authorities to collect information that describes the extent of scale economies in a particular industry.

Second, the engineering studies show that concentration in many markets is often higher than the concentration necessary for exploiting scale economies. It is therefore important that only scale economies *in the relevant range* are used as an argument to support horizontal mergers.

2.3.4.1 Cost Savings Due to Plant Size

Scherer et al. (1975) cover twelve major industries in seven industrialised nations. The results on the minimum efficient plant scales in the U.S. are summarised in Table 2. Two measures are given. The second column reports the size of a minimum efficient scale plant, assuming mid-1960s best-practice technology. Scale is measured by capacity, output or employment. The third column reports the percentage elevation of long run unit cost as a consequence of building a and operating plants at one-third the MES rather than the full MES.

Table 2: Minimum Efficient Plant Scales, and the Cost Disadvantage of Suboptimal Scale Plants.

Industry	**Minimum Efficient Scale**	**Percentage by which unit cost rises at one-third MES**
Beer brewing	4.5 million barrels per year capacity	5.0
Cigarettes	36 billion cigarettes per year	2.2
Cotton and synthetic broadwoven fabrics	37.5 million square yards per year	7.6
Paints	10 million U.S. gallons per year	4.4
Petroleum refining	200000 barrels per day crude oil capacity	4.8
Leather shoes	1 million pairs per year	1.5
Glass bottles	133000 tons per year	11.0
Portland cement	7 million 376-pound barrels per year capacity	26.0
Integrated steel	4 million tons per year capacity	11.0
Anti-friction bearings	800 employees	8.0
Refrigerators	800000 units per year	6.5
Automobile storage batteries	1 million units per year	4.6

Source: Scherer et al. (1975).

It is not obvious how one should summarise these results. It would have been instructive to relate the cost differentials connected to size to, for example, the actual total cost differentials between different plants in these industries. However, even without such a benchmark, it seems reasonable to say that in most the concerned industries, during mid 1960s, the long-run cost disadvantage of operating at sub-optimal scale is not huge. In half of the industries, the elevation is five percent or less. But there are exceptions. The cost disadvantage of operating at one-third MES in the cement industry is 26 percent.

These results indicate that in some industries, but not in most of them, cost savings related to plant economies of scale can be an important effect of horizontal mergers. In other words, there exists a gain from evaluating such cost savings on a case-by-case basis.

> **Observation 2.7:** Engineering approach studies indicate that cost savings related to plant economies of scale differ substantially between different industries. These differences indicate that such cost savings should be evaluated on a case-by-case or at least industry-by-industry basis.

However, the picture provided by the above table is very incomplete. The analysis must also take into account economies of multiple-plant operation. And, perhaps more important, in order to understand the implications for merger policy, the MES must be related to the size of the market.

2.3.4.2 Scale Economies in Relation to the Size of the Market

It is crucial to point out that the cost minimising industry structure, and thus the rationale for mergers, depends on the interplay between scale economies and the size of the market. Consider a single product industry with increasing returns to scale up to the production of q^{MES} and thereafter decreasing returns to scale. Thus, q^{MES} is the minimum efficient scale (MES). If consumers demand exactly q^{MES} units of the good, or less, then the cost minimising organisation of the industry requires that all production is carried out in a single firm. Thus, mergers up this level reduce cost. On the other hand, if consumers demand $2q^{MES}$ units of the good, cost minimisation requires that production is divided between two separate firms. In this case, a merger between two duopoly firms will increase the cost of production. Thus, in order to support a merger with efficiency claims, it is not sufficient to show that the technology exhibits scale economies in some range. It must be demonstrated that there are scale economies in the relevant range. In many industries, it is likely that this qualification is crucial.

Economists have estimated minimum efficient scale and compared the estimates with observed levels of market concentration. Typical results from one study, using the engineering approach, are reported in Table 3, showing how important multiplant economies of scale were in the late 1960s for twelve U.S. industries. Column (1) shows estimates of how many MES plants a firm needs to operate in order to have not more than slight overall handicap vis-à-vis companies securing all multiplant economies. Column (2) translates the column (1) judgements into an estimate of how large a share of the U.S. market an enterprise needed to realise these economies. In effect it summarises the imperatives for high concentration at the nation-wide level. In only three of the twelve industries (refrigerators, brewing, and perhaps cigarettes) where oligopolistic national industry structures (defined as four-firm concentration ratio above 40) compelled. However, in several other industries, transportation costs were sufficiently high to confine the sales of any single plant to a regional market. As a result, exploiting scale economies required oligopolistic industry structures also in many glass bottle, petroleum, steel, and cement markets. Column (3) indicates the average market share held by individual firms among the industries' three leading producers in 1970. By comparing column (3) with column (2) it appears that in the United States, concentration was much larger than the operation of efficient firms would dictate. In weaving and batteries, the Big Three firms were ten times as large as they needed to be to enjoy all economies of scale.

Table 3. Firm size required to experience not more than slight cost handicap in relation to the size of the U.S. market.

Industry	Number of MES plants needed to have not more than "slight" overall handicap (1)	Share of U.S. market required in 1967 (2)	Average market share per U.S. Big Three member, 1970 (3)
Beer brewing	3-4	10-14%	13%
Cigarettes	1-2	6-12	23
Fabric weaving	3-6	1	10
Paints	1	1.4	9
Petrolium refining	2-3	4-6	8
Shoes	3-6	1	6
Glass bottles	3-4	4-6	22
Cement	1	2	7
Ordinary Steel	1	3	14
Bearings	3-5	4-7	14
Refri-gerators	4-8	14-20	21
Storage batteries	1	2	18

Source: Scherer et al. (1975, pp 334-6).

The important conclusion from this table is the following:

> **Observation 2.8:** Comparing minimum efficient scale with the size of the market shows that real world concentration may be considerably higher than the imperatives of scale economies require.

The table does not reveal the reason for the excessive concentration. However, at least one potential explanation is the firms' strategies to build market power. Another potential reason may be that economies of scale in marketing have not been properly accounted for in the studies. Be that as it may, these empirical results show that it is important that only scale economies in the relevant range should be used as a defence for a horizontal merger.

2.3.5 Summary and Conclusions

In this section, we review the empirical literature on economies of scale. We both provide a brief overview of empirical estimates of scale economies in several industries, and discuss the methodologies that are used. However, since many of these studies are rather old, and since scale economies in many industries are likely to change over time, as a result of technological development, we do not attempt to provide a detailed account of the measurement of scale economies for all industries. Rather, the main purpose is to extract the general conclusions, from this literature, that are relevant for the discussion of an efficiency defence in merger control.

The primary finding is that the exploitation of scale economies is a source of substantial cost savings in some, but not all, industries. Economies of scale are often present at low volumes of output, and exhausted at larger volumes. In our view, the empirical results have an important implication for merger policy. The variability of scale economies across industries and different output volumes support the idea that cost savings should be evaluated on a case-by-case basis, i.e. an efficiency defence. Moreover, the empirical literature shows that concentration in many industries has gone further than motivated by scale economies alone. Thus, in evaluating mergers it is important that only scale economies in the relevant range should be used as a defence for a horizontal merger. In sum:

> **Conclusion:** *The variability of scale economies across industries supports the idea that cost savings should be evaluated on a case-by-case basis. Only scale economies in the relevant range should be used as an efficiency defence.*

In order for antitrust authorities to take cost savings related to scale economies into account, they need some methodology to assess the importance of these economies. Therefore, we also provide a short introduction to the empirical methodologies that have been used to estimate economies of scale. Essentially, three different methods have been used to study the relationship between size and unit cost. The econometric approach uses statistical techniques to analyse cost-output data from some collection of firms. The mathematical programming approach is similar in that it also analyses cost-output data, but using linear programming. The engineering approach relies on interviews with engineers who design new production units.

In our view, all of these methodologies are useful for antitrust authorities, albeit in different ways. The already existing studies based on the econometric and mathematical programming approaches have produced useful "background knowledge." Whenever a case has to be decided concerning some industry, these studies can give a first hint if scale economies are important in that particular industry. What is the order of magnitude of potential cost savings? Is it reasonable to carry out a more detailed study?

However, for several reasons the econometric approach is probably not suitable for the antitrust authorities to carry out such more detailed investigation. The main reason is data limitation. In order to estimate scale economies, cost and output data also from firms not participating in the merger is necessary. Another reason is that the econometric approach does not identify the sources of scale economies, which is important in order for assessing merger specificity. Estimates of scale economies do not indicate the time it takes for a merger to achieve cost savings related to scale economies. In contrast, the engineering approach should lend itself for such more detailed studies in particular cases. In particular, the engineering approach may supply an "incremental analysis," that is focus on the factors that will be affected by the merger. In sum:

> **Conclusion:** *Antitrust authorities can use existing econometric evidence on scale economies, to indicate if the order of magnitude of potential cost savings motivate a more detailed study, in particular merger cases. If detailed investigation is warranted, studies based on the engineering approach can be commissioned.*

3. PASS-ON OF COST SAVINGS INTO CONSUMER PRICES

3.1 The importance of pass-on of cost savings

The relevant question for competition policy is to which extend the total cost savings, as discussed in the previous chapter, affect the policy objectives. For this reason, it is necessary to be more explicit about policy goals of merger regulation. As discussed in Röller, Stennek and Verboven (2000), several objectives are frequently mentioned in the competition policy debate.

- Consumer surplus. A first objective upon which merger analysis may be based is the protection of consumer interests. If this is the case, the central focus of merger analysis is on the competitive, or price effects, of mergers
- Total surplus. Another objective may be to further both consumer and producer interests. Total surplus may operationally be defined as the sum of consumer and producer surplus. More generally, one may give different weights to consumer and producer surplus.
- Other objectives. These include the promotion of European integration, employment, regional balance, viability of small firms, and competitiveness of national firms on international markets. The concern with the preservation of employment has often been a political concern in proposed mergers. In principle, employment considerations should be taken into account in a full cost-benefit analysis of mergers. In practice, a merger policy designed to preserve old production structure is presumably not the best way to deal with the employment objective in the long run. See for example Jenny (1997) and Crampton for more on the employment objective.

Depending on which policy objective one has in mind, the total cost savings should be evaluated differently. In our analysis below we focus exclusively on the first policy goals, the protection of consumer surplus. The reason is that this goal is typically the main objective in merger control.

Our focus on consumer surplus implies that we are mainly interested in the price effects from mergers. There are, however, also reasons to be interested in the price effects if one adopts the producers' own point of view. Indeed, understanding the price effects is essential to evaluate changes in the merging firms' competitive position. If a merger leads to a reduction in price, the merging firms typically realize an increase in their market share. In contrast, if a merger leads to a price increase, the merging firms are likely to experience a shrinking market share. While the merging firms' profits may still increase in this case, this would be caused by increased market power and thus not by an increased competitiveness.

The price effects from a merger can be decomposed into two components:

- The price increase from in increase in market power, holding the merging firms' costs constant.
- The price reduction from an improvement in the firms' competitive position.

The price reduction from the improvement in the firms competitive position will be strong to the extent that (i) the realized cost savings are marginal cost savings; and (ii) the cost savings are passed on to at least some extent to consumers.

The total price effect from the merger is therefore negative if pass-on of marginal cost savings is sufficiently high to outweigh the market power effects.[48] This shows the importance of assessing pass-on of cost savings into consumer prices in merger analysis. . Generally speaking, the greater is the pass-on, the more likely a merger is going to decrease price in the presence of cost savings. For example, if 50 percent of the cost savings are passed on to consumers, then a merger decreases price if the realized cost savings are at least twice the percentage amount of the price effects from increased market power. If cost savings are fully passed on to consumers, then a merger decreases price if the realized cost savings are at least the same percentage amount as the price effects from increased market power.

These examples illustrate the importance of understanding the degree of pass-on to evaluate the full price effects from mergers. What are the relevant determinants of pass-on? Do technological conditions affect the degree of pass-on? What is the role played by competition? How important is the distinction between pass-on of firm-specific and industry-wide cost savings? What is the empirical evidence regarding pass-on? We take up these questions in this chapter with the purpose of obtaining a better understanding into the role of efficiencies in mergers.

3.2 Theoretical analysis of pass-on

3.2.1 Two measures of pass-on

As an introductory remark, it is instructive to note that pass-on of marginal cost savings can be measured in various ways. The choice of measure depends on what one wants to emphasize.

One approach is to compute how the absolute price level would change when the marginal cost is changed by, say, one unit. We refer to this measure as *absolute pass-on*. A second approach is to compute how the price would change in percentage terms, when the marginal cost is changed by, say, one percent. This gives a measure of *relative pass-on*, or, in other words, the *pass-on elasticity*. Both measures will be used throughout the text.

In fact, it can be verified that the absolute pass-on is simply equal to the pass-on elasticity, multiplied by the ratio of price to marginal cost. Since this ratio is typically greater than one (when firms have positive markups), the absolute pass-on is typically greater than the pass-on elasticity. Only under perfect competition (when markups are zero) both measures are the same.

---

48 Assume that the merger tends to increase price by dP/P percent as a result of the first effect. Then, for the net effect to be non-positive, it is necessary that $(dc/c)(dP/dc)(c/p) \leq -dP/P$, where (dc/c) is the percentage cost reduction, and $\beta=(dP/dc)(c/p)$ is the pass-on elasticity. The pass-on elasticity indicates by how many percent the price would be reduced if cost is reduced by one percent.

3.2.2 Pass-on of industry wide cost savings

We begin with a discussion of pass-on of cost savings that are realized by all firms in the industry. While cost savings from mergers are typically realized only in the level of the merging firms, understanding pass-on of industry-wide cost savings is a useful starting point. We can draw from a well-established theoretical literature on competition and monopoly and add generalizations to oligopolies. Furthermore, we will show how industry-level and firm-level pass-on are related to each other. This is useful, since the largest part of the empirical literature, which we review in section 3.3, has been concerned with industry-level pass-on.

Sections 3.2.2.1 to 3.2.2.3 consider pass-on of industry wide cost savings under alternative forms of competition, assuming products are homogeneous. Section 3.2.2.4 investigates how the results generalize to differentiated products.

3.2.2.1 A competitive industry

Consider first pass-on of industry-wide cost savings in a competitive industry. This case is treated in the context of taxation by Bishop (1968), and is well explained and illustrated in several Economics textbooks, see e.g. Pindyck and Rubinfeld (1992). It is useful to review this case, before considering more complicated market structures with market power.

In a competitive industry firms choose production levels to maximize profits, taking the market price as given. An individual firm's supply is then given by the production level at which the marginal cost of production equals the market price. The aggregate supply curve is simply the sum of the individual firms' supplies, as depicted by the curve $Q^S=S(P)$ in Figure 1. Aggregate demand is the sum of the individual consumers' demand, shown by the curve $Q^D=D(P)$ in Figure 1. The equilibrium market price is the price at which supply is equal to demand, $Q^S=Q^D=Q^*$. It is given by the solution to $S(P)=D(P)$, as shown by the intersection of the two curves in Figure 1.

We are interested in assessing pass-on of an industry-wide cost saving, i.e. the effect of a decrease in the marginal costs on the price level. More precisely, we want to compute the industry-wide pass-on elasticity, defined as the percentage increase in the market price cause by a percentage increase in marginal costs. To address this question, it will be convenient to use the inverse of the aggregate supply curve $S(P)$, which is the aggregate marginal cost curve, denoted by $P=wMC(Q^S)$. The parameter w is used to analyse the effects of percentage changes in marginal costs. Using this curve, the equilibrium market price is given by the condition:

$$P = wMC(D(P)) .$$

Industry-wide pass-on can be computed by totally differentiating this condition:

$$\left(1 - \mathrm{w}\frac{\mathrm{d}MC}{\mathrm{d}Q}\frac{\mathrm{d}D}{\mathrm{d}P}\right)\mathrm{d}P - MC\mathrm{d}w$$

Rearranging gives the following expression for the industry-wide pass-on elasticity:

$$\tau^R = \frac{dP}{dw}\frac{w}{P} = \frac{1}{1+\varepsilon\omega},$$

which would also be the measure for absolute pass-on.

In this expression $\varepsilon = -\frac{dD}{dP}\frac{P}{Q}$ is the price elasticity of demand, measuring the effect of an increase in the price by 1 percent on the percentage reduction in demand. A high price elasticity of demand (a high ε) means that consumers are very price sensitive, which is depicted by a flat demand curve. The term $\omega = \frac{dMC}{dQ}\frac{Q}{MC}$ measures the effect of an increase in the quantity produced by 1 percent on the percentage change in marginal cost. It is, in fact, the inverse of the elasticity of supply, and it may be positive or negative. If ω is negative, then marginal costs decline as output increases, which can be depicted by a downward sloping supply curve. In contrast, if ω is positive, then marginal costs increase with output, depicted by an upward sloping supply curve. This indicates the presence of capacity constraints. If ω is close to zero firms, face little capacity constraints, which can be depicted by a flat marginal cost curve.

The formula for the industry wide pass-on elasticity under perfect competition thus stresses the importance of the slope of the demand and supply curves. To illustrate, we begin our discussion by assuming that ω is positive or zero. Figure 2a assumes that demand is somewhat elastic while supply is upward sloping, meaning that firms face increasing marginal costs as they increase their production. Consider a reduction in marginal cost by 20 percent. Holding production levels constant, perfect competition would induce firms to lower their prices by 20 percent. However, such a price reduction would induce consumers to buy more. To bring the market back in equilibrium, firms need to produce more. Yet since they face increasing marginal costs (capacity constraints), they are only willing to do this if the price can increase. As a result, price will not decrease by 20 percent but by a lower amount.

Pass-on can only be complete if either one of two conditions is met: (i) demand is perfectly inelastic, consumers are very price insensitive (ε=0 as in Figure 2b); (ii) supply is perfectly elastic, capacity is available in unlimited amounts so that marginal costs are constant (ω=0 as in Figure 2c). For example, if marginal costs are constant, then a cost reduction by 20 percent corresponds to a price reduction by 20 percent, since new demand at the lower price can be easily produced without affecting marginal costs. But more generally, pass-on is less complete the more consumers respond to price changes (high elasticity of demand) and the more producers are capacity constrained (low elasticity of supply). For a perfectly elastic (flat) demand curve and increasing marginal costs (ε=infinity and ω>0), a marginal cost reduction is not passed on at all to consumers. Similarly, there is no pass-on if there are very severe capacity constraints, i.e. for a perfectly inelastic (vertical) supply curve (ω=infinity).

This discussion highlights that the degree of pass-on crucially depends on the consumers' responses to price reductions and on the firms' ability to expand production. This can be summarized in the following observation.

> **Observation 3.1:** In a competitive industry, pass-on of a reduction in marginal costs will be more complete, as consumers respond less to price reductions and as producers can expand production more easily without raising their marginal costs by too much (little capacity constraints).

Several empirical studies have emphasized the importance of capacity constraints in explaining incomplete pass-on. For example, in international trade, studies by Bhagwati (1988), Branson (1989) or Knetter (1995) have investigated whether incomplete pass-through of exchange rates is caused by the presence of quantitative import quota restraints. Import quota constraints may be viewed as an example of severe capacity constraints in the country of import.

This discussion is based on the assumption that the supply curve is upward sloping, i.e. the firms' marginal costs increase as production increases. In practice, when the industry is subject to increasing returns to scale, it is possible that the firms' marginal costs are decreasing as production increases, i.e. $\omega<0$.[49] In this case, there will be overshifting: a reduction in marginal cost by, say, 20 percent would lead to a reduction in price by more than 20 percent (unless demand is perfectly inelastic). The intuition for this is clear. At a constant production level, competitive firms would lower their price by 20 percent. This induces consumers to buy more. The required additional output can be produced at a lower marginal cost since marginal cost is decreasing in production. This induces competitive firms to reduce the price further until a market equilibrium is reached.

> **Observation 3.2:** In a competitive industry, there will be overshooting if the supply curve is downward sloping, i.e. marginal costs decrease as producers expand production.

The extent of pass-on in a competitive industry is thus largely empirical question: there may be incomplete pass-on, complete pass-on, or even overshooting. Nevertheless, there seems to be a common acceptance that pass-on is complete (pass-on elasticity of one) in a competitive industry. This can be rationalized if one is willing to make some additional assumptions: (1) the competitive industry is characterized by free entry and (2) all firms have access to the same technology and production factors. Under these assumptions the marginal cost curve is flat and supply is perfectly elastic (ω=infinity), so that pass-on is complete.

> **Observation 3.3:** In a competitive industry, pass-on of a reduction in marginal cost savings is complete if the marginal cost curve is flat. For some industries, this may occur in the long run, when entry is free and firms have access to the same technologies and production factors.

While these assumptions may be more or less justified in the long-run, they are not likely to hold in the short-run, since entry and exit takes time, imitation of competitors' technologies may be difficult, and production factors are not perfectly mobile. For these reasons, even in competitive industries the assessment of the pass-on rate is an empirical question.

[49] We make the standard stability assumption that marginal cost is not decreasing too much, i.e. $-\omega<\varepsilon$ at the equilibrium, so that the supply curve intersects the demand curve from below..

3.2.2.2 Monopoly

We now take another extreme and assume that there is only one single firm in the industry. We ask the question whether and by how much a monopolist will pass on cost reductions to consumers. In contrast to a competitive firm, which takes the market price as given, a monopolist takes into account that it can influence the market price. Its profits are *P Q* minus total costs *C(Q)*. Assuming as before that demand is equal to supply so that demand is given by *Q=D(P)*, the monopolist's profits are given by total revenues minus total costs:

$$PD(P) - wTC(D(P))$$

The profit-maximizing price is given by the following condition:

$$D(P) + \frac{\mathrm{d}D}{\mathrm{d}P}\left(P - wMC(D(P))\right), \qquad (2)$$

where *wMC(Q)* is as before the marginal cost curve.

The first term measures the gain from a price increase by one unit: it is proportional to the total sales at the current price level. The second term measures the loss from a price increase: this is proportional to the markup multiplied by the reduction in sales caused by the price increase. We can rewrite (2) in elasticity form as:

$$\frac{P - wMC(D(P))}{P} = \frac{1}{\varepsilon} \qquad (2')$$

In words, the percentage markup is inversely related to the price elasticity of demand. Only if the price elasticity of demand is infinite, the markup is zero, so that price is equal to marginal cost. We note from (2') that a monopolist always operates at the elastic portion of the demand curve: since marginal costs are positive, it must be that $\varepsilon>1$. Intuitively, at the inelastic portion of the demand curve ($\varepsilon<1$), it always pays for a monopoly to raise its price.

We can follow a similar approach as under perfect competition and totally differentiate the monopoly condition.[50] This gives the following formula for relative pass-on, or the pass-on elasticity:

$$\tau^R = \frac{\mathrm{d}P}{\mathrm{d}w}\frac{w}{P} = \frac{1}{1 + \varepsilon\omega + \frac{1}{\varepsilon - 1}\eta}, \qquad (3)$$

which measures the effect of a percentage marginal cost change in the percentage price change To measure absolute pass-on, one simply has to multiply the pass-on elasticity by the price to marginal cost ratio, as discussed in section 3.1.

The first two terms in the numerator of expression (3) are exactly the same as in the case of perfect competition. Consequently, it is again true that pass-on is more complete as consumers respond less to price reductions and as producers can expand production without raising their marginal costs by too much (see Observation 3.1). To abstract from these effects, assume the marginal cost curve is flat (ω=0), so that pass-on would be complete under perfect competition (see Observation 3.3).

[50] See, for example, Feenstra (1989), Marston (1990), or Knetter (1993)

We can then focus on the third part in the numerator of expression (3). This contains the term η, which measures the elasticity of the price elasticity of demand, i.e. $\eta = \frac{\mathrm{d}\varepsilon}{\mathrm{d}P}\frac{P}{\varepsilon}$. It may be positive or negative. If η>0, then a price increase leads to a higher price elasticity of demand, i.e. consumers become more price sensitive. This is the most intuitive scenario. If η<0, then a price increase, leads to a decrease in the price elasticity of demand; if η=0, then the price elasticity of demand is independent of price changes.

To interpret the role of the parameter η in assessing pass-on, suppose that the monopolist experiences a reduction in marginal cost of 20 percent and initially completely passes this on to consumers by lowering its price. This will cause an increase in demand, which the monopolist can easily accommodate since we assume a flat marginal cost curve (ω=0). However, the lower price may cause the monopolist to reconsider its markup strategy. There are three possibilities, as can be seen from the markup formula (2'). First, if the elasticity of demand is unaffected by price changes (η=0), then the optimal monopoly markup is a fixed percentage of cost. The monopolist never adjusts its percentage margins, so that (relative) pass-on is complete. Second, if the elasticity of demand is increasing in price (η>0), then a price reduction would decreases the elasticity of demand. A lower price elasticity calls for an upward adjustment of the optimal monopoly markup. The monopolist thus prefers to raise its percentage markup and pass on only part of the cost reduction. Third, if the elasticity of demand is decreasing (η<0), lowering price increases the elasticity of demand, so that the optimal monopoly markup decreases. In this case, the monopolist prefers to pass on more than the cost reduction by lowering its markup, so that there will be overshooting.

We can summarize this discussion as follows.

> **Observation 3.4:** Suppose the marginal cost curve is flat (ω=0). If consumers become more price sensitive when price increases (η>0), a monopolist raises its percentage markup after a marginal cost reduction so that relative pass-on is incomplete. If consumers become less price sensitive when price increases (η<0), a monopolist reduces its percentage markup after a marginal cost reduction so that there is overshooting.

This discussion shows the important role of market power in assessing pass on. A monopolist may respond to cost reductions by lowering or raising its markup, depending on the curvature of the price elasticity of demand. The central question for assessing pass-on when monopoly market power is present is thus whether the price elasticity of demand is increasing or decreasing in price. To gain some further intuition, it is useful to compute the elasticity of the price elasticity of demand η as:

$$\eta = 1 + \varepsilon + \frac{\mathrm{d}D/\mathrm{dP}}{\mathrm{d}^2 D/\mathrm{dP}^2} P \qquad (4)$$

This expression shows that the sign of η depends on the *curvature* of the demand curve, as measured by $\mathrm{d}^2 D/\mathrm{dP}^2$. If the demand curve is concave or linear all terms are positive, so that η>0, i.e. the price elasticity of demand is increasing in price. Hence for concave or linear demand curves ($\mathrm{d}^2 D/\mathrm{dP}^2 \leq 0$), firms increase markups after a marginal cost reduction so that pass-on is incomplete (assuming flat marginal costs). Only if the demand curve is sufficiently convex (d^2D/dP^2 sufficiently high) it is possible that firms reduce their markups after a marginal cost reduction so that there is overshooting.

In practice, while one cannot rule out the case of overshooting, it is more reasonable to expect that pass-on of cost reductions will be incomplete. To see this, consider how the value of the price elasticity of demand for two extreme values of marginal cost. First, if marginal cost equal zero, then by the inverse elasticity rule, the price elasticity of demand is equal to one. Second, if marginal cost is so high that only the smallest possible amount is consumed, then the price elasticity of demand must be very large. Hence, the price elasticity of demand drops considerably as one moves from the highest to the lowest possible cost level. Only for a range of prices it may be that the price elasticity of demand increases as price falls. Assuming that marginal costs are flat, we conclude that overshooting, i.e. lowering the markup in response to a marginal cost reduction, is a rather unusual scenario.

Even if one rules out the possibility of overshooting as implausible, the degree of pass-on cannot easily be pinned down to a small range. To illustrate, consider the commonly used specification of linear demand ($d^2D/dP^2=0$) while still assuming flat marginal costs (ω). Substituting (4) into (3) one can compute that the pass-on elasticity of a monopolist is then equal to $(\varepsilon-1)/2\varepsilon$. This implies that the pass-on elasticity is always less than 0.5. In contrast, the commonly used specification of a constant elasticity demand ($\eta=0$) implies that pass-on is complete. This illustrates that one can obtain drastically different conclusion regarding pass-on, by simply changing the demand specification. Assessing the pass-on of cost savings by a firm with monopoly market power is thus, more than under competition, an empirical question.

3.2.2.3 Oligopoly models

To explain the role of competition for assessing pass-on, we consider an oligopolistic industry where firms can influence the market price. One may distinguish between two types of oligopoly models: the Cournot and the Bertrand model. In the Cournot model, each firm chooses a quantity to maximize profits, taking as given the quantity set by the other firms. In the Bertrand model, each firm chooses a price to maximize profits, taking as given the price set by the other firms.

Consider first a Cournot model with N firms, where each firm i operates at a constant, but not necessarily identical marginal cost c_i. Market demand $P=P(Q)$ is now represented as the inverse of the demand function $Q=D(P)$; it is downward sloping. The constant marginal cost assumptions allows us to abstract from the possibility of incomplete pass-on or overshooting stemming from, respectively upward or downward sloping marginal cost functions. Except for this assumption,, this is a fairly general set-up, as described in more detail by for example Levin (1991).[51] We are interested in finding the industry-wide pass-on elasticity, which we define as the percentage reduction in price when all firms reduce their marginal cost by one percent.

A firm i producing a quantity q_i obtains a profit equal to:

[51] For example, some weak stability assumptions for guaranteeing existence and uniqueness of equilibrium are required.

$$\left(P(Q) - c_i\right)q_i$$

Each firm i chooses a quantity q_i to maximize profit given the quantity chosen by the other firms. The profit maximizing quantity for each firm i must satisfy the condition that the firm's marginal revenue equal its marginal cost, or:

$$\frac{\mathrm{d}P}{\mathrm{d}Q} q_i + P - c_i = 0$$

This condition must hold for every firm to have a Cournot equilibrium. As under monopoly, we can write this in elasticity form:

$$\frac{P - c_i}{P} = \frac{s_i}{\varepsilon}$$

where $s_i = q_i/Q$ is the market share of firm i. Intuitively, a firm obtains a high markup if the price elasticity of demand ε is low (as in the monopoly case) and if its market share is high. In the special case in which $s_i=1$, a single firm has 100 percent of the market and the monopoly condition holds.

We can add up these conditions for all firms to obtain the following equation from which the Cournot market quantity can be solved:

$$\frac{\mathrm{d}P}{\mathrm{d}Q} Q + NP - \Sigma c_i = 0$$

The market price can then be computed by substituting the Cournot market quantity in the inverse demand function $P=P(Q)$. We can compute the pass-on elasticity by totally differentiating this condition. Some calculations show that the industry-wide pass-on elasticity is equal to:

$$\tau^R = \frac{\mathrm{d}P}{\mathrm{d}\Sigma \mathrm{c_i} / N} \frac{\Sigma \mathrm{c_i} / N}{P} = \frac{1}{1 + \frac{1}{N\varepsilon - 1}\eta}, \qquad (5)$$

where η is the elasticity of the price elasticity of demand as defined previously for the monopoly case. To obtain a measure of absolute pass-on, one would need to multiply the pass-on elasticity by the ratio of price to the average of the firms' marginal cost.

One can see that the industry-wide pass-on elasticity depends on the number of competing firms N, but is independent on the distribution of market shares in the industry. It thus does not matter whether the industry consists of, say, four equally sized firms, or whether instead there is one dominant firm and three small ones.

The pass-on elasticity (5) covers monopoly and perfect competition as special cases (for the case of constant marginal costs). More specifically, if N=1 the monopoly formula holds. As the number of firms N increases (and assuming $\eta>0$), the pass-on elasticity increases, until it becomes unity as the number of firms N approaches infinity. As in the monopoly case, when the price elasticity of demand increases with price ($\eta>0$), firms find it optimal to increase their markup after a reduction in marginal cost. We can summarize this as follows.

> **Observation 3.5:..** Suppose that the marginal cost curve is flat and the price elasticity of demand increases with price. Cournot firms raise their markups

after an industry wide marginal cost reduction so that relative pass-on is incomplete. As the number of firms in the industry increases, pass-on becomes more complete.

McCorriston et al. (1995) consider an oligopoly with identical firms and constant marginal costs. Conduct is modeled through a conjectural variation parameter θ. This parameter is often used in empirical work (see Bresnahan, 1989) to cover a broad range of oligopoly models as special cases, including perfect competition, Cournot, Bertrand and cartel. In the symmetric Cournot model the conjectural variation would be equal to 1/N. McCorriston et al. consider the case in which the cost of only one production factor decreases; the share of this production factor is called α. Assuming the production factors are used in a fixed proportion their formula reduces to:

$$\tau^R = \frac{\mathrm{d}P}{\mathrm{d}\Sigma c_i / N} \frac{\Sigma c_i / N}{P} = \frac{\alpha}{1 + \frac{1}{\varepsilon / \theta - 1} \eta}.$$

This is very similar to what we found before. The formula first shows the role of conduct through the conjectural variation parameter: pass-on increases as the market becomes more competitive (lower θ). The formula also shows that it is important to know whether or not the cost reduction applies to all production factors. If the cost reduction only applies to a fraction α of the marginal costs, then the pass-on elasticity has to be adjusted correspondingly.

Observation 3.6: The pass-on elasticity increases as conduct becomes more competitive. Furthermore, the pass-on elasticity is proportional to the share of the production factors for which the cost reduction applies.

It is important to bare in mind this observation when interpreting the results obtained in the empirical literature. In particular, if a low pass-on elasticity is found, this may simply be because the empirical study only considered one component of costs.

Next consider the Bertrand model of oligopoly, where firms compete by choosing prices taking as given the prices of their competitors. Assume again that firms have constant but not necessarily identical marginal cost. If products are homogeneous, then the price will equal to the marginal cost of the firm with the second lowest marginal cost and the market share of the lowest cost firm will be 100 percent. This is true no matter how many firms are in the industry. In this case, one can easily see that pass-on of an industry-wide marginal cost change will be complete, no matter how many firms are in the industry. The Bertrand model with homogeneous goods thus has the rather unappealing property that the number of firms has no influence on the degree of pass-on. In the next section, it will become clear that this is no longer the case if one assumes that product are differentiated.

3.2.2.4 Industries with product differentiation

Up to now we have assumed that consumers view the products offered by different firms as perfect substitutes. As a result, there is only one price in the industry; if there were two prices, than the products at the highest price would obtain zero market share. With homogeneous products pass-on of marginal cost savings is therefore defined straightforwardly in terms of the single prevailing price, independent of the firms market shares. If products are differentiated, things are more complicated. Different products are typically sold at different prices and have different sales. A natural way to define pass-on would be to construct a price index for the industry, where the products are weighted according to sales. Pass-on could then be defined as the effect of a reduction in marginal cost on this price index (holding the weights fixed). A justification for this approach is that the change in this price index when all prices change by one unit is exactly the opposite of the change in consumer surplus, if we ignore income effects. To see this, let consumer surplus be equal to:

$$CS = V(p_1, p_2, ..., p_N, y)$$

where $V(p_1, p_2, ..., p_N)$ is the indirect utility function of a representative consumer in the industry. Applying Roy's identity, it can be checked that the change in consumer surplus is given by:

$$\mathrm{d}CS = q_1 \mathrm{dp}_1 + q_2 \mathrm{dp}_2 + ... + q_N \mathrm{dp}_N ,$$

where q_i are the sales of product i. When we assess pass-on of cost reductions on the sales based price index, we are thus looking at pass-on to consumers.

The pattern of product differentiation in a market may take many forms and can differ from market to market. Generally speaking, one may view competition in industries with product differentiation to be either localized or global. If competition is localized, then products are direct substitutes of only a few other products in the market. Many products then only compete indirectly with each other through a chain of price responses by competitors. In contrast, if competition is global, then all products are direct substitutes for all other products in the market.[52] A simple form of global competition occurs when all products are symmetric substitutes for each other. An example of this is the popular logit model, which we will use to illustrate how to measure pass-on. It will be convenient to focus on absolute pass-on (rather than relative pass-on).

Suppose there are N differentiated products, $i=1...N$; consumers either choose one unit of their most preferred product, or they decide not to buy anything, in which case they choose the outside good 0. According to the logit model, the market share s_i for product i (in the total sales including the outside good) is given by the following formula:

$$s_i = \frac{\exp((v_i - p_i)/\mu)}{\sum \exp((v_k - p_k)/\mu)}$$

where v_i measures the intrinsic utility for the product, and μ is a parameter measuring the degree of substitution. The total sales q_i for product i are equal to the market share s_i times the total number of potential consumers, L. Assuming a flat marginal cost curve c_i for each product i, the profits from product i are given by $(p_i - c_i)s_i L$.

We consider pass-on of an industry-wide marginal cost reduction in two kinds of market structure: a monopoly and a Bertrand oligopoly. A monopolist takes into account that a price increase does not only reduce the sales of the product itself, but also increase the sales of the other products it sells. The monopolist thus sets prices to maximize its total profits, taking into account the profit effects for *all* products it sells.

---

[52] See Anderson, de Palma and Thisse (1993) for a detailed analysis of models of product differentiation. Models of localized competition are sometimes referred to as the Kaldorian approach, whereas models of global competition are referred to as the Chamberlinian approach.

The result is that the monopolist charges a markup that is inversely proportional to the price elasticity of demand for all products. This generalizes the monopoly markup rule for homogeneous products given by (2'). Specifically for the logit model, this rule translates into the following condition:

$$p_i - c_i = \frac{\mu}{s_0}$$

Intuitively, the monopolist charges equal markups for all its products, and these markups are higher the less attractive is the outside good (low elasticity of market demand). One can totally differentiate this condition to obtain the following expression for the absolute pass-on, i.e. the effect of a unit marginal cost change on the industry price index:

$$\tau^A = \sum \frac{\mathrm{d}CS}{\mathrm{d}c_k} = (1 - s_0)s_0 \qquad (6)$$

This simple condition implies that pass-on of a unit cost reduction for all products will not be passed completely into the consumer price index. In particular, a monopolist selling differentiated products of the logit form never passes on more than 1/4 of a cost reduction (which occurs when the outside good has a market share of 1/2).

This can be contrasted with pass-on in an oligopoly. Suppose that there are N firms each selling one of the differentiated products. Each firm takes into account that a price increase reduces the sales of its own product, but does not consider the effect on the sales of the other products. The resulting markup is inversely related to the price elasticity at the level of the individual product, which is greater than the price elasticity of demand for all products together. The closer the products are substitutes, the higher the product-level price elasticities of demand, and the lower the markups. More specifically, the price rule for firm i must satisfy the following condition:

$$p_i - c_i = \frac{p_i}{\varepsilon_i}$$

which for the logit model translates into:

$$p_i - c_i = \frac{\mu}{1 - s_i}$$

From this condition one can derive the following expression for pass-on of an industry-wide unit marginal cost change on the price index:

$$\tau = \sum \frac{\mathrm{d}CS}{\mathrm{d}c_k} = \sum \frac{s_k(1 - s_k)^2}{(1 - s_k)^2 + s_k} \Bigg/ \sum \frac{s_k(1 - s_k)}{(1 - s_k)^2 + s_k}$$

This is a rather complicated formula, yet it has several interesting properties. First, it can be shown that the pass-on expression is greater than the expression in (6). The logit example thus illustrates that also for differentiated products there is more pass-on under oligopoly competition than under monopoly. Second, it can easily be seen that pass-on is always less than one. For symmetric products, it can also be verified pass-on becomes more complete as the number of firms increases; pass-on become complete as the number of products become very large.

Observation 3.7: Suppose that the marginal cost curve is flat and products are differentiated according to the logit model. Absolute pass-on is incomplete, yet greater under competition than under monopoly. As the number of competing products becomes very large, absolute pass-on becomes complete.

Note that Observation 3.7 is only true for the logit model. It is well-known, see for example Froeb and Werden (1994), that the logit model generates price elasticities of demand that are increasing in price. For this reason, it implies that firms absorb cost changes partially by adjusting their markups. In other models of product differentiation, such as the linear demand model or the CES model, markup adjustment may be lower, so that it is possible that absolute pass-on would be more than complete.

3.2.3 Pass-on of firm-specific cost savings

We now go to the question that is most relevant for merger analysis: how are *firm-specific* cost savings passed on to consumer prices? In answering this question we are also concerned with the relationship between pass-on of industry-wide cost savings. Understanding this relationship will prove to be very important since our empirical review will make it clear that previous research efforts have traditionally been most concerned with pass-on of industry-wide cost savings.

3.2.3.1 A competitive industry

In a perfectly competitive industry, each firm's demand curve becomes horizontal. In other words, the elasticity of each firm's perceived demand becomes infinitely large (consumers become very price sensitive). Under such conditions, a firm has no incentive to pass-on any cost changes. It acts simply as a price taker, and would only adjust its produced quantity in response to a marginal cost change. See Yde and Vita (1996) for a graphical illustration.

3.2.3.2 Oligopoly with homogeneous products

First, we consider firm-level pass on in a Cournot oligopoly as introduced in section 3.2.2.3. Each firm chooses an output level to maximize its profits. The condition for profit-maximization can be totally differentiated as before, but now varying only the marginal cost of a specific firm i, holding the marginal costs of the other firms constant. This yields the following formula for the pass-on elasticity of firm i:

$$\tau_i^R = \frac{\mathrm{d}P}{\mathrm{dc_i}}\frac{\mathrm{c_i}}{P} = \frac{(\varepsilon - s_i)/(N\varepsilon - 1)}{1 + \frac{1}{N\varepsilon - 1}\eta} \qquad (7)$$

where the notation is the same as before. To measure the absolute pass-on, one would have to multiply the pass-on elasticity by firm i's price to marginal cost ratio.

The formula (7) looks quite close to the industry-wide pass-on elasticity given by (3). The denominator is in fact the same, showing that the price elasticity of demand ε, the elasticity of the elasticity η and the number of firms has a similar influence on the firm-specific pass-on elasticity as on the industry-wide pass-on elasticity. The numerator is different from (3). It contains the market share of firm i, who experiences the cost reduction. It shows that the pass-on elasticity is larger for firms with a smaller market share. Intuitively, this is because small firms are firms with high marginal costs and little market power.

The relationship between the firm-specific pass-on elasticity τ_i and the industry-wide elasticity τ can be discussed by computing the ratio between the two:

$$\tau_i^R / \tau^R = (\varepsilon - s_i)/(N\varepsilon - 1)$$

Note that this ratio is independent of the elasticity of the elasticity. In fact, the denominator of this ratio is simply the sum of the numerator over all firms. Since ε must be greater than s_i, this implies that this ratio is always less than one, i.e. the firm-specific pass-on is necessarily less than industry-wide pass-on in the Cournot model.

It is possible to say a bit more about the relationship between firm-specific and industry-wide pass-on. The ratio τ_i/τ implies that a firm with a market share equal to the average market share in the industry (i.e. 1/N) has a pass-on elasticity that is exactly equal to a fraction 1/N of the industry-wide pass-on elasticity. For example, if there were four firms in the industry, then a firm with a market share of 25 percent would have a pass-on elasticity of exactly 25 percent of the industry-wide pass-on elasticity. A firm with a higher market share than the average market share in the industry will necessarily pass-on less than the fraction 1/N, so that pass-on of a cost reduction by such firms is likely to be small. In contrast, a firm with a lower market share than the average in the industry will pass-on more than the fraction 1/N. If all firms are identical (with market shares of 1/N), then the firm-specific pass-on elasticity is exactly 1/N of the industry wide pass-on elasticity.

> **Observation 3.8:** Suppose that the marginal cost curve is flat. In a Cournot industry a firm-specific cost reduction is passed on less than an industry-wide cost reduction. The higher a firm's market share is, the lower is the firm specific pass on.

One can also easily compute the pass-on elasticity for multiple firms, i.e. the N^F merging firms. One simply has to reinterpret the market share in (7) as the joint (post-merger) market share of the N^F firms. This shows that pass-on from cost savings is lower the greater is the merging firms' joint market share. The special case of Dornbusch' (1987) analysis obtain if one considers pass-on of cost savings by N^F symmetric firms and a constant price elasticity of demand (η=0).[53] In this case, the pass-on elasticity is equal to the firms joint market share N^F/N: the greater the firms' joint market share, the closer pass-on is to unity.

Now consider firm-level pass-on in a Bertrand oligopoly where firms choose prices rather than quantities. As explained in section 3.2.2.3 in such an industry firms set a price equal to the marginal cost of the second lowest cost firm, and the market share of the lowest cost firm is 100 percent. In such an industry a cost saving by the dominant, low cost firm is not passed on at all to consumers, whereas a cost saving by the second lowest cost firm is passed on completely. This is in line with the results of the Cournot model: larger firms pass-on less than smaller firms and firm-specific pass-on is less than industry-wide pass-on. Pass-on of cost savings by multiple firms will be complete if and only if the second lowest cost firm is involved. Since the Bertrand model with homogeneous products yields a quite unrealistic situation in

[53] Dornbusch (1987) did a Cournot analysis with symmetric firms in the context of analyzing exchange rate pass-through.

which one firm obtains the whole market, we now turn to the case where products are differentiated.

3.2.3.3 Oligopoly with differentiated products

To illustrate pass-on of firm specific cost savings we again use the logit model of product differentiation, which has the property that products compete symmetrically with each other. We will again focus on absolute rather than relative pass-on in the logit model because this is more convenient.

As before, the condition for profit maximization can be totally differentiated, now varying only the marginal cost of a single firm and holding the cost of the other firms constant. First consider the product-specific pass-on rate under the hypothesis that prices are set by a monopolist:

$$\tau_i^A = \frac{\mathrm{d}CS}{\mathrm{d}c_i} = s_i s_0$$

Product-specific pass-on is simply a fraction of industry-wide pass-on, where this fraction is proportional to the product's market share, i.e. $\tau_i^A / \tau^A = s_i / (1 - s_0)$. Consistent with the findings for the Cournot model, a cost reduction for a single product is thus passed on by less into the price index than an industry-wide cost reduction. However, the product-specific pass-on rate is now *greater* as the product's market share increases.

Now consider the product-specific pass-on rate under the hypothesis that prices are set oligopolistically:

$$\tau_i = \frac{\mathrm{d}CS}{\mathrm{d}c_i} = \frac{s_i (1 - s_i)^2}{(1 - s_i)^2 + s_i} \Big/ \sum \frac{s_k (1 - s_k)}{(1 - s_k)^2 + s_k}$$

This rather complicated formula has again the property that product-specific pass-on is only a fraction of industry-wide pass-on. A cost reduction for a single product is thus passed on by less than an industry-wide cost reduction. However, in contrast to the monopoly case, there is now no longer a simple relationship between the product's market share and the degree of pass-on. To gain some further insights we computed firm-specific pass-on for alternative values of the firm's market share and alternative numbers of competing firms. We assume the competing firms are all identical (so that they have an equal market share. The results of the computations are given in the Table 4. below.

Table 4. Firm-specific absolute pass-on with symmetric competitors

Market share	Number of competing firms: 1	2	3	4	5	1000
0.001	0.000	0.000	0.000	0.000	0.000	0.000
0.100	0.197	0.138	0.120	0.111	0.107	0.090
0.200	0.287	0.222	0.200	0.189	0.183	0.160
0.300	0.327	0.271	0.250	0.240	0.234	0.210
0.400	**0.337**	**0.293**	**0.276**	0.267	0.262	0.240
0.500	0.325	0.292	0.279	**0.272**	**0.268**	**0.250**
0.600	0.293	0.271	0.262	0.257	0.254	0.240

0.700	0.243	0.230	0.224	0.221	0.219	0.210
0.800	0.176	0.170	0.167	0.166	0.165	0.160
0.900	0.095	0.093	0.092	0.092	0.091	0.090
0.999	0.000	0.000	0.000	0.000	0.000	0.000

Reading Table 4 row by row, one can see that a firm with a given market share passes on less as the number of competing firms increases. Intuitively, when there are a large number of firms, competition between these firms is strong and market power is weak, so that they do not respond to price changes initiated by one firm. Reading the Table column by column, one can see the role of the market share of the firm experiencing the cost reduction. Pass-on is very limited when the market share of the firm is either very small or very large. The case of no pass-on under a very small market share corresponds to the case of a price taking firm as under perfect competition (section 3.2.3.1). Intermediate market shares lead to the greatest pass-on.

> **Observation 3.9:** In a product-differentiated industry, product-specific pass-on is less than industry wide pass-on, both under monopoly as under oligopoly.
>
> Market share matters in assessing the magnitude of firm-specific pass-on under product differentiation. The logit model indicates that intermediate levels of market shares yield the greatest product-specific pass-on.

We note that the specific results depend on the assumption of the logit model, which implies that firms have a tendency to adjust their markups when cost changes. In other models, such as the linear demand model or the CES model, it may be expected that higher levels of firm-specific pass-on would obtain. A robust finding would nevertheless be that firm-specific pass-on would be less than industry wide pass-on. Also, it should be clear that the market share of the firm experiencing a cost reduction is a key determinant of firm-specific pass-on; the logit example suggests that pass-on is largest when the firm has an intermediate market share.

It is useful to compare this preliminary analysis on differentiated goods industries with the theoretical results obtained by Feenstra et al. (1996). They study exchange rate pass-through, which is equivalent to pass-on of cost changes. In particular, they investigate the role of the market share of the exporting firms who experience the exchange rate shock (the cost change). We stress that their results are not directly comparable to ours: they treat each individual exporting firm as small and look at the role of the total market share of all exporting firms. They find that under some modest assumptions pass-through is equal to one half if the exporting firm has a very small market share. As the market share of the exporting firm increases, pass-through may initially decrease or increase; yet as the market share becomes very large, pass-through starts to increase and ultimately becomes complete. These results on the role of market share are not necessarily inconsistent with ours. Indeed, as the market share of the exporting firms increases in Feenstra et al.'s model, the exchange rate (or cost) change becomes more like an industry-wide cost change. Their conclusion that pass-on is complete as the market share of the exporting firms becomes very large is then similar to our Observation 3.7 regarding complete pass-on of an industry-wide cost shock.

3.2.4 Summary

This section has reviewed the theoretical determinants of pass-on of cost savings. It is important to bear in mind that one may measure either absolute pass-on or relative pass-on (the pass-on elasticity). We begin by looking at pass-on of industry wide cost savings, i.e. cost savings realized by all firms in the industry. Under perfect competition, the price elasticity of demand and the shape of the marginal cost curve are the main determinants. As consumers become more price inelastic, and as the supply curve becomes more elastic (little capacity constraints), pass-on of industry wide cost savings will be more complete. If supply is perfectly elastic (no capacity constraints), pass-on of industry-wide cost savings is complete under perfect competition. The picture is different under monopoly or oligopoly with market power. In this case, pass-on of industry wide cost savings may be incomplete even if supply is perfectly elastic. The reason is that firms charge markups, which depend on the price elasticity of consumer demand. In the "typical" case, consumers become more price sensitive as the price increases. If this is the case, then firms will absorb cost changes by adjusting their markups and pass-on will be incomplete. This is especially true in the monopoly case. Nevertheless, it is worth stressing that even in the monopoly case (or the dominant firm case) firms pass on at least part of the cost savings onto consumers. As the number of firms increases, pass-on of industry wide cost savings generally becomes more complete.

This section finally emphasized the importance of distinguishing between industry-wide and firm-specific pass-on. An understanding of firm-specific pass-on is especially relevant in the context of mergers, since it are typically only the merging firms who benefit from any cost reduction. Our analysis shows that firm-specific pass-on is generally less than industry wide pass-on. For example, when firms are identical, firm-specific pass-on is equal to industry-wide pass-on divided by the number of firms in the industry. Hence the more firms there are, the lower is the degree of firm-specific pass-on. When firms are not identical, no general results on firm-specific pass-on are available. Nevertheless, it is clear that the market share of the firm realizing the cost reduction matters. On the one hand, a high market share of the cost-reducing firm makes the cost saving close to an industry-wide cost saving and thus makes pass-on more complete. On the other hand, a high market share also means a lot of market power, which may provide an incentive to pass on cost changes incompletely. The results of these forces is that pass-on in an industry is typically the largest for firms with intermediate market share, i.e. not the very small firms but also not the dominant firms.

3.3 Empirical evidence

We do not aim at this point to provide a complete review of the empirical results on pass-on. Instead, we consider some of the most important recent contributions in more detail. This will give the opportunity to explain strengths and weaknesses of alterative approaches in measuring pass-on. Nevertheless, we will also discuss the specific empirical results obtained in various papers. This should be useful to put further results in perspective.

The empirical literature on pass-on of cost savings from mergers is, to our knowledge, virtually non-existent. It is thus necessary to draw lessons from empirical work on pass-on of cost changes that are triggered by other events than mergers. Various fields in economics have, in fact, been concerned with quantifying pass-on. First, there is a large literature in International Economics on exchange-rate pass-through. This deals with the question whether and to which extent exporting firms pass-through exchange rate fluctuations in consumer prices. Related to the exchange rate pass-through literature is the pricing to market literature (Krugman, 1987). The focus is different in that the implications of incomplete exchange rate pass-through on local prices are investigated when exporting firms sell to different markets. Also in International Economics, there has been research on the impact of tariffs on consumer prices. Second, in Public Economics, there has been a great amount of research on tax incidence. A central question here is which agents are most affected by a tax policy, taking into account behavioural responses. The literature relevant for our purposes deals with the incidence of excise taxes. It asks how consumer prices respond to changes in excise taxes. Third, in various specialized applied areas in Economics, there has been considerable interest in the transmission of intermediate goods prices into the prices for final goods. For example, in Agricultural Economics, empirical research has looked at the price transmission of intermediate food products, such as cacao, coffee beans, sugar, etc…; in Energy Economics empirical research considers the transmission of price fluctuations for crude oil, etc.

Exchange rates, tariffs, excise taxes and intermediate goods prices have in common the important properties that they affect the marginal costs of firms, often fluctuate considerably and can be well observed.[54] They thus have the potential to provide useful information on the pass-on of cost changes onto consumer prices. Nevertheless, special care has been taken when drawing inferences. First, it is essential to know whether the measured cost changes are specific to some firms only, or whether they apply to all firms in an industry. In the previous section we emphasised that firm-specific pass-on may differ significantly from industry-wide pass-on. Changes in excise taxes or intermediate goods prices typically affect the costs of all firms in the industry. Exchange rate changes, in contrast, only affect the costs of the foreign firms, leaving the costs of the domestic firms constant.[55] Second, one has to know whether the measured cost change applies to the whole product, or only to part of it. Excise taxes apply to the full product. An exchange rate fluctuation affects the full marginal cost incurred in the exporter’s country, yet have no impact on local distribution or other costs of services. Intermediate goods typically constitute only a fraction of the marginal costs. Third, one should be sure whether one is measuring short run or long run effects of cost changes on consumer prices. If short-run effects are considerably lower than long-term effects, and if adjustment goes quickly, then results on short-run effects may be highly misleading.

[54] Feenstra (1989) goes into a detailed theoretical and empirical investigation of the “symmetry” hypothesis, which states that exchange rate fluctuations, tariff and factor prices should hve the same impact on consumer prices.

[55] This would at least be the case in the short run The long-run, exchange rates may also affect the costs of the domestic firms, because exchange rate changes affect the prices of inputs used by the domestic firms.

3.3.1 Evidence on industry-wide pass-on

3.3.1.1 Excise taxes and pass-on

Several papers have looked at the pass-on of industry wide cost changes by looking at the relationship between excise taxes and consumer prices. A first study in this spirit is by Sumner (1981). He regresses the price for cigarettes in 47 states of the U.S. during 1954-78 on the per unit excise tax and some control variables. The regression thus takes the following form:

$$P_t = \alpha + \beta T_t + \gamma W_t + u_t \qquad (8)$$

where P_t is the consumer price for observation t, T_t is the per unit excise tax, and W_t are control variables. Of our immediate interest is the coefficient β. Sumner interpreted this coefficient also as a measure of the price to marginal cost markup. This interpretation was criticized by Bulow and Pfleiferer (1983) and Sullivan (1985) since it relies on strong functional form assumptions. We instead interpret the parameter simply as a reduced form parameter measuring the extent to which excise taxes are passed on into the consumer price. More precisely, since the variables are expressed in levels, the parameter β measures the pass-on of a unit cost (tax) change on the unit price change. This differs from the pass-on elasticity, which measures pass-on in percentage terms rather than in absolute levels.

Sumner finds a precise estimate of β of 1.074 (with a standard error of 0.013). This means that an increase in the excise tax by, say, 10 cents per unit leads to an increase in the price for cigarettes of 10.74 cents. (Or, since the analysis uses a panel, the price in a state with a 10 cents higher excise tax will be 10.74 cents higher.) Pass-on in absolute terms is thus greater than one. This implies that one can reject the hypothesis of perfect competition under constant marginal cost. To compute the pass-on elasticity, referring to pass-on in percentage terms, one would need to divide the number 1.074 by the price/marginal cost ratio. This number is not known, but it is generally greater than one. It is thus possible that the pass-on in percentage terms is incomplete.

Sullivan (1985) also estimates a reduced form version of (8). He allows for a more flexible specification, since he adds a quadratic term for the excise tax:[56]

$$P_t = \alpha + \beta_1 T_t + \beta_2 (T_t - \overline{T})^2 + \gamma W_t + u_t$$

where $\overline{T}$ is the average tax rate. The effect of an increase in the excise tax by one unit on the consumer price is thus $\beta_1 + 2\beta_2 (T_t - \overline{T})$. Sullivan obtains a parameter estimate of β_1 equal to 1.089 (standard error of 0.26) and an estimate of β_2 equal to 0.0090 (standard error of 0.0035). For example, if the tax in one state is 10 cents higher than the average, the price in that state will be higher by the amount of 10 cents times 1.089+2*0.009=1.107, so it will be 11.07 cents higher. In contrast, if the tax rate is 10 cents lower than the average, then the price will be 10.71 cents lower. Pass-on is thus found to be asymmetric, yet of the same order of magnitude as in Sumner (1981).[57]

[56] He also considers a reduced form demand equation to draw inferences on market power.
[57] See also the preceding studies by Barzel (1976), Johnson (1978), Harris (1983), Wohlegant and Sumner (1985), Sumner and Ward (1981).

Barnet, Keeler and Hu (1995) estimate a structural model of oligopoly to assess the pass-on (“incidence”) of excise taxes. This approach allows them to estimate the cost parameters and demand parameters that potentially explain pass-on. The advantage of this approach is thus that is can unravel the determinants of pass-on. A main disadvantage is that the results are more dependent on the functional form assumptions. Their estimates show that the price elasticity of demand is equal to 0.71, implying that a price increase by 1 percent leads to a demand reduction of 0.71 percent. They also find that manufacturers are subject to increasing returns to scale, and that conduct is consistent with Cournot oligopoly behaviour, indicating a moderate degree of market power. They use their e-parameter estimates to simulate the effects of a 1 cent increase in the excise tax, holding all other variables constant. They find that such an increase leads to an increase in the consumer price in the range of 0.9-1.0 depending on whether it concerns a federal or a state tax increase. This result is due to a combination of two factors. First, marginal costs are declining which induces overshifting of the taxes into consumer prices. Second, firms have market power and adjust their markups downward when taxes increase (and upward when they decrease).

In a structural model of the automobile industry, Verboven (1998) considers the effects of the differential taxation on gasoline and diesel cars on the manufacturers pricing decisions. Verboven shows that the premium charged for diesel cars can be largely explained by higher markups and less so by higher marginal costs for producing diesel cars. Firms thus exercise significant market power on their premium models (though not necessarily on their base models). Simulations on the effects of increasing the tax differential between gasoline and diesel cars show that firms adjust their markups significantly. In other words, although marginal costs are assumed constant, pass-on is small because markup adjustment.

In sum, the studies on cigarette excise taxes show that pass-on of industry-wide cost changes (excise taxes) in the cigarette industry is more than complete in absolute terms. It is asymmetric in that an increase in costs is passed on more than a decrease. It is not possible to draw strong conclusions on pass-on in percentage terms, i.e. the pass-on elasticity. All we can say is that the pass-on elasticity has to be adjusted by the price/marginal cost ratio; it will thus be smaller than the reported range of 1.07-1.09. Nevertheless, it seems that pass-on is rather large in the cigarette industry, which is confirmed by the structural analysis by Barnett et al. The results follow the moderate amount of market power and the possible presence of increasing returns to scale (declining marginal costs). In contrast, in the automobile market, the study by Verboven indicates that pass-on of cost changes is rather limited, at least regarding differential costs between gasoline and diesel cars. This follows from the significant market power firms have regarding the pricing of their premium models.

3.3.1.2 Intermediate goods prices and pass-on

Several papers have looked at how the price for a final consumer good responds to the movements in the price of a single intermediate good, e.g. crude oil in Energy Economics, or unrefined sugar in Agricultural Economics. For simplicity, we assume in the following discussion that the intermediate good is used in a fixed proportion to the other inputs. This means that producing one unit of the final consumption good always requires the use of x units of the intermediate good, where x is called the transformation rate. In contrast to an excise tax, a movement in the price of an intermediate good only affects a fraction of the marginal cost. As explained in section 3.2.2.3, to know the full pass-on elasticity, one needs to adjust the pass-on elasticity by the cost share of the intermediate good. Without such an adjustment, one could falsely conclude that pass-on is low, whereas in fact only the cost share of the intermediate good is low. To illustrate, in unpublished results we estimated the following pass-on regression relating the price of roasted coffee on the price of its main intermediate input, coffee beans:

$$p_t = \alpha + \beta m_t + \gamma w_t + u_t$$

where p_t, m_t and w_t refer to the consumer price of roasted coffee, the price of the intermediate good, coffee beans, and control variables (e.g. time and country dummies). All variables are now in lower case letters, to denote that they are expressed in logarithms. The coefficient β can then be interpreted as a pass-on elasticity, rather than measuring pass-on in absolute levels. Regression results based on either quarterly or annual data for a set of 15 OECD countries since 1970 show that the pass-on elasticity varies between 0.5 and 0.7, depending on the country. This should not, however, be seen as evidence that pass-on of marginal cost savings is incomplete. One rather has to ask what is the share of coffee beans is in the full marginal cost. Taking into account that labour and packaging costs also affect the marginal costs, most industry experts agree that coffee beans constitute slightly more than 60 percent of the marginal costs. If this is true, then one could conclude from the estimates that the pass-on elasticity for industry-wide cost savings is close to one in the coffee markets of OECD countries.

Several studies have looked at pass-on of intermediate goods price fluctuations in levels, rather than in percent.

$$P_t = \alpha + \beta M_t + \gamma W_t + u_t \qquad (9)$$

This is similar to the experiments of Sumner and Sullivan on the pass-on of excise taxes in levels. There is, however, one important difference. An excise tax directly translates into a marginal cost. An intermediate goods price needs to be multiplied by its transformation rate to measure the impact on marginal cost. For example, for the production of 1 kg roasted coffee 1.2 kg of coffee beans is required, so the transformation rate is 1.2.[58] To correctly assess whether pass-on in levels is complete one thus has to compare the estimate of β to the transformation rate (if this is known).

Borenstein, Cameron and Gilbert (1997) consider a version of regression (9) to assess the effect of fluctuations in crude oil prices on gasoline prices. They propose a lag adjustment specification, which distinguishes between short-run and long-run responses to fluctuations in crude oil prices. Our interest is mainly in the long-run response. They find that that an increase in the crude oil price by 1 cent leads to an increase in consumer prices by 0.55 cent in the first two weeks, and a further increase of 0.12 cent in the next two weeks; after 6 weeks pass-on is about 0.71 cents without any further effects at later points in time. Pass-on is asymmetric: decreases in the crude oil prices are passed on to consumers more slowly, but eventually at the same rate. A decrease in the crude oil price by 1 cent has no immediate effect; it leads to a

[58] The transformation rate is 1.2 because water is evaporated during the production process with an amount of 20 percent.

0.29 cent price reduction after 4 weeks, a 0.43 cent price reduction after 6 weeks and a 0.70 cent price reduction only after 8 weeks. Overall, in the long run (after 10 weeks) pass-on is less than one, about 0.81 cent. This is despite the fact that they are looking at pass-on in levels and not in percentage terms, so that one cannot use the fact that crude oil is only one component of the marginal cost of producing gasoline. Furthermore, the transformation rate of crude oil into gasoline is presumably greater than one. One explanation for the incomplete pass-on is that the marginal cost curve is upward-sloping (capacity constraints). As explained in the theoretical section, a reduction in price following a crude oil price decrease would lead to greater demand. To accommodate this demand, a larger production is needed. This could lead to a partially offsetting increase in cost, explaining incomplete pass-on. An alternative explanation would be that consumers become more price sensitive as prices increase (increasing price elasticity of demand). If this is so, then it is optimal to pass-on cost changes incompletely by adjusting markups.

In a related model von Cramon-Taubadel (1997) considers the transmission of producer prices for pork into the wholesale prices in Northern Germany. He also distinguishes between short term and long-term adjustments. Pass-on turns out to be complete in the long run: over a period of about 10 weeks. the absolute margin between the producer and the wholesale price remains constant at 1.30 DM/kg. In the short run, however, pass-on is found to be incomplete: a unit increase in the producers price leads to an increase in 0.58 units during the first week. Furthermore, it turns out that the adjustment process is asymmetric. The response to an increase in the producer price for pork is significantly faster than the response to a decrease. The most relevant message for our purposes is that pass-on of producer price into whole prices is complete in the long-run, i.e. when considering a period of at least 10 weeks. This result is confirmed in various other studies relating producer to wholesale prices in agricultural economics, see e.g. Kinnucan and Forker (1987) for dairy product price transmission or Pick, Karrenbrock and Carman (1990) for the citrus market.

Bettendorf and Verboven (1999) consider the price transmission for coffee beans into consumer prices for roasted coffee. In contrast to the other studies on price transmission in agricultural products, they consider a structural model to explain the causes underlying price transmission. They find that the rather weak relationship between bean and consumer prices follows largely from the fact that coffee beans are only a fraction of marginal costs. Nevertheless, they also show that markup adjustment is present, although it has only a modest importance since market power is limited. Genesove and Mullin (1997) consider a structural model of price setting for refined sugar. Although they do not explicitly consider the pass-on question, their estimation results imply similar results on pass-on as those obtained by Bettendorf and Verboven.

From the analysis of price transmission of intermediate goods prices common to all firms in the industry, we conclude that pass-on in the long run is relatively large, and in many cases complete. In the short-run pass-on may be less complete, yet adjustment takes place relatively quickly, within a period of 10 weeks for the studies we reviewed. Structural studies suggest that the large degree of pass-on follows from the relatively low degree of market power in the industries considered.

3.3.2 Evidence on pass-on by a fraction of the firms

3.3.2.1 Exchange rate pass-through in general

The large empirical literature on exchange rate pass-through also provides interesting evidence on pass-on of cost savings. This literature aims to measure by how much foreign firms pass on exchange rate fluctuations into the import price charged to local consumers. Understanding exchange rate pass-through is very relevant for our purposes, since a fluctuation in the exchange rate may be viewed as a change in the marginal costs of the foreign firm. Note that an exchange rate fluctuation only affects the cost position of the foreign firms. An exchange rate fluctuation should thus no longer be viewed as an industry-wide cost shock in contrast to the above discussed studies considering the effects of taxes and fluctuations in intermediate goods prices.

To analyse exchange rate pass-through, a common approach has been to regress the local consumer price to the exchange rate of the foreign firm, after controlling for other variables:

$$p_t = \alpha + \beta e_t + \gamma w_t + u_t \qquad (10)$$

In this specification, p_t is the logarithm of the consumer price at time *t*, expressed in the local currency of the importing country. The variable e_t is the logarithm of the exchange rate, measured in units of the local currency per unit of the foreign firm's currency. An increase (decrease) in e_t thus means an appreciation (depreciation) of the foreign firm's currency relative to the currency of the importing country. It may equivalently be viewed as an increase (decrease) in the marginal cost of the foreign firm. The vector w_t is a control vector of cost or demand shifters and u_t is a disturbance term.

In its simplest form the parameters α, β and γ can be estimated through ordinary least squares (OLS). Our main interest is the parameter β, which measures the extent of exchange rate pass-through, or the pass-on elasticity as discussed before in the theoretical analysis. For example, if β=0, then a foreign firm does not pass on any of the exchange fluctuations in the local consumer prices. In this case, we have local currency price stability or zero pass-on of cost changes. In contrast, if β=1, then a foreign firm would increase its price by 10 percent whenever its currency appreciates by 10 percent; and vice versa for a currency depreciation. In this case, we have complete exchange rate pass-through or full pass on of changes in marginal costs.

The empirical literature has considered various alternative forms for regression (10). One alternative is to estimate the regression in first differences of the included variables:

$$\Delta p_t = \alpha + \beta \Delta e_t + \gamma \Delta w_t + u_t$$

This regresses the change in local consumer prices on the change in the value of the foreign firm's currency and a change in the control variables.

More recently, authors have estimated "error correction" versions of the model. Such regressions take into account that responses to exchange rate fluctuations may not necessarily be instantaneous. One can then distinguish between exchange rate pass-through in the short term and in the long term. Because of adjustment costs, one may expect exchange rate pass-through to be larger in the long term. One simple version of an error correction model is:

$$\Delta p_t = \beta^S \Delta e_t + \gamma^S \Delta w_t + \varphi(p_t - \alpha - \beta e_t - \gamma w_t) + u_t$$

The parameters β^S and γ^S measure short term responses of prices, whereas β and γ measure the long term responses.[59]

Among the most interesting studies in the spirit of the above regressions are the contributions by Kreinin (1977), Woo (1984), Hooper and Mann (1989), Feenstra (1989) and Athukorala and Menon (1994). All studies estimate some form of the above regressions, to find out the degree of exchange rate pass-through, after controlling for other factors.

Kreinin (1977) considers exchange rate pass-through to 8 countries, using detailed commodity data. He estimates that exchange rate pass-through by foreign firms to U.S. import prices is 50 percent. In comparison, pass-through by foreign firms to import prices is 60 percent in Germany, 70 percent in Japan, 90 percent in Canada and Belgium, and 100 percent in Italy. One interpretation of the differences across the importing countries (which is also advanced by Kreinin) relates to differences in the size of the countries. For a large country like the U.S., the market share of the foreign firms is likely to be small. An exchange rate shock thus only affects the marginal costs of a relatively small fraction of the firms. In contrast, foreign firms are likely to have a significant market share in smaller countries like Canada, Belgium or Italy. An exchange rate shock then affects the marginal costs of a larger fraction of firms. In an extreme case, when there are only foreign firms, an exchange rate fluctuation would affect the marginal cost of all the firms in the market, so that one may speak of an "industry-wide cost change". The fact that pass-on is smaller into the U.S. than into Belgium, Canada or Italy, is thus consistent with the hypothesis that firm-specific pass-on is smaller than industry-wide pass-on.

In our theoretical analysis we discussed that incomplete pass-on of cost changes can be explained by either markup adjustment, or by adjustments in marginal costs if these are not constant. Since Kreinin does not include direct controls for marginal cost changes, one cannot determine whether the observed incomplete pass-through can be explained by markup or marginal cost adjustment. Woo (1984) studies exchange rate pass-through by foreign firms exporting to the U.S. He estimates pass-on to be between 40 and 75 percent, in line with the results by Kreinin. He attributes the incomplete pass-through fully to markup adjustments, since he controls for changes in the marginal costs.

Hooper and Mann (1989) control for cost changes more completely, by accounting for foreign labour and materials costs, foreign capacity and U.S. domestic costs. They also try to distinguish between long run and short run effects using an error correction approach. They find that the long run exchange rate pass-through elasticity by foreign firms in the U.S. is 60 percent. Because they control for costs, they interpret this incomplete exchange rate pass-through as markup adjustment. Interestingly, they also impose the symmetry restriction. The symmetry restriction says that a change in all factor prices (labour, materials) by, say, 10 percent, should have the same effect on

[59] We note that to estimate an error correction model, one has to check whether there is a co-integrating relationship between the (nonstationary) variables.

local consumer prices as a change in the exchange rate of the foreign firm by 10 percent. Intuitively, this is because a change in *all* factor prices has the same effect on marginal costs as a change in the exchange rate, and should therefore also lead to the same degree of pass-on. They find that imposing the symmetry restriction does not affect the estimated pass-on elasticity of 60 percent.

Feenstra (1989) investigates the symmetry restriction further. He tests whether a change in all factor prices has the same effect on local consumer prices as a change in the exchange rate and as an import tariff. Intuitively, also an import tariff has the same effect on marginal costs, and should therefore have a symmetric effect on local consumer prices. In contrast to the previous studies, Feenstra considers disaggregated data on individual industries. The symmetry hypothesis is tested for Japanese firms selling into three U.S. markets: trucks, cars and heavy cycles. With respect to exchange rate changes, Feenstra finds a pass-on elasticity between 0.52 and 0.63 for trucks, between 0.7 and 0.8 for cars and between 0.7 and 1.1 for heavy cycles. With respect to tariffs, the pass-on elasticity is 0.57 for trucks, and between 0.95 and 1.39 for heavy motor cycles (no tariffs on cars). Feenstra finds statistical support for the symmetry hypothesis. Imposing the symmetry hypothesis[60], Feenstra obtains estimates of the pass-on elasticities of 0.58 for trucks, 0.73 for cars and between 0.97 and 1.27 for heavy cycles. It is instructive to consider the interpretation provided by Feenstra for his differential findings regarding the pass-on elasticity with respect to tariffs. For trucks, the tariffs only applied to Japanese models produced in Japan, with a relatively small market shares. In contrast, for heavy cycles, the tariff also applied to U.S. based production by the Japanese companies. Furthermore, the Japanese firms have a modest market share in the U.S. truck market, and a strong market share in heavy cycles, the only U.S. competitor being Harley-Davidson. Once again, this indicates that it is important to distinguish between firm-level and industry-wide pass-on.

Another study that looks at more disaggregate data is by Athukorala and Menon (1994). They consider exchange rate pass-through in various sectors, in particular textiles, chemicals, metal products, general machinery, electrical machinery, and transport equipment. They also take into account that an exchange rate fluctuation affects the marginal costs because inputs need to be imported at a different cost to the exchange rate change. The following table shows the pass-on coefficients for the various sectors, after controlling for changes in factor prices for imported inputs:[61]

Table 5. Pass-on elasticities for Japanese firms in various industries, 1980-1992.

Industry	*Export share in 1985*	*Pass-on elasticity*
Textiles	3.37	0.140
Chemicals	5.47	0.483
Metal products	10.82	0.678
General machinery	21.13	0.461

[60] Feenstra also imposes a homogeneity restriction.

[61] Athukorala and Menon prefer to call these coefficients pricing to market coefficients, since they take out exchange rate effects stemming from changed factor prices for imported inputs. For our purposes, the exchange rate pass-through coefficient that includes factor price effects is not relevant. The differences between both measures are generally small.

Electrical machinery	15.95	0.549
Transport equipment	34.23	0.605
Miscellaneous products	9.03	0.566
Total Manufacturers	100.00	0.788

Source: From Table 2 of Athukorala and Menon (1994).

First, Athukorala and Menon note that the aggregate pass-on elasticity of 0.788 is larger than the average pass-on elasticity across the industries, which is 0.47. From this they suggest that there is an upward bias in the pass-on elasticity due to aggregation. Second, going to individual sectors, they argue that the low pass-on elasticity for textiles, compared to the other sectors, is due to the limited market power of the Japanese firms in the textile industry.

3.3.2.2 Exchange rate pass-through due to markup adjustment

A new approach to estimating incomplete exchange rate pass-through was proposed by Knetter (1989). He criticises the previous approaches because they only partially control for cost changes that coincide with the exchange rate fluctuation. According to Knetter, the previous approaches cannot very well identify whether pass-on is incomplete because of markup adjustment, or because of cost changes that coincide with the exchange rate change. Consider for example an exchange rate appreciation of a foreign firm's currency by 10 percent. This has the direct effect to make the firm less competitive, similar to an increase in its marginal cost by 10 percent. However, two additional cost effects may occur. First, an appreciation makes the imported inputs cheaper, so that marginal cost decreases. Failure to account for this would lead to an estimate for the pass-on elasticity that is biased downward. This was in fact shown to be the case by Athukorala and Menon (1994). Second, if an appreciation is at least partly passed on to consumers by raising import prices, then demand decreases. Only if marginal costs are flat this has no further implications for costs. If, however, marginal costs are upward sloping, then the demand reduction also means a reduction in marginal costs, which partly compensates for the initial cost reduction.

To measure costs accurately, Knetter proposes to consider exchange rate pass-through behaviour by an exporting firm selling into multiple local markets, as in the following regression:

$$p_{it} = \alpha_i + c_t + \beta_i e_{it} + \gamma w_{it} + u_{it}$$

This regression model differs from the regression (10) in that now data are used for multiple local markets, as indicated by the additional subscript i. One thus uses a panel data set of countries and years. The term α_i is a fixed market effect to control for the fact that prices in one country may be persistently higher. The marginal cost c_t is unobserved, but it is assumed that it is common across all local destination markets to which the firm exports. Marston considered a two-country version of this model, and used the price in the country of production as a measure for the marginal costs in the local destination market. Knetter proposes to estimate the model for multiple local markets, and estimate the common cost effects c_t as a fixed effect using time dummies. The regression model then is:

$$p_{it} = \alpha_i + \alpha_t + \beta_i e_{it} + \gamma w_{it} + u_{it}$$

or in first differences:

$$\Delta p_{it} = \alpha_t + \beta_i \Delta e_{it} + \gamma \Delta w_{it} + u_{it}$$

The time effects thus aim to control for annual changes in marginal cost common for all countries. If one uses these fixed effects, then the coefficient β_m may be interpreted as a pass-through coefficient, after holding marginal costs constant; it thus purely aims to capture markup adjustment following exchange rate fluctuations. A coefficient for β_m equal to 1 does thus not mean that exchange rate pass-through is complete; it only means that markups are not adjusted after exchange rate movements. In contrast, a coefficient for β_m equal to 0 means that markups are fully adjusted after an exchange rate movement.

The advantage of the panel data approach proposed by Knetter is thus that it is possible to identify one key component of pass-on: the possibility that firms adjust their markups.[62] Without such an identification, it is not clear whether one should explain incomplete pass-on by markup adjustment or by the presence of an upward sloping marginal cost curve.

Knetter (1993) applies his approach to study a rather large set of import markets and various industries. Knetter thus also aims to bring a synthesis of the existing empirical evidence using his methodology. One may summarize his findings as follows. First, there is a tendency by firms to pass-on incompletely by adjusting markups in response to exchange rate fluctuations. This tendency is stronger for Japanese, German and U.K. firms than for U.S. firms, with pass-on elasticities of respectively 0.52, 0.64, 0.64 and 1. Second, incomplete pass-through stemming from markup adjustment depends on the specific industry. For example, consider industries that have a common classification across all countries. Pass-on is complete for bourbon, 84 percent for titanium dioxide, 40 percent for synthetic dyes and film, 25 percent for paper, and absent for aluminium foil. One can conclude from this work that an important source of incomplete pass-on of cost savings is in fact markup adjustment. Thus even an industry with constant marginal costs, it is possible that firms will pass on cost changes incompletely into consumer prices. The exact magnitude will, however, depend on the specific industry. In a subsequent paper, using data on the same industries, Knetter (1994) tests whether pass-on is asymmetric, i.e. different for exchange rate appreciation than depreciation. He does not find evidence on this possibility.

In subsequent work, Gagnon and Knetter aim to distinguish between short-term and long-term adjustments of markups in response to exchange rate changes, following an error correction approach.[63] In a simple form, one may write the error correction model as:

$$\Delta p_{it} = \beta_i^S \Delta e_{it} + \gamma^S \Delta w_{it} + \varphi\left(p_{it} - \alpha_i - \alpha_t - \beta_i e_{it} - \gamma w_{it}\right) + u_t$$

[62] Marston provides an early approach in the spirit of Knetter. He controls for cost changes by looking at the ratio of the local import price to the foreign export price.

[63] If the variables are non-stationary, then significant parameter estimates may be a consequence of spurious correlation. One then has to test whether there is a co-integrating relationship between the variables. If so, one can sensibly apply an error correction model to distinguish between a short-run and long run relationship.

They apply their analysis using data on Japanese, German and U.S. automobile exporters, in three engine size categories (less than 1 liter, 1-2 liter, and greater than 2 liter). They find that pass-on due to markup adjustment is especially low in the short run due to adjustment costs (price rigidities). In the long run, pass-on tends to be more complete, especially by U.S. firms. Japanese firms have pass-on elasticities in the range of 0.09-0.12, meaning that they tend to pass-on only about 9 to 12 percent of an exchange rate fluctuation into final consumer prices. This is true for all three engine size categories, and most local import markets. German firms tend to pass-on more. In the low and medium size engine categories, pass-on is usually 50 percent or more. In the large size engine category, pass-on is close to being complete, with even overshooting to the import markets of France and Sweden. U.S. firms show an even larger tendency to pass-on exchange rate movements, with elasticities ranging between 0.80 and 1.2. Gagnon and Knetter interpret the different results regarding the three source firms, based on the type of demand they face. Japanese exporters attract only few consumers with high reservation prices (high willingness to pay). This implies that an increase in the local price to compensate for a Yen appreciation makes most consumers more price sensitive (higher price elasticity). This forces Japanese firms to adjust their markup. German and U.S. exporters are more active in niche markets with few substitutes and many consumers with high reservation prices. They can more easily raise prices after an appreciation without making consumers more price sensitive.

Gosh and Wolf (1994) also emphasize the importance of distinguishing between the short and the long run. They look at weekly prices set for the magazine the Economist. If one does not make a distinction between the short and the long run, then one may obtain the “spurious” result that pass-on is close to zero, since adjusting to new conditions takes time.

Feenstra, Gagnon and Knetter (1996) bridge the gap between empirical studies that consider industry-wide pass-on (section 3.3.1) and studies that consider fraction by only a fraction of (exporting) firms. In particular, they consider the role of the market share of the foreign firms in determining pass-on in the long run, using data on the automobile industry. They find that exporting firms with a very small market share pass on about 70 percent of an exchange rate fluctuation by adjusting markups. As market share of the exporting firms increases, the pass-on elasticity decreases, to reach a minimum of about 0.30 when the market share is about 40 percent. When the market share increases beyond 40 percent the pass-on elasticity increases and becomes roughly complete when the exporting firms have 100 percent of the market. These findings are very intuitive and consistent with the theory on pass-on of cost changes. If the market share of the exporting firms is 100 percent, then the exchange rate fluctuation is essentially an industry-wide cost shock. One may then expect complete pass-on if the marginal cost curve is flat, and if firms act rather competitively. In contrast, if the market share of the exporting firms is small, then an exchange rate fluctuation corresponds more to a firm-specific cost shock.[64] In this case, pass-on is expected to be smaller.

[64] Note however that even when the exporting firms have a small market share, the exchange rate fluctuation cannot be viewed as a firm-specific cost shock, since there may still be several competing exporting firms.

Schembri (1989) analyses pass-on of exchange rate fluctuations for a Canadian export industry (using confidential data) selling an intermediate good to the U.S. and Canada. In contrast to the previous studies on exchange rate pass-through, Schembri explicitly estimates the cost and demand parameters of the export industry, from which the pass-on elasticity can then be computed using formulas similar to the ones we provided in section 3.2. This approach is feasible because there are more data than simply prices and exchange rates. Good cost data (factor prices and factor demands) and demand data (sales) are also available. Schembri reports that the pass-on elasticity by the Canadian export industry to the U.S. is approximately 0.85. This incomplete pass-on is partly due to the fact that marginal costs are not flat but increasing, and partly because of markup adjustment. Unfortunately, we do not know what is the market share of the Canadian export industry in the U.S. market.

In a study of the U.S. automobile market, Goldberg (1995) estimates a model with product differentiation, the nested logit model, which classifies the market in various segments, according to class and country of origin. Using the estimated demand and cost parameters, Goldberg simulates the pass-on of exchange rate fluctuations into consumer prices for various exporting firms selling into the U.S. She computes exchange rate pass-through for Japanese and German firms, following both appreciations and depreciations. She distinguishes between the time when the Japanese firms where capacity constrained due to import quotas, and the time when they were unconstrained. The following summarizes the results:

Table. Pass-on by Japanese and German firms, with and without quota constraints

	Without Japanese quota		With Japanese quota	
	$ appr.	$ depr.	$ appr.	$ depr.
Jap. Subcompact	.155	.250	.015	.01
Jap. Compact	.18	.30	.010	.02
Jap. Trucks	.195	.245	.235	.255
Ger. Intermediate	.700	1.21	.600	1.20
Ger. Luxury	.755	1.01	.551	1.11
Ger. Sports	.65	.805	.60	1.06

The first finding from this table is that Japanese firms typically pass-on a small amount, and considerably less than German firms. This is consistent with e.g. Knetter's study who uses a different methodology. Japanese firms cannot raise (lower) the prices in the U.S. in response to a cost increase (decrease) induced by the exchange rate, and need to adjust markups. German firms have a higher ability to adjust their prices, since they are able to adjust their markups by less. The second finding is that pass-through is asymmetric: a cost decrease induced by an appreciation of the dollar is passed on less than a cost increase induced by a depreciation of the dollar. The third finding is that pass-on by the Japanese firms is much smaller when they where subject to import quota constraints. This is very intuitive: an import quota constraint means that Japanese firms are capacity constrained, i.e. they face a steeply increasing marginal cost curve. Under these conditions, any pass-on of cost changes has to be small since supply cannot be adjusted to demand responses.

Goldberg and Verboven (1998) consider the role of exchange rates in the European automobile market. They find evidence of incomplete pass-on, with pass-on elasticities of about 0.5. They show that 1/3 of the incomplete pass-on can be explained by markup adjustment. About 2/3 of the incomplete pass-on follows from the fact that exporting firms also have important local costs (transportation, distribution, marketing) which do not depend on the exchange rate fluctuations of the exporting firms. Abstracting from the presence of local costs, one can thus say that pass-on is relatively large, i.e. markups are only adjusted moderately. This is intuitive, since an exchange rate fluctuation of, say, the pound, affects the competitive position of all exporting firms selling into the U.K.

3.3.3 Firm specific pass-on

Studies that have looked at firm-specific pass-on of cost savings are scarce, especially when compared to the relatively large literature on pass-on of industry-wide cost changes (taxes, intermediate goods prices) and on pass-on of cost changes for a fraction of firms (exchange rates). In the following, we discuss two articles, one reduced form approach to assessing firms specific pass-on, and one structural approach.[65]

Ashenfelter, Ashmore, Baker and McKerman (1998) propose a methodology to estimate firm-specific pass-on as opposed to the pass-on of an industry-wide cost change. They apply their approach to the Staples/Office Depot merger, a prominent U.S. case in applying quantitative analysis.[66] Their analysis was a response to the merging firms' expert who had asserted, without providing empirical evidence, that Staples reduced price by two-thirds of any cost reduction (see Baker, 1998). In essence their approach amounts to regressing the price an individual firm charges on both its own costs and the costs of other firms in the industry. Consider the following simplified regression, using the notation we introduced in section 3.3.1:

$$p_t = \alpha + \beta_1 m_t^F + \beta_2 m_t + \gamma w_t + u_t$$

where m_t^F measures the cost of the individual firm, i.e. Staples; m_t measures industry-wide costs; and w_t captures control variables. Industry-wide cost m_t is approximated by the cost of Staples' competitor Office Depot. All variables are in logarithms, so the parameters may be interpreted as elasticities. The coefficient β_1 measure the pass-on of a cost change specific to the firm, controlling for industry-wide cost changes, which are measured by the coefficient β_2. Ashenfelter et al. also consider a restricted

---

65 The residual demand literature, in particular, Baker and Bresnahan (1985); Spiller (1987), has also offered results that are of related interest. A detailed discussion would however lead us too far.

66 The case is not only known for its careful analysis of assessing pass-on. Also a detailed analysis on market power effects was conducted. This was done by showing that Staples prices were significantly lower in cities where Staples competed with Office Depot, than in non-competitive zones, where there were no other superstore chains. This market power analysis indicated that prices could increase by 9 percent in markets where there was competition.

specification in which β_2 is set equal to zero, to see what bias would result if one does not control for industry-wide cost savings.

The data are provided by Staples and Office Depot. They are average monthly prices and variables costs for 30 identical products sold during the years 1995 and 1996 in various stores. The products were 17 pens, 7 paper items, 5 toner cartridges and 1 computer diskette. Store, product and time dummies are included to control for price variation due to differences across stores, products and months.

In the regression where β_2 is excluded, they find a pass-on elasticity of 0.571 (t-value of 194). This relatively high number does not distinguish between firm-specific and industry-wide cost changes. In the regression where β_2 is included they find a much lower elasticity of 0.149 (t-value of 38). It implies that when Staples costs fall by 10 percent, without a change in the rivals costs, then Staples would lower price by roughly only 1.5 percent. The difference between the two regression models demonstrates that the bias from estimating firm-specific pass-on without controlling for industry-wide cost changes can be quite large. Ashenfelter et al. also consider whether there was any asymmetric pass-on of cost changes (i.e. increases versus decreases), yet they did not find evidence on this.

Kadiyali (1997) studies strategic pricing and exchange rate pass-through in the U.S. photographic film industry. There are two main competitors: Kodak and Fuji. A structural econometric model is constructed, in which the demand and cost are directly estimated. Kadiyali then simulates the effects of changes in the Yen/$ exchange rate on the prices set by Fuji. Since there are only two firms, one can view his simulation exercise as an analysis of firm-specific pass-on, in contrast to all the other articles on pass-on reviewed above. Kadiyali finds that the firms-specific pass-through elasticity for Fuji is relatively low. During the period 1980-1984, when it had an average market share of 7.17 percent, the pass-on elasticity was 0.076. During the period 1985-1990, when its average market share increased to 15.59 percent, the pass-on elasticity was 0.178.

3.3.4 Summary

In this section we have started by reviewing the rich empirical literature on tax incidence, intermediate goods price transmission and exchange rate pass-through. The empirical results vary sometimes substantially from sector to sector. Yet it is still possible to make some empirical generalizations. It seems fair to say that the literature on the effects of excise taxes and intermediate goods prices finds that pass-on is close to 100 percent, at least when one considers a sufficiently large time horizon (10 weeks or more). The literature on exchange rate pass-through tends to find incomplete pass-on, of an order of magnitude of 60-70 percent. Part of this literature also relates the extent of exchange rate pass-though to the market share of the exporting sector. Finally, the scarce literature on firm-specific pass-on finds a relatively low degree of pass-on, especially when the market share of the firms is small. The estimates are in the range of 10-20 percent.

This evidence makes clear the empirical importance of considering firm-specific rather than industry-wide pass-on. While industry-wide pass-on is more or less complete, firm-specific pass-on may be substantially smaller. At the same time, however, it is not sensible to draw strong general conclusions about the extent of firm-specific pass-on. The extent of firm-specific pass-on is likely to vary substantially from merger to merger, and it seems unreasonable to aim for a general presumption on the extent of firm-specific pass-on. In particular, there is a central role for market share in explaining firm-specific pass-on. Low firm-specific pass-on may be due to low market shares; yet in those cases the market power effects from mergers may also be low.

Our empirical review does not only stress the importance of assessing pass-on on a case-by-case basis; it has also introduced several possible methodologies for assessing pass-on of cost savings. We hope our analysis makes it clear that the methodology lends itself quite well for a fast implementation, as is required in merger cases. This is especially true for the reduced form analysis, yet also structural approaches could be feasible in some cases. The major potential constraint to assess pass-on in merger cases is thus not necessarily in the application of sophisticated technical analysis. Rather, it is important to make sure that the data required for the analysis are collected from the parties in an efficient way.

4. AN APPLICATION OF PASS-ON

In the previous chapter we have shown that the degree of pass-on depends on a variety of factors. For example, one cannot make general claims that pass-on of cost savings will be complete, unless one can be sure about certain market conditions (such as the absence of capacity constraints). Similarly, one cannot make a general claim that a monopolist (or a dominant firm) would not pass on any cost savings; to the contrary, it is likely that even a monopolist will pass on at least part of its cost saving to consumers. The extent of pass-on thus ultimately becomes an empirical matter. The previous chapter showed that there are several general tendencies, confirming the theory. An example is that that firm-specific pass-on is typically less than industry-wide pass-on. Nevertheless, it is also clear that one cannot hope to have completely general empirical conclusions on the extent of pass-on, applicable to any firm in any industry. One is thus left with the task to empirically measure the extent of pass-on.

One approach would be to try to quantify all the relevant structural market conditions, for example the price elasticity of demand, the extent of capacity constraints, etc… While such an approach may be feasible in certain applications, it is often impractical, especially in light of the time frame that competition agencies have to respect. The goal of this chapter is to illustrate that there is a practical empirical alternative to measure pass-on of cost savings in competition cases. The approach can be based on data that are often readily available to the firms, and can be carried out within a reasonable time using econometric techniques.

Using data for refrigerated juices, we will illustrate how one can practically estimate the extent of pass-on using standard econometric techniques. The econometric framework further elaborates on the framework of Ashenfelter et al. (1998). They have used it to assess firm-specific pass-on in the Staples/Office Depot merger, as discussed above. We show how the framework can be extended to also consider the multiproduct nature of firms. Yet to estimate firm-specific pass-on, we would also need data on competing firms.

We begin by outlining the required data and the actual data in section 4.2. We next discuss the econometric framework in section 4.1. Finally, we present and interpret the empirical results in section 4.3.

4.1 The data

4.1.1 Data requirements

Price and cost data

Whether one measures absolute pass-on or relative pass-on (i.e. the pass-on elasticity), two key types of data are required: price data and cost data. The price data refer to the price for the products sold by the merging firms. The cost data in principle refer to marginal cost, though in practice variable costs may be a reasonable proxy.

In merger analysis, it is important to focus on firm-specific pass-on, rather than industry wide pass-on, since cost efficiencies are typically only realized by the merging firms. In the previous chapter it was shown that firm-specific pass-on is typically lower than industry-wide pass-on. We also discussed there the framework of Ashenfelter et al. to measure firm-specific pass-on. The required data for their approach are price and cost data for several firms in the industry, for example the two merging firms. We do not have this information. Instead, we show how one can easily extend the Ashenfelter framework to account for the multiproduct nature of firms. We also show that it is important to focus on the price level in the whole industry, rather than on the prices of the individual products.

Sales data

In addition, as will be shown in section 4.2, it will be useful to have sales data. With sales data, it is possible to summarize the firm-level price data to an industry price index. This may be more relevant from a competition policy point of view, especially if the price index is closely related to measures of consumer surplus. Note that the pass-on measures in the previous chapter were often closely related to consumer surplus or the industry price index.

The size of the data set

Given price and cost data (and possibly sales data), the remaining question is how many of these price and cost observations are required. In principle, one could follow the prices and costs over an extended period in time. For example, one could collect annual price and cost data over a period of 30 years. Yet in practice, it is probably asking too much to obtain the price and cost information dating back thirty years. When information is available over a more limited historical period, the number of observations to identify pass-on can be increased by increasing the frequency of the observations. For example, one could collect quarterly data on prices and cost over the last 10 years, to have 40 observations to identify pass-on. Similarly, one could gather monthly data over three years and have 36 observations on prices and costs. One could increase the number of observations even further by having a higher frequency, say weekly, but in practice, the informational value of this would depend on the actual variation of price and cost data in short intervals. For example, costs and prices for gasoline prices show a high variation, even within a month, so that a high frequency of data may provide useful additional information. In contrast, the prices for cars are typically adjusted by the manufacturers only a few times a year, so that there is probably little to be gained from increasing the frequency beyond the quarterly level.[67] Note also that when one increases the frequency of the data, it may become useful to specify a dynamic model, which distinguishes between short-run and long-run pass-on.

In addition to following the price and cost observations over a certain time period, there is the possibility of *pooling* the data. To see how this works, note that data are often available at a finer level of detail than at the level of the firm. A firm may sell different products, in different stores. Suppose that each firm operates in 3 different product categories and sells in 10 different stores. One could then collect historical price and cost data to estimate pass-on for each product in each. This would amount to 30 (3 times 10) separate pass-on studies. But alternatively, one could pool the information on products and stores and construct a large panel data set in three dimensions: time, stores and product categories. If one is willing to make certain assumptions, for example that the degree of pass-on is the same across stores and/or product categories, then one can conduct a single large study of pass-on instead of 30 small ones. By pooling the data, one can thus also try to identify pass-on by comparing price and cost observations across stores and products, in addition to following the evolution of price and cost over time. This can greatly increase the amount of information, so that one does not necessarily have to go back very far in time. In our example of 3 product categories and 10 stores, the number of observations can be multiplied by a factor of 30 when one pools the data. For pooling to be effective it is necessary that prices and costs show enough variation across stores and product categories, in addition to the previous requirement of sufficient variation over time.

To summarize, a suitable data set to estimate the extent of pass-on would contain the following variables for several firms in the industry: prices, (marginal or variable) costs and (possibly) sales. The information can be collected as a panel data set with the following dimensions: time, stores and product categories. This information should be available for many cases.

4.1.2 Actual data

The panel data set collected by Ashenfelter et al. (1998) consisted of the following.

- There was information on two firms, Staples and Office Depot. Since there was no information on the other firms, the information on Office Depot was treated as a good proxy for all other firms.
- There was information on prices and variable cost.
- The following dimensions of the panel were included:
    - Time: 24 months during 1995-1996

[67] Note that, as the frequency of the data is increased, dynamic models that distinguish short term from long term effects may be necessary.

- Stores: 500 stores
- Product categories, here called "stock keeping units" (SKU): 30 SKUs, including 17 pens, 7 paper items, 5 toner cartridges and 1 diskette.

The total number of observations in the panel on the prices and costs for both firms can thus be up to 360000 (24 times 500 times 30).

The analysis in this chapter has made use of an alternative data set on the market for refrigerated juice. In contrast to Ashenfelter et al. we have only information available on one supermarket: Dominick's Finer Foods, one of the two largest retail chains in Chicago metropolitan area.[68] There are three main brands sold by the retail chain: Tropicana, Minute Made and the private label. There are 146 different stores and the brands are sold in many different varieties (called UPCs).

Our panel data set on the refrigerated juice market thus consists of the following:

- There is information on three brands, Tropicana, Minute Maid and the private label. The other brands have a negligible share of the market and are ignored (only 11 percent of the market for the remaining brands).
- There is information on prices, variable cost (measured by the manufacturer price) and sales.
- The following dimensions of the panel are included:
  - Time: 218 weeks between 1989-1993.
  - Stores: 146 stores
  - Product categories, here called UPCs: about 200.

  The total number of observations in the panel is thus also very large. To keep the data manageable, we followed Kadiyali et al. (2000) and averaged the data across stores and across varieties. This reduces the number of observations to only 218 (the number of weeks). It thus appears that we no longer make use of panel data information, and that all identification of pass-on would become because of variation in prices and cost over time. Yet this is not the case. Because our price and cost data are now averages (across store and UPC), the variation across stores and UPC also plays a role in identifying pass-on.[69] We emphasize that in principle we could also have followed the same approach as in Ashenfelter et al., yet in practice it turned out to be more convenient to follow the "averaging approach".[70] This may also be the case in competition policy cases.

The main difference between the application of Ashenfelter et al. and our current application lies in the ownership structure of the firms. In Ashenfelter et al., two independent firms are chosen that are competing with each other. In our application, three brands are considered. At the retail level, the prices of these brands are set jointly by the retailer, rather than competitively (see Kadiyali et al, 2000). One could

[68] The data set is maintained by the marketing group at the university of Chicago.
[69] In the panel data terminology, we have something resembling a "between-groups" estimator.
[70] Otherwise, it would have been necessary to precisely match the product varieties (UPCs) of the various brands.

thus view our analysis as complementary to Ashenfelter: we do not focus on the extent of pass-on of cost changes by competing firms, but rather on the extent of pass-on of cost savings by brands whose prices are set cooperatively (by the retailer). For both applications, the same econometric framework can be used.

4.2 The empirical framework

Suppose there are three products, product 1, 2 and 3 and we are interested in studying pass-on of cost changes into consumer prices. A "traditional" approach would be to estimate the following basic regression equation for firm 1:

$$p_{ijt}^{1} = \alpha_i + \alpha_j + \beta^1 m_{ijt}^{1} + \gamma w_{ijt} + u_{ijt} .$$

In this specification, p_{ijt}^{1} is the price charged by product 1 for a product sold in store i, belonging to product category j at time t. Similarly, m_{ijt}^{1} is the marginal (or variable) cost incurred by product 1 in store i in product category j at time t. The variable w_{ijt} measures other factors that may affect the price. The terms α_i, α_j, β^1 and γ are parameters that are to be estimated using common econometric regression techniques. For example, the ordinary least squares method (OLS) may be used under certain assumption.

The parameters α_i and α_j are called fixed effects for, respectively, the store i, the product category j. (Ashenfelter et al. also include a time effect α_t .). It is important to include them as parameters to be estimated, since they measure the systematic effect that a store, product category or time period could have on the price. (For example, one product category can be systematically more expensive than the other product category.)

The main parameter of interest is β^1, which measures pass-on. It may either reflect absolute or relative pass-on, depending on how the variables are measured. If the price p_{ijt}^{1} and the cost m_{ijt}^{1} are measured in monetary units, e.g. in Euros, then β^1 measures absolute pass-on: the effect of an increase in the variable cost by one Euro on the price in Euros. In contrast, if the price and cost are measured in logarithms, then β^1 measures the relative pass-on, or the pass-on elasticity.

Ashenfelter et al. showed that the above specification fails to distinguish between firm-specific pass-on and industry-wide pas-on. The reason is that the specification looks at the effects of cost increases by firm 1 on the price charged by firm 1, without controlling for any possible accompanying cost changes by the other firms. In practice, this will lead to biased estimates for firm-specific pass-on.

A more general approach is therefore to modify the regression equation and also include the costs of the other products in the regression, i.e. the variables m_{ijt}^{2} and m_{ijt}^{3}. In Ashenfelter et al these other products were from the merging partner. Here we include other products of the same firm.

Pass-on equation for firm 1's price p_{ijt}^1 *:*

$$p_{ijt}^1 = \alpha_i + \alpha_j + \beta^1 m_{ijt}^1 + \beta^2 m_{ijt}^2 + \beta^3 m_{ijt}^3 + \gamma w_{ijt} + u_{ijt} \qquad (11)$$

More specifically, one can interpret the parameters as follows:

- β^1: This reflects product-specific pass-on: the pass-on of a cost increase by product 1 on the price of product 1's product.

- $\beta^1 + \beta^2 + \beta^3$: This measures product 1, 2 and 3's overall pass-on: the pass-on of a cost increase by product 1, product 2 and product 3 together on the price of product 1's product (if the costs are highly correlated).

As before, the pass-on may be either in absolute terms, or in elasticity form (in relative terms), depending on whether one measures the prices and costs in monetary values (Euros) or in logarithms.[71]

One may easily extend the analysis beyond pass-on by product 1. One can thus also ask how a cost changes will be passed on to the price of product 2 and product 3. It is straightforward to formulate similar regression equations for product 2 and 3. To avoid too much notation, these equations are not shown here.

We now discuss how one can extend beyond the Ashenfelter et al. framework. Instead of asking what will happen to the individual price of product 1, or product 2 or product 3, one may wonder what would happen to a composite price index when the cost of one product (or several products) change. This may indeed be the more natural question to ask if one is interested in consumer welfare questions relating to the industry as a whole. In fact, consumer theory has discussed extensively the choice of an appropriate price index to measure changes in consumer surplus. A detailed discussion on the choice of an appropriate price index is beyond the scope of this study. Yet it is fair to say that there is a consensus that one should include representative basket of related products into the price index, using sales as weights. In the theoretical discussion on pass-on in the previous chapter the emphasis was also on what happens to the price level for the industry as a whole.

Suppose the price index for store i of product category j at time t is measured by the variable p_{ijt}^I. Such a price index is typically some weighted average of the prices of product 1, product 2 and product 3 (p_{ijt}^1, p_{ijt}^2 and p_{ijt}^3), where the weights may be the

[71] Note that this set-up slightly from Ashenfelter et al. They assumed that only the cost changes of the other firms should be interpreted to capture firm-specific pass-on, whereas the current framework also includes the cost change of firm 1 itself.

sales at a certain base period. Using the price index as the relevant focus it is possible to formulate the following variation of the above regression equation:

Pass-on equation for a composite price index p^I_{ijt}:

$$p^I_{ijt} = \alpha_i + \alpha_j + \beta^1 m^1_{ijt} + \beta^2 m^2_{ijt} + \beta^3 m^3_{ijt} + \gamma w_{ijt} + u_{ijt} \qquad (12)$$

The only difference with the previous regression equation is that one now looks at the effects of cost changes on the whole price index, rather than on the price of a single product. We now have the following interpretation of the included parameters:

- β^1: Product 1's specific pass-on into the price index.
- β^2: Product 2's specific pass-on into the price index.
- β^3: Product 3's specific pass-on into the price index.
- $\beta^1 + \beta^2 + \beta^3$: Product 1, 2 and 3's overall pass-on.

The above equations (11) and (12) can be estimated directly using the large panel data set. Alternatively, the data can be averaged across the store and/or the product category. As discussed in section 4.1.2, in our application we average the observations across both dimensions of the panel.[72] The equations (11) and (12) then become:

Pass-on equation for product 1's averaged price $\overline{p}^1_t$:

$$\overline{p}^1_t = \overline{\alpha} + \beta^1 \overline{m}^1_t + \beta^2 \overline{m}^2_t + \beta^3 \overline{m}^3_t + \gamma \overline{w}_t + u_t \qquad (11')$$

Pass-on equation for an averaged composite price index $\overline{p}^1_t$:

$$\overline{p}^I_t = \overline{\alpha} + \beta^1 \overline{m}^1_t + \beta^2 \overline{m}^2_t + \beta^3 \overline{m}^3_t + \gamma \overline{w}_t + u_t \qquad (12')$$

One may of course also estimate the pass-on equations for product 2 and 3's price. These equations will be the regression equations to which the data set is applied in the next section.

---

[72] This can be referred to as a between-groups estimator, as discussed before.

4.3 The results

4.3.1 Base results

We have estimated the pass-on regressions with the variables expressed either in monetary values or in logarithms. The main results are displayed in Table 6.

Table 6. Pass-on of cost changes –
Effect on price of the Private Label, Minute Maid, Tropicana, and the market price index.

		level specification		logarithmic specification	
	variable	estimate	stand. error	estimate	stand. error
Private Label	constant	-0,113	(0,144)	0,279	(0,042)
	cost PL	**1,493**	(0,056)	**1,120**	(0,043)
	cost MM	-0,003	(0,041)	-0,014	(0,030)
	cost TR	-0,044	(0,050)	-0,038	(0,050)
Minute Maid	constant	0,649	(0,168)	0,523	(0,048)
	cost PL	-0,135	(0,065)	-0,121	(0,049)
	cost MM	**1,297**	(0,048)	**0,945**	(0,035)
	cost TR	-0,097	(0,059)	-0,108	(0,058)
Tropicana	constant	-0,354	(0,142)	0,139	(0,035)
	cost PL	-0,038	(0,055)	-0,017	(0,035)
	cost MM	-0,087	(0,041)	-0,055	(0,025)
	cost TR	**1,631**	(0,050)	**1,275**	(0,041)
Market price index	constant	-0,369	(0,157)	0,177	(0,041)
	cost PL	**0,505**	(0,061)	**0,359**	(0,042)
	cost MM	**0,488**	(0,045)	**0,332**	(0,030)
	cost TR	**0,549**	(0,055)	**0,493**	(0,049)

Table 6 shows the estimates of pass-on when cost increases and cost decreases have symmetric effects on the price (as was implicitly assumed in the empirical framework of section 4.2). The first twelve rows show the effects of product-specific cost changes on the price of the private label, Minute Maid, and Tropicana, respectively. They are based on the specifications (11'). The final four rows show the effect of cost changes on the general market price index, based on the specification (12'). The first two columns show the estimates (and standard errors) when all variables are expressed in monetary values; the final two columns show the estimates when the variables are expressed in logarithms.

Most relevant for our purposes are the numbers printed in bold. They measure the product-specific pass-on rate. The other numbers refer to the parameters estimates for the control variables, namely the competitors' costs. They are not of immediate interest, but as was discussed in the previous section, it is necessary to include them as control variables, since otherwise the parameters of our main interest will be biased.

We begin by discussing the pass-on regressions for the private label, Minute Maid and Tropicana, as shown in the first twelve rows. These specifications are most comparable to the one considered by Ashenfelter et al., especially the logarithmic specification. Consider first the regression for the private label (first four rows), with the variables in monetary values (levels). It can be seen that a cost increase for the private label by one unit would lead to a price increase by 1.493 units.[73] This is thus the absolute pass-on, which is estimated with a high precision as can be seen from the low standard error. The effect of the rival firms costs is negligible (insignificantly negative). The logarithmic specification gives an estimate of the relative pass-on estimate, or elasticity, namely 1.120. It says that a cost increase for the private label by one percent would lead to a price increase of 1.12 percent, after controlling for cost changes for the competing firms.

The pass-on estimates for Minute Maid and Tropicana are shown by the bold numbers in the next eight rows. The absolute pass-on numbers for Minute Maid and Tropicana are estimated to be 1.297 and 1.631, respectively. The relative pass-on numbers, or the pass-on elasticities, are estimated to be 0.945 and 1.275, respectively.

One can conclude from these numbers that product-specific pass-on of cost changes is considerably larger in this market than in the market considered by Ashenfelter et al. For example, the product-specific pass-on elasticities are around one, so that pass-on is roughly complete. This compares to the product-specific pass-on elasticity of around 0.15, as had been obtained by Ashenfelter et al. One interpretation for this finding is the different focus of our analysis. Ashenfelter et al. took into account also cost changes by competing retail chains. In contrast, the analysis here only focuses on competing brands within the same retail stores. If the prices are set more or less cooperatively within a retail store (as in Kadiyali et al.), then a higher degree of pass-through would be reasonable, and also consistent with the theoretical results derived in the previous chapter. Nevertheless, it is possible that the estimates would go down if the analysis could also control for the cost increases by a competing retail chain (which was not feasible in our application because of limited data).

How can one interpret the results in light of the theoretical discussion of chapter 3.2? To address this question, it is necessary to consider product-specific pass-on with respect to the market price index, rather than with respect to the individual prices. In our opinion, this approach is also more appealing from a competition policy point of view, at least if the market price index is a reasonably good approximation for consumer surplus. The estimates are shown in the last four rows of Table 6.

Looking at the extent of absolute product-specific pass-on, one can see that a cost increase for the private label by one unit would lead to a rise in the market price index by 0.505 units. A cost increase by one unit for Minute Maid would lead to a slightly lower rise in the market price index (0.488 units), whereas a cost increase by one unit for Tropicana would have a larger effect (0.549 units). The product-specific pass-on estimates expressed in relative terms, or the elasticities, are lower. A 1 percent cost increase for the private label would lead to a general price increase by 0.359 percent, whereas a 1 percent cost increase for Minute Maid would lead to a general price increase by only 0.332 percent, and a 1 percent cost increase for Tropicana would lead to a general price increase by 0.493 percent.

If cost shocks are highly correlated, the sum of these pass-on rates is a rough measure of the industry-wide pass-on. In absolute terms this amounts to a number of 1.542, whereas in relative terms it amounts to 1.184. This says that a joint cost change by 1 unit or by percent would lead to a change in the market price index by respectively

[73] Note that the effect of other product cost changes is negative, though it is insignificant.

1.542 units and 1.18 percent. This indicates some tendency of overshifting, at least in absolute terms.

To which extent are these results consistent with the theoretical analysis in the previous section? First, note that it is confirmed that product-specific pass-on is smaller than the overall pass-on of all product's costs.[74] This underscores the importance of formulating a pass-on regression after controlling for the cost changes by all firms, if one is interested in an unbiased estimate for product-specific pass-on.

Second, note that product-specific pass-on is the smallest for the products with the smallest market shares in this example. The average market shares of the private label, Minute Maid and Tropicana during the sample period were respectively 28.0, 24.8 and 47.2 percent (ignoring the market shares of other brands, which constitute about 10 percent of the sales, see Kadiyali et al. 2000). Tropicana is the firm with the highest market share and the highest pass-on, especially when expressed in elasticity terms. The theoretical analysis in section 3.2.3.3 was ambiguous about the role of market shares, with a tendency of the highest degree of pass-on for intermediate market shares (see Observation 3.9). The empirical analysis correspondingly provides important information that theoretical considerations cannot bring.

Third, the overall pass-on estimates in absolute terms indicate a tendency of overshifting. This appears to be in contradiction with the theory, in particular the predictions on pass-on with differentiated products, unless if one would assume strong increasing returns to scale (declining marginal cost curve). For example, Observation 3.7 established that industry-wide pass-on is incomplete in a logit model of product differentiation. Yet it was also noted that this was dependent on the functional form assumptions of the logit model, which generate price elasticities that are increasing in price, or equivalently, markups that are declining in price; under the logit assumptions firms thus absorb cost changes by adjusting their markups. Under other models of product differentiation, such as the CES model, markup absorption may be lower (or absent) so that industry-wide pass-on in absolute terms can be greater than 1. This application thus shows that since we do not know in general how the elasticity or markup varies with price, it is better to take an empirical approach and measure the extent of pass-on.

4.3.2 Extensions

The above analysis of pass-on takes the most stylised form. In merger investigation, it is desirable to do various extensions so as to confirm the robustness of the results. We have already pointed out that it would be useful to have information on another retail chain, in particular if the proposed merger was between two retail chains (in which case it should not be too difficult to collect the data, as demonstrated by Ashenfelter et al.). Nevertheless, our analysis may be a reasonable approximation if competition between retail chains is relatively weak, for example due to consumer search costs. Our application may also be a reasonable point of departure if one is analysing a

[74] Note however that this only holds in this application for considering pass-on with respect to the market price index. As can be confirmed from the first nine rows of the Table, it does not hold for pass-on with respect to the individual firms' prices, in contrast to Ashenfelter et al.'s result. This is because Ashenfelter consider *competing* products, whereas we consider products that are priced by the same firm.

merger between two manufacturers (say Minute Maid and Tropicana) rather than between two retail chains.[75]

Regarding the econometric specification, several possible extensions may be undertaken. First, it is possible to estimate the model on the full data set, rather than averaged across stores and product categories. Furthermore, it is possible to consider other control variables for the reduced form pass-on equation, for example a measure for the macro-economic state of the economy (GDP, interest rate, etc.). It could also be possible to introduce dynamics in the analysis, for example by adding lagged price and cost variables, or by specifying a complete error correction model, which can distinguish between short-run and long-run effects of cost changes.

Finally, one may consider the role of alternative functional forms. We have already considered two possible functional form specifications: the linear and the logarithmic specifications. Note that both functional forms generated comparable measures for goodness of fit. An R^2 in the range of 0.6-0.75 was obtained for the pass-on regressions on the individual prices; an R^2 of around 0.5 was obtained for the pass-on regressions on the market price index, whether one consider the linear or the logarithmic functional form. A further extension is to consider the role of asymmetries in pass-on regarding cost increases and cost decreases. We have considered this possibility in Table 7, for the linear case (The results for the logarithmic case are analogous).

Table 7. Pass-on regressions with possible asymmetries

		level specification	
		estimate	stand. Error
Private Label	constant	-0,091	0,144
	cost PL	**1,442**	0,062
	cost PL up	**0,026**	0,015
	cost MM	0,006	0,042
	cost MM up		
	cost TR	-0,031	0,051
	cost TR up		
Minute Maid	constant	0,675	0,167
	cost PL	-0,143	0,066
	cost PL up		
	cost MM	**1,253**	0,058
	cost MM up	**0,026**	0,020
	cost TR	-0,082	0,059
	cost TR up		
Tropicana	constant	-0,196	0,153
	cost PL	-0,050	0,055
	cost PL up		
	cost MM	-0,089	0,040
	cost MM up		
	cost TR	**1,545**	0,061
	cost TR up	**0,032**	0,013
Whole Market	constant	-0,475	0,170

[75] Yet for this case a complete analysis would also require data on the manufacturers' cost.

cost PL	**0,518**	0,069
cost PL up	**-0,014**	0,016
cost MM	**0,505**	0,055
cost MM up	**-0,015**	0,019
cost TR	**0,606**	0,068
cost TR up	**-0,026**	0,015

The interpretation of Table 7 is similar as in the previous Table the new parameters are the parameters related to the cost "up" variables. These parameters indicate whether there is a different effect on prices (in absolute terms) when costs go up, than when costs go down. In merger analysis, the relevant estimates refer to what would happen to prices when costs go *down* (since the claims by the parties concern efficiencies rather than inefficiencies). We limit our discussion here to the last eight rows; the discussion for the other results straightforwardly generalizes the previous Table. One can see that a one unit cost decrease by the private label would decrease the price by 0.518. In contrast, a one unit cost increase has a smaller effect: it would increase the price by only 0.518 – 0.014 = 0.504 units. Note, however, that the difference is not significant. Similarly, for Minute Maid and Tropicana the effects of cost increases are somewhat less pronounced than for decreases, though the effects are not significant. In sum, this extension would indicate that the pass-on estimates remain more or less robust whether or not one allow pass-on to differ when the effect of cost increases is allowed to differ from the effect of cost decreases.

4.4 Concluding remarks

The above empirical analysis illustrates how an analysis of pass-on can be implemented in practice. The analysis complements the original work of Ashenfelter et al. They demonstrate that relatively simple regressions can be estimated, to measure firm-specific pass-on using data on price and variable firm-specific costs to be provided by the two merging firms. We extended the analysis to include other products sold by the same firm to account for the multiproduct nature of the firm.[76]

Our analysis was done in the context of retailing. Note that this has also implications for analysing mergers by (upstream) firms in manufacturing. If two upstream firms merge, one should not simply ask to which extent these firms will pass-on cost savings into the wholesale prices charged to the downstream retailers. One should also ask whether these downstream retailers will pass-on the changed wholesale prices onto the final consumer prices.

[76] Yet unlike Ashenfelter et al. we could not account for competing products due to data limitations. These problems are not likely to be present with an actual merger investigation, since then one can request data by the competing merging firms.

5. FIGURES

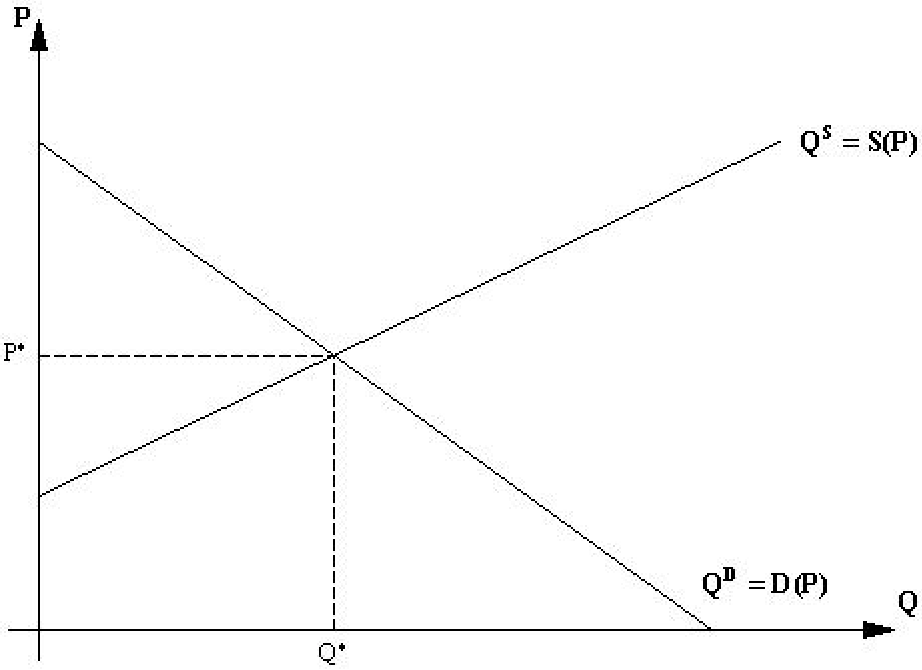

Figure 1

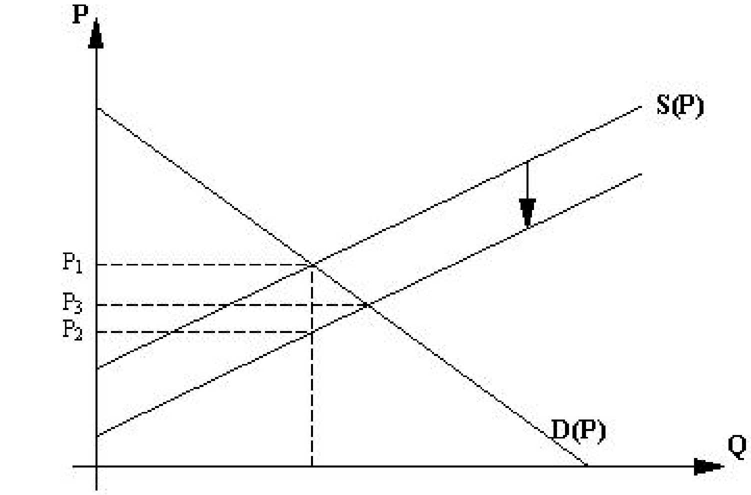

Figure 2a

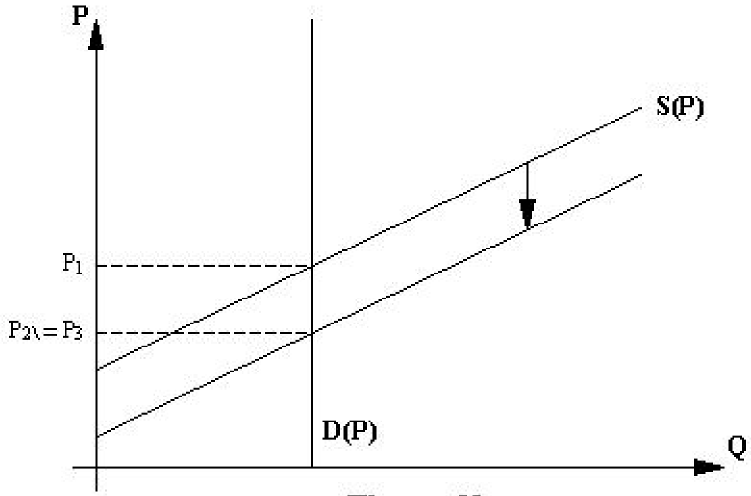

Figure 2b

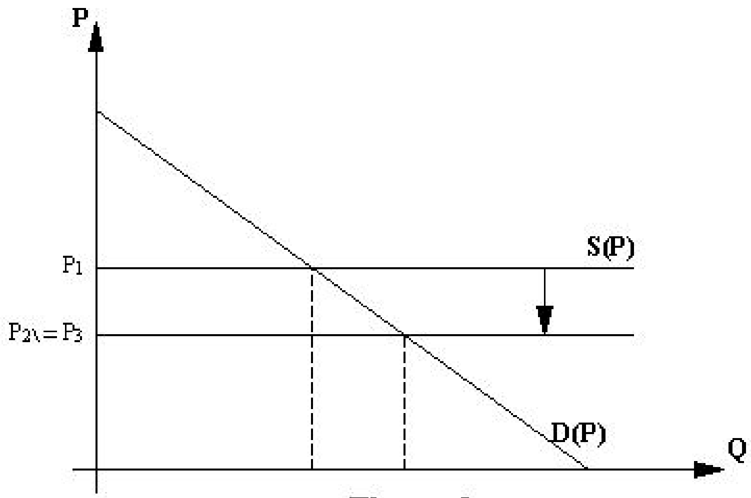

Figure 2c

6. REFERENCES

Ashenfelter, Orley, David Ashmore, Jonathan B. Baker and Signe-Mary Mc Kernan, 1998, Identifying the Firm-Specific Pass-Through Rate, FTC Discussion paper.

Bailey, Elizabeth E.; Friedlaender, Ann, F.: Market Structure And Multiproduct Industries, *Journal of Economic Literature*, Vol. 20 (September 1981), pp. 1024-1048.

Baumol, W.J.; Panzar, J.C.; Willig, R.D.: *Contestable Markets and the Theory of Industry Structure*. New York: Harcourt Brace Jovanovich, 1982.

Bain, Joe S.: *Barriers to New Competition*, Cambridge: Harvard University Press, 1956, especially pages 71-83 and 227-249.

Baker, Jonathan B., 1998, Econometric Analysis in FTC versus Staples, Prepared remarks before American Bar Association's Antitrust Section, 1997.

Baker, Jonathan B., Policy Watch: Developments in Antitrust Economics, Journal of Economic Perspectives; 13(1), Winter 1999, pages 181-94.

Baker, Jonathan and Timothy Bresnahan, Estimating the Residual Demand Curve Facing a Single Firm, International-Journal-of-Industrial-Organization; 6(3), 1988, pages 283-300.ges 25-31.

Barzel, Yoram, An Alterative Approach to the Analysis of Taxation, Journal of Political Economy, 1976, 84, 1177-1197.

Baumol, W. J.; Panzar, J. C.; Willig, R. D.: *Contestable Markets and the Theory of Industry Structure*, New York: Harcourt Brace Jovanovich, 1982.

Bettendorf, Leon, and Frank Verboven, 2000, Incomplete transmission of coffee bean prices: evidence from the Netherlands, European Journal of Agricultural Economics, 27, 1-16.

Bishop, Robert L., 1968, The Effects of Specfic and Ad Valorem Taxes, Quarterly Journal of Economics, 82, 198-218.

Borenstein, Severin, A. Colin Cameron and Richard Gilbert, 1997, Do Gasoline Prices Respond Asymmetrically to Crude Oil Price Changes?, Quarterly Journal of Economics, 305-339.

Braeutigam, R. ; Daughety, A. F.: On the Estimation of Returns to Scale Using Variable Cost Functions, *Economic Letters* 11, 1983, pa

Bresnahan, Timothy F.: Empirical Studies of Industries with Market Power, in: Schmalensee, Richard; Willig, Robert D., eds. *Handbook ofindustrial organization*. Volume 2. Handbooks in Economics, no. 10,Amsterdam; Oxford and Tokyo: North-Holland; 1989, pages 1011-57.

Bulow, Jeremy and Paul Pfleiderer, 1983, A Note on the Effect of Cost Chnages on Prices, Journal of Political Economy, 91, 182-185.

Chappell, Jr., H. W.; Wilder, R. P.: Multiproduct Monopoly, Regulation, and Firm Costs: Comment, *Southern Economic Journal* 52(4), 1986, 1168-1174.

Christensen, L.; Greene, W.: Economies of Scale in U.S. Electric Power Generation, Journal of Political Economy 4, 1976, pages 655-676.

Coelli, Tim.: Assessing the Performance of Australian Universities using Data Envelopement Analysis, mimeo. Centre for Efficiency and Productivity Analysis, University of New England, 1996.

Coelli, Tim.; Rao, D.S. Prasada; Battese, George E.: An Introduction to Efficiency and Productivity Analysis, Dordrecht: Kluwer Academic Publishers, 1998.

Davis, S.; Lyons, B. (eds.): *Industrial Organisation in the European Union: Structure, Strategy and the Competition Mechanism*, Oxford University Press, 1996.

Dixit, Avinash K., Comparative Statics for Oligopoly, International Economic Review; 27(1), February 1986, pages 107-22.

Farrell, Joseph; Shapiro, Carl : Horizontal Mergers: An Equilibrium Analysis, *American Economic Review*; 80(1), March 1990, pages 107-26.

Foreman, R. Dean, and others: Scale Economies in Cellular Telephony: Size Matters, *Journal of Regulatory Economics* v16, n3 (November 1999): 297-306.

Friedlaender, Ann F.; Spady Richard H.: Freight Transport Regulation: Equity, Efficiency, and Competition in the Rail and Trucking Industries. Cambridge, MA: MIT Press, 1981.

Friedlaender, Ann F.; Whinston, Clifford; Wang, David Kung**:** Cost, Technology and Productivity in the U.S. Automobile Industry, *Bell Journal of Economics* v14, n1 (Spring 1983): 1-20.

Fuss, M. A.: A Survey of Recent Results in the Analysis of Production Conditions in Telecommunications, in: L. Courville, A. de Fontenay and R. Dobell, eds., *Economic Analysis of Telecommunications: Theory and Applications.* Amsterdam: North-Holland, 1983.

Gold, Bela: Changing Perspectives on Size, Scale, and Returns: An Interpretive Survey, *Journal of Economic Literature*, Vol. 19 (March 1981), pp. 5-33.

Harris, Jeffrey E. The 1983 Increase in the Federal Cigarette Taxes, Reprinted from Tax Policy and the Economy, 1987, edited by L.H. Summers, Cambridge, MA: M.I.T. Press.

Hausman, Jerry A. and Gregory K. Leonard, 1999, Efficiencies from the Consmer viewpoint, George Mason Law Review, 7, 707-727.

Jara-Diaz, Sergio; Whinston, Clifford: Multiproduct Transportation Cost Functions: Scale and Scope in Railway Operations, Eighth European Association of Research in Industrial Economics, Vol. 1. Eds: Nicklaus Blattner, et al. U. of Basle, 1981.

Jara-Diaz, Sergio; Cortes, Cristian E.: On the Calculation of Scale Economies from Transport Cost Functions, *Journal of Transport Economics and Policy* v30, n2 (May 1996): 157-70

Johnson, Terry R., Additional Evidence on the Effect of Alternative Taxes on Cigarette Prices, Journal of Political Economy, 1978, 86, 325-328.

Kadiyali, Vrinda, 1997, Exchange rate pass-through for strategic pricing and advertising: an empirical analysis of the U.S. photographic film industry, Journal of International Economics, 43, 437-461.

Kadiyali, Vrinda, Pradeep Chintagunta and Naufel Vilcassim, 2000, Manufacturer-Retailer Channel Interactions and Implications for Channel Power: an Empirical Investigation of Pricing in a Local Market, Marketing Science, 19(2), 127-148.

Kinnucan H.W. and O.D. Forker, 1987, Asymmetry in Farm-Retail Price Transmission for Major Dairy Products, American Journal of Agricultural Economics, 69, 285-292.

Knetter, Michael, 1989, Price Discrimination by U.S. and German Exporters, American Economic Review, 79, 198-210.

Levin, Dan, 1990, Horizontal Mergers: the 50% Benchmark, *American Economic Review*, 80(5), 1238-1245.

Martin, Stephen: *Advanced Industrial Economics*, Oxford: Blackwell, 1993.

Mayo, J.W.: Multiproduct Monopoly, Regulation, and Firm Costs', *Southern Economic Journal* 51(1) (1984): 208-218.

Mayo, J.W.: The Technological Determinants of the U.S. Energy Industry Structure, *Review of Economics and Statistics* 66 (1986): 51-58.

McCorriston, S., C.W. Morgan, and A.J. Rayner, 1998, Processing Technology, Market Power and Price Transmission, Journal of Agricultural Economics, 49, 185-201.

Mulligan, James, G.: The Economies of Massed Reserves, *American Economic Review* 73, September 1983, pages 725-734.

Narnett, Paul G., Theodore E. Keeler, The-wei Hu, 1995, Oligopoly Structure and the Incidence of Cigarette Excise Taxes, Journal of Public Economics, 57, 457-470.

Nerlove, M.: Returns to Scale in Electricity Supply, in: C. Christ, ed., *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*. Stanford: Stanford University Press, 1963.

Panzar, J. C.: Determinants of Firm and Industry Structure, in: Schmalensee, Richard; Willig, Robert D., eds. *Handbook of industrial organization*. Volume 2. Handbooks in Economics, no. 10,Amsterdam; Oxford and Tokyo: North-Holland; 1989, pages 3-59.

Panzar, J. C.; Willig, R. D.: Economies of Scope, American Economic Review 71(2), 1981, pages 268-272.

Perry, Martin K.: Vertical Integration: Determinants and Effects, in: Richard Schmalensee and Robert D. Willig, eds., *Handbook of Industrial Organization*, Vol. 1, Amsterdam: North-Holland, 1989.

Pick, D.H., J. Karrenbrock, and H. F. Carman, 1990, Price asymmetry and Marketing Margin Behavior: an Example for California-Arizona Citrus, Agrobusiness 6, 75-84.

Pratten, C. F.: *Economies of Scale in Manufacuring Indstry*, Cambridge: Cambridge University Press, 1971.

Pratten, C.: A Survey of the Economies of Scale, in: *Research on the "Cost of Non-Europe", Basic Findings, Volume 2, Studies on the Economics of Integration*, Luxembourg: Office for Official Publications of the European Communities, 1988.

Röller, L-H.; Stennek, J.; Verboven, F.: *Efficiency Gains from Mergers*, Report to the European Commission, 1999.

Sharkey, W. W.: The Theory of Natural Monopoly. Cambridge U.K.: Cambridge University Press, 1982.

Scherer, F. M.; Beckenstein, Alan; Kaufer, Erich; Murphy, R. D.: The *Economics of Mult-Plant Operation: An International Comparison Study*, Cambridge: Harvard University Press, 1975.

Scherer, F. M.; Ross, David: *Industrial Market Structure and Economic Performance*, Third Edition, Houghton Nifflin Company, Boston, 1990.

Scheffman,-David-T.; Spiller,-Pablo-T., Geographic Market Definition under the U.S. Department of Justice Merger Guidelines, Journal-of-Law-and-Economics; 30(1), April 1987, pages 123-47.

Shapiro, Carl: Mergers with Differentiated Products, Address by before the American Bar Association International Bar Association, November 9, 1995.

Spady, Richard H.; Friedlaender, Ann F.: Hedonic Cost Functions for the Regulated Trucking Industry, *Bell Journal of Economics* 9(1) (Spring 1978), pp159-79.

Sumner, Daniel, 1981, Measurement of Monopoly Behavior: an Application to the Cigarette Industry, Journal of Political Economy, 89, 1010-1019.

Sumner Daniel and Michael K. Wohlegant, 1985, Effects of an increase in the Federal Excise Tax on Cigarettes, American Journal of Agricultural Economics 67, 235-242.

Sullivan, Daniel, Testing Hypotheses about Firm Behavior in the Cigarette Industry, Journal of Political Economy, 1985, 93, 586-598.

Sutton, John: *Sunk Costs and Market Structure*, Cambridge: MIT Press, 1996.

Sutton, John: *Technology and Market Structure*, Cambridge: MIT Press, 1998.

Sumner, Michael and Robert Ward, 1981, Tax Changes and Cigarette Prices, Journal of Political Economy, 89, 1261-1265.

Stigler, George J.: The Economies of Scale, Journal of Law & Economics vol. 1, October 1958, pages 54-71.

Teece, D. J.: Economies of Scope and the Scope of the Enterprise, Journal of Economic Behavior and Organization 1, 1980, pages 223-247.

Von Cramon-Taubadel, Stephan, 1997, Estimating asymmetric Price Transmission with the Error Correction Representation: An Application to the German Pork Market, European Review of Agricultural Economics, 1998, 25, 1-18.

Wang Chiang, J. Shaw-er: Economies of Scale and Scope in Multiproduct Industries; A Case Study of the Regulated U.S Trucking Industry, Ph.D. Dissertation, MIT, July 1981.

Weiss, Leonard W.: Optimal Plant Size and the Extent of Suboptimal Capacity, in: Robert T. Masson and P. D. Qualls, eds., Essays on Industrial Organization in Honor of Joe S. Bain, Cambridge: Ballinger, 1975, pages 128-131.

Werden, G. J. and L. M. Froeb (1994): “The Effects of Mergers in Differentiated Products Industries: Logit Demand and Merger Policy”, *Journal of Law, Economics, and Organization*; 10(2), 407-26.

Willig, Robert D.: Merger Analysis, Industrial Organization Theory, and Merger Guidelines, Brookings Papers: Microeconomics, 1991.

Yde, Paul and Micael G. Vita, 1996, Merger Efficiencies: Reconsidering the Passing-On Requirement, Antitrust Law Journal, 64, 735-747.