Chapter 5

Transport costs

Learning Outcomes:

In the course of this chapter, you will learn about:

- The theory of production
- Definitions of technical, cost and allocative efficiency
- How time is divided in economics
- How transport costs behave in the short and long runs
- Returns to scale and economies of scale in the transport industries
- The relevance of production costs to the supply of transport services.

INTRODUCTION

Whilst elasticity of demand impacts upon the demand side of the market, a major factor affecting the quantity supplied is the cost of production. Transport costs however fall into a variety of different classes. There are those costs that impact on the individual user of a particular mode of transport who directly benefits from undertaking a journey. These are known as private costs and would include both the financial costs involved, such as the fare in the case of public transport, as well as non financial costs, such as the time involved in undertaking the journey. Taken together these are known as the 'generalised cost'. There are then the costs of transport that fall on non users of the transport service who do not benefit from that transport service. This includes what could almost be termed the unwanted by-products from the undertaking of transport activities, such as polluted air, the congested road, noise and visual intrusions. These are generally referred to as public costs, and as these are a significant factor in the provision of transport services these will be examined in some depth in various chapters later in the text. Finally there are production costs that fall on the operators of a transport service or in the case of private transport the financial costs incurred when undertaking the activity. In many ways these are essentially private costs, as the individual that incurs the cost (the operator/road user) is the one that benefits from the provision of that service (i.e. profit/benefit from the journey). This chapter is specifically concerned with costs that fall into this last group, particularly in the production of public transport services. Road user issues are examined later in Chapter 8 under pricing.

Public transport costs have received a large amount of attention over the years, as these services are a vital component of the economy and society and have been to a greater or lesser extent subsidised by national and local governments. In simple terms, high cost levels restrict transport authorities' options and constrain the level of public transport services that can be provided. Even where such services are not subsidised and are provided by profit-making private companies, the constraints placed on supply arising from costs still apply: you should remember from Chapter 3 that the cost of production is one of the main determinants of supply. Following that logic, if costs can be reduced more transport services can be provided, i.e. supply can be increased. In subsidised markets, of course, lower costs reduce the need for subsidy.

Attempting to reduce and maintain downward pressure on public transport costs has therefore been a main concern of government policy. This has been one of the main reasons for on-going moves by authorities to shift transport operators more towards 'the market' rather than operating as state owned and controlled public enterprises. This move has resulted in a substantial shift from publicly owned operators towards privately owned companies and the introduction of market principles, most visibly, through competition in one form or another in the supply of transport services.

The actual size of costs and authorities' attempts to keep them as low as possible however are not the only issues surrounding costs. How costs are incurred is also important. For example, in many instances there may be a large initial overhead incurred by the firm but the actual cost of providing one extra service or carrying one extra passenger may be very small. For example, the extra cost to a bus or train operator of transporting an additional passenger, if they are operating at less than full capacity, is negligible. Alternatively, there are other cases where there is a very low initial overhead but most costs are incurred as a direct result of providing the transport service. Even at a more basic level, are large companies more cost effective than small companies or does an 'optimum' size exist, i.e. not too big and not too small? There may for example be certain advantages to being a large-scale operation that leads to certain cost savings; however, in other cases the sheer size of the firm may cause certain inefficiencies to arise. All of these issues will heavily influence the structure of the transport sector and determine how transport services are provided to the market.

THE EFFICIENT PRODUCTION OF TRANSPORT SERVICES

The costs of transport operations are primarily dependent upon a combination of the production processes used and the efficiency of the management of that production process. Not only that, however, but the physical characteristics of the operating environment in which transport services are provided will also impact upon costs. The start/stop nature of urban bus operations due to high densities of population, for example, tend to make such an environment more costly to provide public transport services per vehicle kilometre than in rural areas – DfT statistics (DfT, 2007) for example show the cost per bus kilometre to be some two and a half times higher in London than for the rest of England outside of the metropolitan areas. The reason for this is the considerably longer running distances between stops and overall faster running speeds in rural areas. There is little that transport management can do however about these aspects of operation. In terms of controllable costs, therefore, it is only the production process that is under the direct control of management, and consequently is the main determinant of costs to the operator.

It may be difficult to reconcile transport operators with the idea of the traditional firm, where inputs are fed in at one end of the factory and finished goods emerge from the other. The main difference of course is that transport is a service, and thus production is concerned with producing 'service units'. In order to measure the output of the transport firm, therefore, such 'service units' need to be quantitatively defined. This first raises the question of what exactly do transport operators produce, which is not as easy a question to address as it may first seem. What needs to be considered is what are transport operators actually attempting to achieve. The answer may appear to be obvious; a public transport operator, for example, simply moves people from A to B. Output would therefore be measured in terms of the journeys undertaken, either directly by the number of journeys or by the total number of passenger kilometres to also account for differences in distances travelled. Strange as it may appear, however, in many cases moving people from one location to another is not the aim of a public transport operator. Many instances exist where operators are contracted by transport authorities to provide transport services within a given area or location, irrespective of the number of people moved. Rather than journeys, therefore, the output would be measured in terms of the number and level of services produced, and hence could be expressed in terms of vehicle kilometres. How, therefore, should the output of the transport firm be measured, by journeys or by vehicle kilometres produced?

These two different outputs can be equated by the idea that what operators attempt to do is move people from A to B, but how this is achieved is through the production of vehicle kilometres. The idea of producing vehicle kilometres is far more consistent with the traditional view of production, but this whole topic needs to be related to the aims of the producer, which you may remember is one of the determinants of supply. These can vary from profit maximisation in a completely free transport market where revenue is only related to the carriage of passengers, to sales maximisation in a tightly regulated market where the transport company is directly paid to operate the service and hence profit maximises by attempting to fill the available capacity. Passenger journeys as the output would be more consistent with the former as revenue is only connected to the carriage of passengers, whilst vehicle kilometres specified as the output would be more consistent with the latter, as the firm is paid to provide services and hence profits are maximised through filling the available capacity, i.e. sales maximise. What is important in this context is that there is some measurable output at the end of the production process, whether that be the number of journeys or the number of vehicle kilometres produced, and this will largely be dependent upon the type of market the firm is operating in.

You should already be aware of the concept of production; however, in this context it specifically relates to the process whereby economic inputs or resources in the form of the factors of production (recall from Chapter 1 that this relates to inputs in the form of land, labour and capital) are brought together by entrepreneurs (or firms) in order to produce goods and services. This basic process is outlined in Figure 5.1.

As illustrated by the figure, the production process consists of converting the inputs of Land/Raw Materials, Labour and Capital, by way of a combination process, into the outputs of Transport Services and By-Products. Transport Services are then sold for an economic return, with the aim being that the transformation process should convert the inputs into something of a higher economic value. The revenue generated from the sale of the output therefore should be greater than the payment for the inputs. Also shown in Figure 5.1 is 'By-products'. Not all of the output of the combination process is sold for an economic return, as the process also results in the output

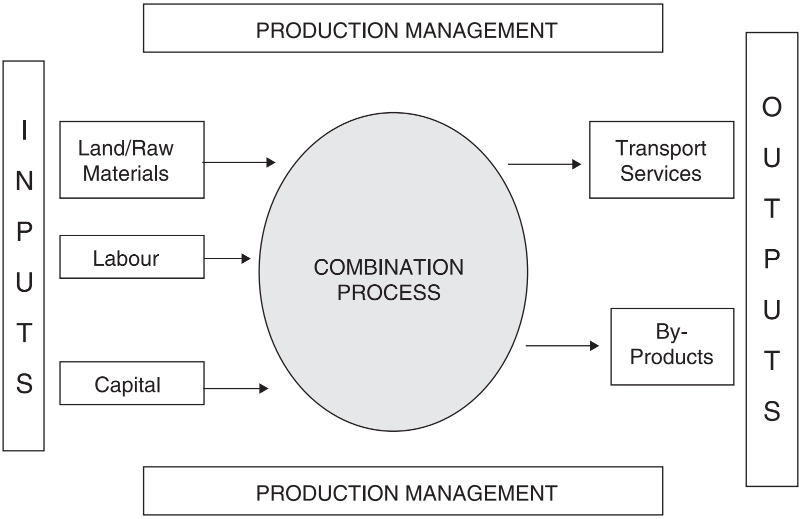

Figure 5.1 *The production process*

of other factors such as wastage or pollution. Often overlooked, this has to be included in the outputs of the production process and has important implications for transport markets.

To directly relate Figure 5.1 to transport industries, 'land' is an important factor in the production process, but of the three it is the one that is least under the control of transport operators. This is because land in many cases relates to the 'suitability' of the prevailing geographical environment to the provision of transport services. This was exemplified above by the urban/rural split, although many other examples exist. The provision of rail services, for example, will be relatively easier in areas of flat terrain (such as many parts of Belgium) rather than in mountainous regions (such as many parts of Switzerland). This is because the latter will require a far higher level of tunnels and bridging, i.e. less 'land' and more capital, in the production of services.

The other two factors of production are far more straightforward and more under the direct control of the firm. Put simply, labour relates to all staff involved in the production of transport services, whether that be operational or administrative staff, whilst capital relates to any goods that have been manufactured in order to be put into the production process. This obviously not only includes the vehicle stock, but also any other physically made equipment, e.g. terminal buildings, infrastructure, bridges, tunnels, handling equipment, depots and IT facilities.

The transformation process outlined above is known as production, and the relationship between the level of inputs and the level of the outputs achieved known as the production function. This is formally specified below:

$$Q = f(A, L, K)$$

Where:

Q = quantity of output produced
f = 'some function of'

A = quantity of land and raw materials used in the production process
L = quantity of labour used in the production process
K = quantity of capital used in the production process

This equation simply states that the level of output is some function of land/raw materials, labour and capital, with the letters L and K always used in economics as the shorthand notation of these two inputs. Where an increase in the inputs leads to an increase in the outputs, which should virtually always be the case, then production would be said to be monotonic. Whilst appearing to be common sense, this is actually an important theoretical consideration as real life data does not always fit with what would appear to be a very basic theoretical concept. The inputs/outputs ratio is also one of the main bases for assessing whether a given operation can be described as 'efficient' or not. The idea of efficiency in the production of public transport services is an important concept, and as noted above has received considerable attention in the academic literature. 'Efficiency' however is an often over-used term and has different meanings to different people. Within the field of transport, it has been used in the past to describe issues such as reliability, punctuality, pricing, costs, subsidy levels, number of passengers carried and so on. Within economics, however, there are only three basic 'types' of efficiency although these may be seen elsewhere under slightly different terms. However, they are:

Technical efficiency – this relates to the outputs to inputs ratio, with a technically efficient operator being one that uses the minimum level of inputs to produce the maximum level of outputs. Alternatively, this may be achieved where the minimum level of inputs is used to produce a given level of output. Both measures are highly relevant in the study of transport industries, as in many transport markets the output level is set by a transport authority and hence for the operator the technical efficiency question is one of input minimisation.

Cost efficiency – sometimes referred to as productive efficiency, or even cost allocative efficiency, cost efficiency arises because there may be several different ways to produce the output, all of which would be technically efficient. For example, a high level of capital and a low level of labour could be employed, or alternatively a high level of labour and a low level of capital employed. Both production processes may be technically efficient. The issue therefore becomes which one is the 'best'. In order to answer this question, the relative prices of labour and capital are examined and the one that produces the lowest cost combination deemed to be the 'best' or more exactly the most cost efficient.

Allocative efficiency – even within the study of economics, there is much confusion over the term 'allocative efficiency', which is quite surprising given it is almost the holy grail of the economics discipline! Allocative efficiency however relates to usage. As many of the former Eastern European countries showed in the past, there is little point in producing goods and services that are technically efficient in the lowest cost combinations if no one wants them. This is a total waste of resources and hence could never be considered as efficient. Allocative efficiency is therefore said to exist where goods and services are produced cost efficiently and in the 'right' quantities. Where this exact position is found is where the price paid for the good, which can be used as in indicator of the additional benefit received from that good, equals the extra cost of producing that good. This however is an issue which will be returned to in greater depth in the following chapter.

The first two of these efficiency concepts can be illustrated graphically, which is done in Figure 5.2.[1]

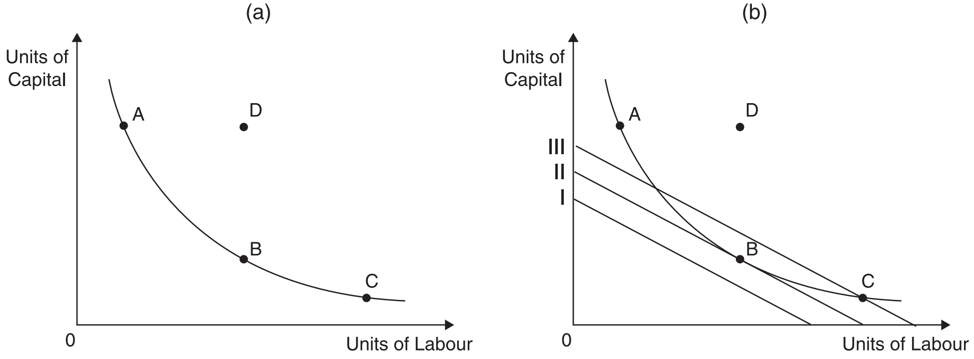

Figure 5.2 *Technical and cost efficiency*

On the left, Figure 5.2 (a) illustrates the combinations of labour and capital required to produce a single unit of output. Consequently, points nearer the origin are more efficient, as relatively speaking less of the inputs are used in the production process. Also shown are the relative positioning of four hypothetical firms, A to D. In this example there exists a large range of combinations of the two inputs that can be used; this varies from a high element of labour and a low amount of capital (firm A), to a high level of capital and low amount of labour (firm C). Whilst this may appear to have little relevance to transport services where the relative quantities of labour and capital are unlikely to vary greatly across firms in the same industry, some variation does exist and this is often the situation with regard to the carriage of people and freight across a number of different modes. These tend to employ varying levels of capital and labour in order to achieve the same purpose. In Figure 5.2 (a) firms A, B and C have the lowest possible combinations of labour and capital, thus an efficiency frontier can be drawn between these points. Consequently points above that 'frontier' would represent inefficient firms and points below simply unobtainable with the level of today's technology. You should hopefully recall that this is identical to the idea of the production possibility curve introduced in Chapter 1, only expressed in a different way: in this case the focus is on minimising inputs rather than maximising the output. The frontier is curved for theoretical reasons that surround variations in the way that one input is substituted by the other across different levels of the inputs, known as elasticities of substitution. As firms A, B and C outline the actual technical efficiency frontier, all three would be deemed to be technically efficient. This is because there is no way of assessing which is the 'best' combination of the two inputs – technical efficiency is simply concerned with the lowest combination of units used in the production process, not whether this represents the best 'mix' of the inputs to use. Firm D however lies above the frontier and thus would be said to be technically inefficient. This is because D is using a higher amount of capital than firm A and also a higher level of labour than firm B in the production process.

As regards the best combination of labour and capital to use in the production process, this can only be assessed once the prices of the individual factor inputs are introduced into the evaluation. Assuming that all firms within the industry face identical cost conditions, this is done in Figure 5.2(b). Added to the figure are budget lines, which are linear combinations of the costs of employing labour and capital and are drawn as straight lines out from the origin. These are straight because the combination of prices of the inputs are fixed, hence the cost of substituting one input for another is fixed over the various combinations of capital and labour that can be employed. The slope is determined by the relative prices of the two inputs of labour and capital.

The first 'viable' budget line on Figure 5.2(b) is line II, as this is the first that is tangential to the technical efficiency frontier. Firm B therefore has the lowest cost combination of inputs, as it lies both on the technical efficiency frontier and the lowest budget cost line. Note that in the case of the two other technically efficient points illustrated in Figure 5.2(a), firm C is on a higher budget line and hence would have a higher cost combination, and point A would lie on a higher budget line again.

These are the basic principles concerning the efficient production of transport services, but note these only concern production and not actual usage, i.e. allocative efficiency. This could be measured based on profit, as that would account for usage and whilst incredibly naive, profitable firms must be efficient as the output being produced is not only being used by consumers, but is also valued by them at a higher price than it cost to produce – hence they make a profit. It would be very difficult in practical terms however to use profit as a measure of efficiency, due to the presence of other (external) factors. Indeed these lead to the conclusion that it is virtually impossible to accurately assess the level of allocative efficiency in transport markets. As an example, the provision of a little used rural bus service may be essential to the daily functioning of a local community. However, if the level of allocative efficiency was to be assessed based on profit, then due to the relatively large number of resources (inputs) being used to move a relatively small number of people, this would always make a loss and hence be considered to be highly inefficient. Nevertheless, removal of the service on the grounds of allocative inefficiency may reduce overall welfare because a significant percentage of the benefits arising from the service do not accrue to the direct users of the service and hence would not be included in any such evaluation. In an entirely free market, however, that is exactly what would happen. This raises many issues, which we return to later in the text; however, the rest of this chapter concentrates on costs and production.

The economist's definition of time

We begin our examination of costs by firstly considering the issue of time. This needs to be examined in order to devise formal definitions for what is a relatively 'short' period of time and what is a relatively 'long' period of time. The reason for doing this is that costs may behave quite differently depending upon whether a relatively short period of time or a relatively long period of time is being considered. Time is defined in terms of the extent to which the factors of production can be varied in order to produce a different level of output.

The short run

In economics the short run is formally defined as that period of time during which at least one of the factors of production is fixed. This could in theory be any of the inputs; however, it normally relates to capital. The implication therefore is that variations in the level of output can only be achieved through variation in one or more of the other inputs, normally labour. This may be achieved through overtime working, employment of agency labour and so on.

The long run

In the long run, variations in the level of output can be achieved through variation of all of the outputs, thus the firm is not restricted to using only one of them. Capital can therefore be expanded to achieve such gains.

As regards an actual time period for the short and long runs, i.e. months, years and so on, this can vary considerably from industry to industry. In relatively labour-intensive industries, i.e. those that employ a relatively high degree of labour, such as the bus industry, the length of time in the short run can probably be measured in days or weeks. Hence a few days or a couple of weeks is roughly the length of time it would take a bus company to increase the capital stock of the firm, i.e. purchase a second-hand bus, assuming they are available. Procurements of new buses, however, will probably take considerably longer and would be measured in months. In more capital-intensive industries, i.e. those that employ a relatively high degree of capital in the production process, such as the rail industry, the short run can probably last up to a number of years. Thus for a railway company, the gap between ordering new rolling stock and it actually being introduced into service will take a minimum of roughly two years and can take anything up to five. There is therefore no fixed period of time as such with regard to the short and long runs, it will vary from industry to industry dependent upon how long it takes to vary all of the factors of production.

The very long run

The final definition of time in economics is the very long run. In simple terms the very long run is that period of time where all factors of production are variable, including the level of technology. Thus production levels that are not possible today may be possible in the very long run due to an increase in the level of technology.

COSTS AND PRODUCTION IN THE SHORT RUN

Having considered what constitutes ‘efficient’ production and the time dimension in production economics, the actual production process in the short run is now considered and the underlying economic concepts underpinning production theory examined. Short run production is considered first as this is the most basic form of the production function outlined earlier, as only one input is varied to produce variations in the level of output. The total output produced is known as the total product, with other important concepts in short run production being the average and marginal outputs (of the variable factor), known as the average and marginal products. The average product is simply the total product divided by the number of units of the variable factor, and hence in the case of labour would be more commonly known as labour productivity, i.e. the average amount produced by each person employed. The marginal product is the change in the total product that results from adding one more unit of the variable factor into the production process. Thus, for example, say two people are employed to drive a taxi which results in a daily mileage of 100 miles. The average product would be 50 miles. If however daily coverage was to be expanded, in the short run as one of the factors of production is fixed this could only be achieved by either increasing the number of drivers, or increasing the number of taxis, but not by increasing both. If therefore say one further driver was employed and the daily mileage rose to 120 miles, then average product would be 40 miles and the marginal product 20 miles. The theoretical relationship between total, average and marginal products is illustrated in Figure 5.3.

The main issue to consider here is the shape and relative positions of these three curves. Firstly, the total product curve – this is in the form of an 'S' shape. Hence, as more units of labour are added to the fixed amount of capital, production not only increases but increases at an increasing rate. This is up to point a in Figure 5.3, which is the point of the highest marginal product as shown in the lower diagram. Beyond point a, as more of the variable factor is added production still rises, but at a decreasing rate, reaching the highest point of average product curve at point b before eventually total product declines after point c. Thus in the short run some form of maximum output is reached. This tailing off effect after point a is known as the law of diminishing marginal returns. Peppers and Bails (1987) clarify that the marginal product of the variable factor of production will eventually decline if enough of it is combined with the fixed factor. In other words, a variable input cannot endlessly be added to a fixed factor to continually achieve ever increasing levels of output. A second point to note is that the marginal product curve cuts the average product curve at the latter's highest point. This relationship between the marginal and the average is an important concept in economics and one that will appear in many other contexts. It is however a simple mathematical relationship. In this example, while extra units of labour are increasing total output by more than the average, then this will increase the average. If however the

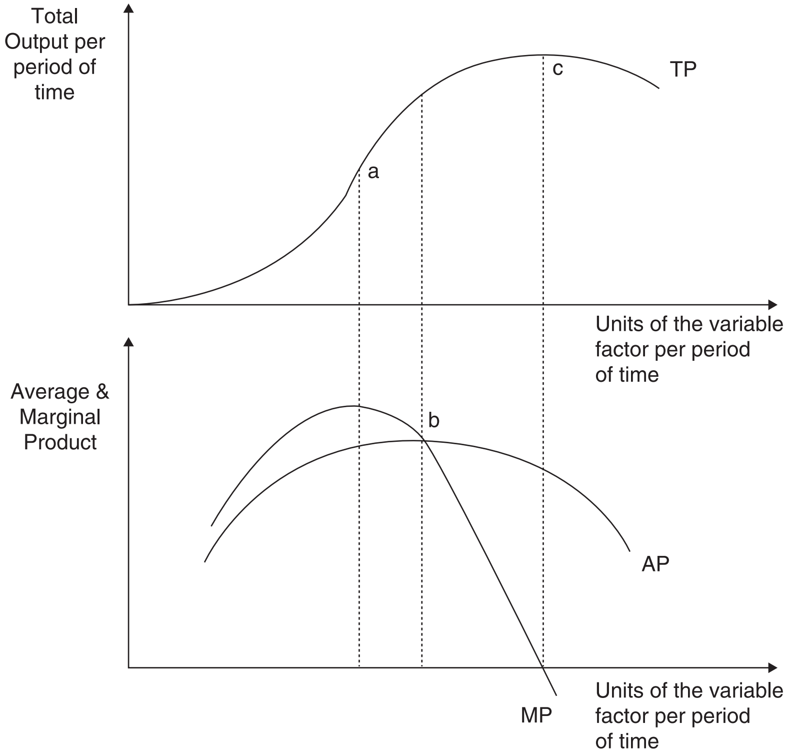
Total Output per period of time
c
TP
a
Units of the variable factor per period of time
Average & Marginal Product
b
AP
MP
Units of the variable factor per period of time

Figure 5.3 *Short run production, total, average and marginal product*

last person employed should add less to the total output than the average, this will pull the average down.[2]

Points b and c in Figure 5.3 can be used to break down the production process into three stages, known conveniently as Stage 1, Stage 2 and Stage 3 production. The divisions are based upon the total and marginal products. In Stage 1 production, the marginal product is always increasing, hence total product is increasing at a rising rate. This would be up to point b in Figure 5.3. In Stage 2 production, diminishing marginal returns set in. Thus although the marginal product is positive it is falling in value, and thus total product is increasing at a declining rate; in other words from point b to point c. The earlier taxi example was a case of Stage 2 production, as the last driver added reduced average product from 50 to 40 miles per employee. Finally in Stage 3 production the marginal product becomes negative and total product is decreasing, as shown by all points beyond point c in Figure 5.3.

All of these ideas are best underpinned by a practically orientated example. Table 5.1 outlines a hypothetical case of a bus operator and the short run production of bus services. It shows what happens to output levels, in this example measured in vehicle kilometres, as more units of labour (between 0 and 9) are added to a fixed capital stock, i.e. a fixed fleet size.

The first column in Table 5.1 simply gives the number of labour units employed in the production process. The second column is the resultant output from combining the variable units of labour with the fixed capital stock. The third column introduces the idea of the average product. This is simply the second column divided by the first, and as highlighted above, would usually be referred to as labour productivity, i.e. the average units produced per employee. The last column calculates the marginal product. In simple terms this is the extent to which output rises as a result of employing one extra unit of labour in the production process. These three measures therefore give the total, average and marginal products which follow the same general shape of the theoretical curves shown in Figure 5.3.

As all output variations are produced by the addition of the variable input, all changes in the level of output are ascribed to that variable input. Starting at zero, no output is produced, as the only factors of production that are employed are the fixed factors, in this case capital. However,

Table 5.1 *Variable labour and the production of bus services*

(1) *Labour units*	*(2)* *Total product* *(thousands)*	*(3)* *Average product* *(thousands)*	*(4)* *Marginal product* *(thousands)*
0	–	–	1
1	1	1.0	6
2	7	3.5	11
3	18	6.0	8
4	26	6.5	6
5	32	6.4	5
6	37	6.2	3
7	40	5.7	2
8	42	5.3	–1
9	41	4.6	

all factors are required to produce any output – buses cannot operate without drivers. Even after a single unit of labour is employed, however, still no output is produced. This is because by the time the single employee has opened up the bus station/depot, cleaned the facilities, carried out all of the administrative and maintenance tasks, there is no free time left to actually drive a bus!

When a second labour unit is employed, however, then a degree of specialisation can occur – one can look after the bus station/depot whilst the other drives a bus. As more units are employed, then more specialisation and/or the better scheduling of inputs can occur and hence the level of output increases at an increasing rate. This is shown by an increase in both the average product and the marginal product figures in Table 5.1. Output levels are therefore increasing at an increasing rate (as evidenced by the ever increasing figure of the marginal product), i.e. we are in Stage 1 production. This continues up to where five labour units are employed, at which point the marginal product reaches its highest value. With the employment of a sixth person, total output is still rising but now at a declining rate. The marginal product therefore begins to fall and diminishing marginal returns set in, which would be at point a on Figure 5.3. Although each person employed is still increasing the total output, they are not increasing the output by as much as their predecessor and Stage 2 production is entered, i.e. beyond point b on Figure 5.3. Note that this has nothing to do with individual 'productivity', i.e. the sixth person not being as hard working and/or as skilled as the fifth person employed, but rather is related to the combination of inputs employed in the production process. By the time nine labour units are employed, total output actually falls, thus Stage 3 production is entered as shown at point c on Figure 5.3. This can be for a number of reasons many of which are related to the productivity of the individual. For example, employees may now be getting in each other's way (remember, all other inputs are fixed). This may not be physically but rather figuratively, such as in the case where one employee has to wait until a colleague returns the bus to the depot before they can start using it and hence being productive.

As highlighted above, there is only so much output that can be produced by the fixed factor. In this example, buses could only be run for around 20 hours a day maximum (allowing for refuelling and maintenance). Whilst that may be the maximum level, there will also be an optimum level, i.e. a level of the variable factor (labour) that the fixed factor was 'designed' to be operated by. As this level of output is approached, productivity will increase. Once that point is exceeded however and the maximum level is approached, overall productivity will be reduced as the fixed factor is being 'overworked' whilst the variable factor in many respects is being underutilised.

Case study 5.1 Costs and production in transport operations

Although costs are not introduced until the next section, it is worthwhile at this stage to look at the actual inputs used in terms of the factors of production that are employed in the production of transport services through an examination of their respective costs. Irrespective of the mode, at the most basic level all transport operators will employ the same inputs in the production of transport services. At this crude level, these will consist of a combination of a vehicle, a driver/operative and a power source. The output would normally be measured in terms of the vehicle kilometres produced through the combination of these inputs, and the inputs would be viewed as being complementary, i.e. all three are required to produce transport services, rather than substitutable, i.e. using one rather than the other. There will nevertheless be a degree of substitutability between these basic inputs at the margins. The Docklands Light Railway in London for example uses driverless vehicles, which technically is a substitution of labour for capital.

Unfortunately with regard to the production inputs involved in transport modes, beyond this basic level is where the similarity ends. Starting at the most basic split, passenger versus freight operations, one of the major advantages of passenger services are that passengers by and large load themselves on to the vehicle, whereas freight has to be physically loaded. Freight operations, in terms of producing vehicle mileage therefore, will be less 'productive' as the level of inputs required will be higher. Little research exists in this area, however, as few operators serve both freight and passenger markets. This initially would suggest that there are little if any advantages in running both passenger and freight operations; however, railways and airlines are the two notable exceptions. In the case of railways, companies have tended to run both types of service due to economies of scope, and airlines carry large volumes of freight in the cargo hold of passenger aircraft due to goods in joint supply (both these concepts are examined further in Chapter 12). The only major direct comparative study on the topic was undertaken by Professor Chris Nash in 1985, who found that a higher labour component of around 1.45 was required for freight as opposed to passenger rail operations (Nash, 1985). Although limited, this is consistent with the view that freight operations will tend to be less productive in terms of vehicle kilometres. For airlines, however, the loading of freight onto passenger aircraft will be part of the general servicing of the aircraft whilst at the airport, and as all of this cargo is containerised the extra inputs required are likely to be relatively low.

In order to bring out the differences in factor inputs further, however, the costs of five different transport operators in five different markets are examined in Table 5.2. These have

Table 5.2 *Mode cost comparisons*

Operating Cost	*Airline*		*Ferry Operator*		*Bus Company*		*Railway Company*		*Parcels*	
	British Airways 2005/6		*Caledonian Mac. 2005/6*		*First Glasgow 2005/6*		*Virgin West Coast Value 2005/6*		*Parcelforce Value 2005/6*	
	Value	*%*	*Value*	*%*	*Value*	*%*	*Value*	*%*	*Value*	*%*
Labour Costs:	2346	30.0%	45.0	51.7%	45.4	69.2%	102.2	17.7%	5968	71.8%
Vehicle Costs:	1302	16.6%	18.0	20.7%	5.8	8.8%	171.7	29.8%	1392	16.7%
Infrastructure Costs:	0	0.0%	0.0	0.0%	0.0	0.0%	141.4	24.5%	0	0.0%
Fuel Costs:	1632	20.8%	9.5	10.9%	10.0	15.2%	15.6	2.7%	0	0.0%
Terminal Costs:	1514	19.3%	12.6	14.5%	0.0	0.0%	17.5	3.0%	530	6.4%
Other Overheads:	1034	13.2%	1.9	2.2%	4.4	6.7%	128.5	22.3%	426	5.1%
Totals:	7828	100.0%	87.0	100.0%	65.5	100.0%	576.9	100.0%	8316	100.0%
Fixed Inputs:	2816	36.0%	29.0	37.7%	5.8	8.8%	330.5	57.3%	1922	23.1%
Variable Inputs:	5012	64.0%	48.0	62.3%	59.7	91.2%	246.4	42.7%	6394	76.9%

Sources – Compiled from Company Annual Reports, 2005/6 been adapted from the original accounts by grouping costs into six common headings. It is important to stress even at this stage that the actual costs themselves are not what is important. This book for example is not entitled 'Transport Accounting' and hence this should not be viewed as an exercise in cost accounting. Rather, what is important is the balance between the different types of costs and the likely impact such a balance will have on the structure of production, the firm and the market. This exercise after all is about attempting to understand the economics of transport operations, hence the choice of title of the book!

The above examples of transport operators clearly illustrate through cost differences variations in the structure of production in the modes examined. What is important are not the actual amounts, as these reflect the different scale of operations of each company, but rather the relative levels spent on each input. This reveals a very high level of variation of input factors between modes. As staff costs are a direct measure of the relative proportion of labour used in the production of transport services, the above figures indicate that parcel and bus operations are more labour-intensive industries than the railways, ferries and airlines. In turn, railways, ferries and airlines would appear to be more capital intensive than the others shown. What is also interesting to note is the relative level of fuel costs – although only shown for four out of the five modes (due to accounting differences), these account for somewhere between around 3 and 20 per cent of all operating cost. Interestingly, fuel costs in the bus and airline companies are around the two highest relative shares of operating costs, suggesting that the level of fuel costs is independent of the labour and capital intensity of the production process within the industry. Note however that all transport modes employ a high level of capital equipment in the production of transport services, as for example few rickshaws exist as a mass mover of people in the Western world! What is being compared here therefore is the relative levels of labour used to work the capital equipment, and hence some transport modes are more highly capital intensive than others.

Listed at the bottom of Table 5.2 is a very rough division where costs have been arbitrarily split between those that relate to factor inputs that may be considered to be fixed and those considered to vary directly with production. Of the five modes shown, passenger railways appear to have the highest level of fixed inputs, at around 57 per cent, ferries and airlines have around 36 to 37 per cent of fixed costs, whilst parcels (23 per cent) and the bus company (9 per cent) appear to have by far the lowest level of fixed costs. The real significance of this division between fixed and variable is that variable inputs are only employed when transport services are actually provided, whilst fixed inputs will incur a cost even where no output is produced and hence accumulate with the simple passage of time. If a high proportion of operating costs relate to variable inputs, this means that in a service industry such as transport most costs are only incurred when in revenue earning service. In many ways this considerably simplifies the planning of operations and also in theory makes entry into a new industry/market easier, as all else being equal, start up finance requirements should be relatively smaller as revenue streams are more evenly matched with cost outgoings.

This division between fixed and variable factors and the associated costs has major implications on the structure of the market, as a high level of fixed costs coupled with a capital-intensive production process would suggest large firms, which would act against market entry and competition in the market. On the other hand, it may be expected that in more labour-intensive transport industries, such as the bus and parcel markets, competition in the market should be both achievable and sustainable. To some extent this would certainly appear to be the case with parcels, where a number of national and international couriers compete fairly intensively in the market place. In the bus industry, however, outside of a few isolated cases, there is only limited competition in the British deregulated market and the country tends to be divided into distinct bus company 'territories', hence little direct competition exists. This suggests that the structure of costs, whilst an important determinant, is not the only factor in determining market structure and that other characteristics need to be considered when examining such markets. This will be taken up in the next two chapters which examine the issues of market structure and competition in transport markets.

Costs in short run production

Before considering how costs can be classified, it is important to stress that costs in economics include profit, or to be more exact, what is known as normal or economic profit. A simplified definition of normal profit would be the opportunity cost of being in business plus some form of risk premium in recognition of the risks that the investor is taking. Hence for example say the next best alternative was for the investor to deposit the funds in the bank rather than invest in the business, and for this they received a 5 per cent annual return. Then in order to be worthwhile investing in the business, the investor would have to earn more than the 5 per cent per year earned at the bank, as that gain is virtually guaranteed. Normal profit is therefore the cost to the firm of the investor's outlay, and this is normally paid in the form of a dividend. As such, this has to be included in costs in much the same way that staff salaries are included in costs. Anything earned above the level of normal profit would be termed abnormal or supernormal profits, as these are rewards in excess of the risks of being in business.

Classifying costs

White (2008) highlights a number of possible ways of classifying costs in transport operations. Firstly, costs can be classified by input, hence all labour costs are grouped together, all vehicle costs and so on; this essentially is what was done in Case study 5.1. Secondly, costs can be classified by the associated production of various outputs, such as passenger operations and freight operations, or scheduled and chartered flights in the case of an airline. Finally, costs can be classified by the activity performed, thus a railway could group its costs under the headings of sales and marketing costs, train services costs, infrastructure costs, station costs, administration costs and so on, with each classification of costs relating to a specific activity of railway operation. All of these enable the analysis of different costs to be made. Within transport economics, however, costs are generally classified into fixed and variable. As the names suggest, a fixed cost is one that does not vary with the level of output, whilst a variable cost does. As seen in Case study 5.1 above, this division and the balance between fixed and variable costs can have considerable implications on the structure of transport markets.

Fixed costs are input costs that are sometimes classified as 'indivisibles' or unavoidable costs, as these costs must be paid even if no output is produced or service provided and cannot be divided or bought in parts, e.g. you cannot purchase half an aircraft or bus. This includes costs such as leasing charges; rents and rates that relate to offices, depots and stations; management salaries and costs and administrative expenses such as telephone line rentals. Variable costs on the other hand relate to the direct expense of providing the output, such as the wages of employees, fuel costs, power and electricity for heat and light. Many costs however fall somewhere between the two, as they partially vary in relation to output, hence technically these are semi-variable. Labour costs, specifically wages, although normally classified as a variable cost, would actually be a good example of such a semi-variable cost, as within a wage there is a basic weekly sum that can be supplemented with overtime working. Only the overtime element would be directly variable, hence wages technically are a semi-variable cost.

Depreciation

Depreciation is an important concept that it is worth spending some time on in order to clarify the issues. One of the main problems associated with this cost is that 'depreciation' is the reduction in economic value to the firm of using an asset in the production process. There will come a point in time where it will be more cost effective in the longer term to replace that asset rather than continue to patch it up through maintenance and keep it running. The more it is used, the quicker that point will be reached, hence technically depreciation as a concept is a variable cost. To return to transport accounting, however, this reduction in value is assessed in company accounts as only an approximation of the real reduction in value. There are two main methods used. Firstly, the straight line method. Under this method, the scrap value is subtracted from the purchase price and then divided by the number of expected years of usage. That then gives the value of 'depreciation' to be written off annually. The second approach is the reducing balance method, where a percentage of the value is written off each year until the scrap value is reached. Both approaches are best illustrated through the use of an example, and again a bus company is used.

Say for example a new bus costs £110,000, has a scrap value of £10,000 and has an expected useful life of ten years. The value to be written off therefore is £100,000 and over the useful life of ten years this would be done by subtracting a straight value of £10,000 a year. In ten years therefore it would be shown in the books as having a value of £10,000, i.e. its scrap value. This would be the straight line method. Alternatively, a percentage of the book value could be written off each year, in this case 21.3 per cent, and this will achieve the same effect of reducing the value to (almost) £10,000 in ten years' time. This would be the reducing balance method. The annual depreciation charge is then written off against profits, as this is the cost to the firm of using the capital equipment. Both approaches are fully illustrated in Table 5.3.

Note that under the reducing balance method higher values are written off in the earlier years, and this is said to better reflect the decreasing value of such assets in that earlier period. Irrespective of what method is used, however, neither takes into account usage. For example, as illustrated under the straight line method £10,000 is written off the value of a bus each year irrespective of the extent to which that bus is used. In this sense, therefore, it is a fixed cost, as it is one that does not vary with the level of output produced. In the longer term, however, if that bus was little used it would retain its economic value (obsolescence accepted) beyond the ten years, and hence would still be useable and thus of value to the firm after it had been written off in the books. There are many such examples of transport assets that have long been written off in the company accounts but are still in use today, such as the Severn Rail Tunnel and the Forth Bridge.

Table 5.3 *Illustration of depreciation by straight line and reducing balance methods*

Year	*Straight line method*			*Reducing balance method*		
	Value at the beginning of year	*Annual depreciation charge*	*Value at the end of year*	*Value at the beginning of year*	*Annual depreciation charge*	*Value at the end of year*
1	110,000	10,000	100,000	110,000	23,430	86,570
2	100,000	10,000	90,000	86,570	18,439	68,131
3	90,000	10,000	80,000	68,131	14,512	53,619
4	80,000	10,000	70,000	53,619	11,421	42,198
5	70,000	10,000	60,000	42,198	8,988	33,210
6	60,000	10,000	50,000	33,210	7,074	26,136
7	50,000	10,000	40,000	26,136	5,567	20,569
8	40,000	10,000	30,000	20,569	4,381	16,188
9	30,000	10,000	20,000	16,188	3,448	12,740
10	20,000	10,000	10,000	12,740	2,714	10,026

Short run average and marginal costs

As total, average and marginal products are plotted against labour, costs are normally plotted against the output produced by that labour. This then illustrates how costs vary over different levels of the output. These are known as the average and marginal costs curves, in this case those relating to the short run. The ‘S’-shaped production function outlined earlier will in turn produce an ‘S’-shaped total cost curve, therefore consistent with this the average cost curve is ‘U’ shaped. In simple terms, as total productivity increases (Stage 1 production), then average costs fall, whilst when total productivity decrease (Stages 2 and 3 production), average costs increase. This is shown in Figure 5.4.

This is the stylised textbook version of the short run marginal and average cost curves. As more of the variable input is added, average costs at first fall, are then minimised at the optimal output level of production at point b, and then increase beyond that optimum point. Having outlined the basic shape of these curves, what is important is why the short run average cost curve should be U shaped. As this relates to the short run, one of the inputs in the production process is fixed and hence will be subject to the law of diminishing marginal returns, which in turn will be related to utilisation of the fixed input. Say for example in the production of bus services, the number of buses is fixed. As highlighted previously, this means that there is an optimal production level, i.e. an ideal output size. Any production level that is under that point will lead to the under-utilisation of the fixed resource and consequently higher average costs. As that optimum point is approached, i.e. point b on Figure 5.4, these fixed resources are increasingly being utilised, hence average costs fall. Any point beyond the optimum however will lead to the over-utilisation of resources. Furthermore, as the level of output can only be altered by varying the level of the variable factor, in this case labour, then in order to increase labour this may involve paying over-time rates, hiring agency labour at a higher cost and so on, thus incurring a higher overall average unit cost.

In order to examine average and marginal costs further, Table 5.4 expands the previous

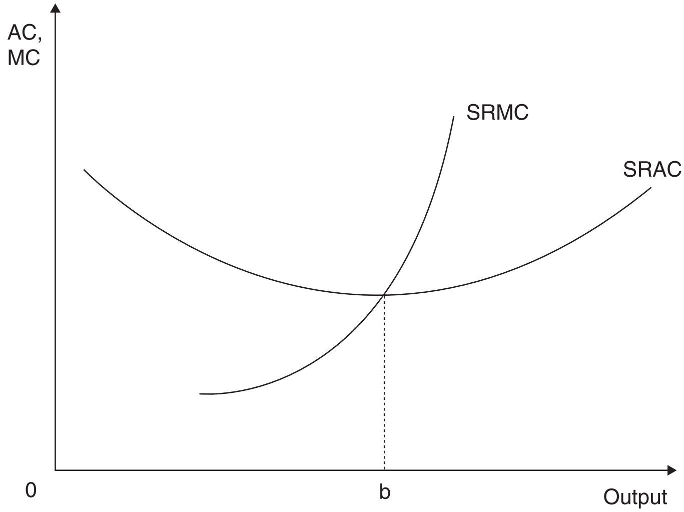

Figure 5.4 *Short run average and marginal cost curves*

Table 5.4 *Variable and fixed costs of short-run production of bus services*

Production				*Costs*				
Labour units	*Output produced (000s)*	*Average product (000s)*	*Marginal product (000s)*	*Total fixed costs*	*Total variable costs*	*Total costs*	*Average total costs*	*Marginal costs*
0	–	–	1	80000	–	80000	–	30.00
1	1	1.0	6	80000	30000	110000	110.00	5.00
2	7	3.5	11	80000	60000	140000	20.00	2.73
3	18	6.0	8	80000	90000	170000	9.44	3.75
4	26	6.5	6	80000	120000	200000	7.69	5.00
5	32	6.4	5	80000	150000	230000	7.19	6.00
6	37	6.2	3	80000	180000	260000	7.03	10.00
7	40	5.7	2	80000	210000	290000	7.25	15.00
8	42	5.3	–1	80000	240000	320000	7.62	–
9	41	4.6	–	80000	270000	350000	8.54	

example regarding the production of bus services in the short run by including cost data. In this case it has been assumed that fixed costs are £80,000 and the variable factor, labour, costs £30,000 per unit.

The first four columns are taken from Table 5.1, and to this have been added five new columns to include the relevant cost data. Total Fixed Costs, as outlined above, do not vary with the level of output, hence remain fixed at £80,000 irrespective of the level of output produced. These are the costs associated with the fixed factor, in this case capital. Total Variable Costs are simply the number of labour units employed times the cost of each unit, and total costs the addition of fixed and variable. Average total cost is therefore the total cost divided by total output, and marginal cost the cost of the last unit produced. When for example labour is increased from one to two units, output increases by 6,000 units and costs rise by £30,000. The marginal cost of the last unit therefore is the difference in output divided by the difference in costs and hence is £5.

Notice also that when only production is examined the highest level of labour productivity occurs at 4 units; however, the lowest average cost occurs when 6 labour units are employed. The reason for this difference is because the average cost also includes the capital cost, whilst no account is taken of capital inputs under the previous calculation as all variations in output were apportioned to the labour factor. This is one reason why productivity measures based on a single input, such as labour in this case, can be misleading.

The importance of the idea of the average cost is illustrated in the next case that looks at the operational characteristics of low-cost airlines.

Case study 5.2 The importance of average cost in the business model of low-cost airlines

It is difficult, if not impossible, to have a chapter on transport costs without having a specific look at the model of the low-cost airline. As the name of these carriers strongly suggests, this is a business model of airline operation that is heavily based on achieving low average costs in the operation of services. It should be stressed however that whilst 'low cost' is a philosophy that is central to these companies' operations, this is only one part of a complete business model that is geared towards achieving 'low cost' through not only cutting costs directly but also by a number of other measures as well.

The first acknowledged low-cost carrier was the American airline, Southwest, which began operating along low-cost principles in 1978, not long after deregulation of the US Domestic airline market. This business model however did not spread to Europe owing largely to the heavily regulated air market that had been exempt from both the initial Treaty of Rome in 1957 and the Single European Market in 1986. As a result, the European Air Market operated restricted capacity on routes and regulated prices and was largely dominated by state-owned National Airlines. In the early to mid 1990s, however, the EU passed a number of packages of air regulation reforms that opened up the market to potential new entrants. By doing so, this also presented an opportunity for the entry of low-cost operations, and this was initially taken up with considerable success by the Irish operator Ryanair. This model of operation has since been followed by a large number of other airlines, including easyJet, Flyglobespan.com and Flybe. These airlines operate on the basis of offering low priced fares based on a low average cost. Note that this is slightly different from an out and out low cost, as in this business model it is the achievement of low average cost that matters, not just 'low cost'. This obviously has implications on the operations of the airline, and in particular also puts great emphasis on aircraft utilisation. The general 'model' of the low-cost carrier can be summarised as consisting of the following main elements:

- Low staff costs
- Low aircraft turnaround times

- Route network based around secondary or regional airports rather than major hubs since they are usually less congested
- On line ticket sales only with ticketless travel
- Cabin crew perform other duties during turnarounds, such as cleaning the inside of the aircraft
- Simplified point-to-point operations rather than complex hub and spoke
- All 'extras' beyond the basic seat incur an additional charge, e.g. in flight catering or priority in boarding
- No spare aircraft capacity held in reserve to cover in the case of unforeseen breakdowns or delays
- Fleet based on a single aircraft type to reduce maintenance costs.

In order to tease out these differences, the costs of three airlines are examined below in Table 5.5. These have been grouped under five common headings in order to attempt to bring out these differences. The first airline, British Airways, can be viewed as a 'traditional airline', whilst the other two, easyJet and Ryanair, operate largely along low-cost principles, the latter more so than the former.

All costs are taken from the respective 2006 Annual Reports; however, not all costs are directly comparable. Despite the existence of Standard Statements of Accounting Practice designed to ensure consistency across company accounting practices, differences will always remain. Comparison differences therefore cannot be taken as 'exact'. As an example, easyJet only include Crew Costs (cabin and flight deck) as specific staff costs in their annual reports, whilst British Airways (BA) include all staff costs, e.g. cabin and flight deck crew, administration staff, sales and marketing staff and so on. Nevertheless, the above costs do give some basis for an overall appraisal of different costs.

BA provides an interesting starting point to examine firstly the overall cost structure of airlines and secondly where any savings in operating costs could be made in what would be perceived as a traditional airline. Starting with labour costs, these account for around 30 per cent of all operating costs. Whilst this is a significant proportion, in transport operating terms this suggests a relatively capital intensive industry – as was seen earlier, for bus companies staff costs account for around 65 per cent of operating costs. Examining the other costs for BA, direct operational costs such as fuel and oil, engineering and maintenance costs, landing fees

Table 5.5 *Operating costs, British Airways, easyJet and Ryanair, 2006*

Airline:	*British Airways*		*easyJet*		*Ryanair*	
	Actual	*%*	*Actual*	*%*	*Actual*	*%*
Staff costs	2346.0	30%	75.2	11%	171.4	13%
Selling costs	449.0	6%	26.0	4%	13.9	1%
Aircraft costs	2446.0	31%	366.8	54%	590.1	45%
Fuel costs	1632.0	21%	165.9	25%	462.5	35%
Other costs	955.0	12%	42.2	6%	85.6	6%

Source: Adapted from the respective company accounts

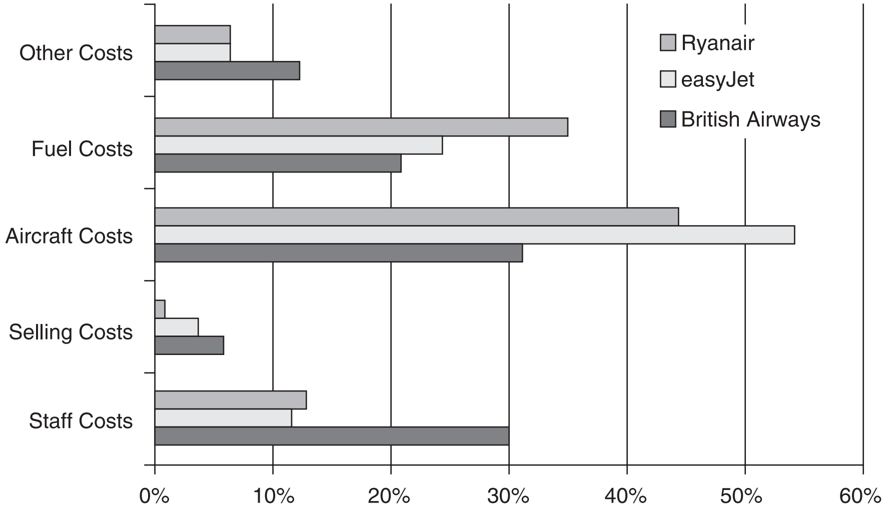

Figure 5.5 *Percentage breakdown, operating costs, British Airways, easyJet and Ryanair, 2006*

Source: Compiled from the respective company accounts

and so on account for just under 50 per cent of operating costs. It is in these two areas, staff costs and operational costs, where the biggest differences between BA and the two LCAs emerge. BA has a far higher proportion of operating costs ascribed to staff and a far lower proportion accounted for by direct operational costs than either easyJet or Ryanair. The other area of difference is in selling costs, although these on the whole are a small proportion of total operating costs.

This is where an understanding of the differences between fixed and variable costs becomes useful, as also does knowledge of how employment contracts are drawn up. Labour, for example, would virtually always be classified as a variable cost. Where employees are salaried, however, as in this case, then in the short run this is a fixed cost. A similar argument applies to depreciation and aircraft lease costs, as again in the short run capital equipment will be fixed and hence the depreciation or lease charge relatively fixed. Following this logic then, in fairly rough terms all operational costs are variable whilst all other costs, including labour and aircraft costs, are fixed. In the case of BA, therefore, around half of the costs are variable whilst for the other two airlines this is far higher at nearer 70 per cent.

Whilst the airline can do little about variable costs such as fuel and oil as these vary directly with output, these costs are zero while the aircraft is stationary. Fixed costs however are not. For example, cabin crew are still employed (being paid) after the aircraft has reached the airport gate and all passengers have disembarked, hence such costs are incurred even although the aircraft is technically not in revenue earning service. This is achieved only when the aircraft is in flight, therefore what becomes crucial in any such business model of low-cost operation is aircraft utilisation – the higher the utilisation, the more fixed costs are spread over output, the average cost reduces and the proportion of fixed to variable cost also reduces. Thus high turnaround times become the vital element in this type of operation. easyJet for example operate something like an absolute maximum of a 45 minute aircraft turnaround time and in most cases considerably less than that. Also what is important is that staff while on the ground, such as cabin crew, are actively employed to service the inside of the aircraft while it is stationary, hence they are being 'productive' during what would normally be 'dead' time. Note also that 'dead time' as such is far shorter due to fast turnaround times.

Another important aspect in reducing average cost is the lack of any slack in the whole system of operation. Spare aircraft are simply not an option, as this would constitute a fixed cost with zero output, hence any faults that develop in the fleet during the daily operation can have considerable and long-lasting (i.e. all day) knock-on effects on the whole system. Those with some experience of LCAs for example will be familiar with the announcement of 'flight delayed due to the late arrival of the incoming aircraft' which although a completely meaningless statement does highlight the lack of any slack in the whole system. Why LCAs can get away with such practices of course is due to the charging of low air fares.

The key to success in the operation of LCAs, therefore, is not 'low cost' per se but rather the attainment of a low average cost (per passenger carried). Few pilots for example work for low wages; however, the key to effectively reducing a pilot's wage is through increasing the number of hours worked (within regulatory limits of course).

In many senses, LCAs have changed the whole economics of airline operation where traditional thinking was in terms of an industry with a high proportion of capital costs and a relatively low level of variable costs. Furthermore, most variable costs were associated with take-off and landing, with respect to landing and terminal charges and most importantly fuel costs – a high proportion of the fuel consumed on a flight is during take-off and landing phases, sometimes as high as 50 per cent. All of these factors would suggest far higher unit prices on short-haul flights as the average cost of such a flight would be far higher.

All of these factors surrounding air transport economics however were known for some time, particularly the aspect of aircraft utilisation being the key to success in the airline business. Why they had not been fully exploited before is due to other factors, most notably market conditions and regulatory regimes. Facing such conditions, it made economic sense for the operator to restrict supply as this increased profits. In the LCA model, however, profits are maximised through low profit margins and high passenger volumes.

COST AND PRODUCTION IN THE LONG RUN

The first point to note about costs and production in the long run is that because all input factors are variable there is no division between fixed and variable costs. To briefly recap, whilst a firm may be planning a new production facility or entering a new market, it is operating in the long run. Once however it builds the new factory or sets up the new depot to service the new market, it is then operating in the short run (because at least one of the factors of production is fixed).

How costs behave in the long run is closely related to the behaviour of production and thus what happens when more inputs are added to the production process. There are some important differences between production and costs; however, these will be further developed later. To begin, Figure 5.6 illustrates a long run production function.

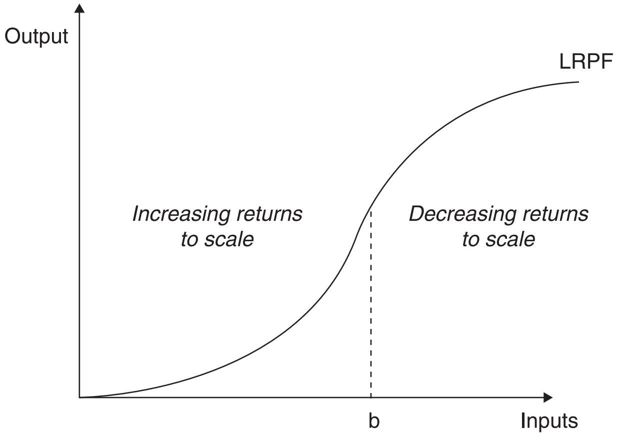

Figure 5.6 The long run production function

As with the short run, the long run production function is S shaped in nature. At first there are large gains when firm size increases – the relative percentage gain in output is greater than the relative percentage increase in inputs. Note also that this effect increases as firm size increases. These gains in total productivity, or increasing returns to scale, continue up to point b on Figure 5.6. Once firm size (as measured by the level of inputs) rises past point b, however, the proportionate gains from adding more inputs are not as large as before, hence the firm experiences decreasing returns to scale. This decline continues until increasing firm size has very little effect on the level of output. The main reasons for this general pattern of increasing and decreasing returns to scale that define the shape of the long run production function are outlined below.

Sources of increasing returns to scale

Specialisation of labour – larger firms allow more specialisation of the workforce to occur. If this is considered at the most basic level of a one-person sole-trader business, then the owner has to undertake all of the tasks involved in the running of the business. They therefore become a 'jack of all trades and master of none'. As firm size increases, more labour can be employed in specialised tasks and thus become more proficient at those tasks. Consequently productivity would be expected to increase. Note also that as this is the long run, there is potentially no ceiling on this effect and is probably best exemplified in Fordism large-scale production, where individuals can become 'experts' at very specific tasks.

Scheduling of inputs – this is similar to the specialisation of labour source but refers to the scheduling of all of the inputs, a factor that is particularly prevalent in the transport industries. As firm size increases, there exists greater flexibility in how the inputs can be combined and hence better utilisation of all of the inputs may be expected, i.e. higher total productivity. For example, in larger road haulage companies, there may be more flexibility in the scheduling of drivers to ensure that the vehicle stock is operated over the longest possible number of hours. This would ensure higher total productivity.

Capital inputs – this concerns a number of issues that are broadly grouped together here under the title 'capital inputs'. Firstly, some capital inputs can be very expensive to purchase. Only larger firms can afford to spend on these inputs, but they only do so on the basis that this will lead to improved efficiency (i.e. higher productivity) in the longer term. For example, increasing a railway line from single to double track increases capacity by a factor of four, hence potentially significantly increasing the productivity of rail services.

The second issue also relates to specialisation. Using the sole-trader example, then it may well be that the company vehicles consist of a single solitary van. This van will have to fulfil all of the transport requirements of the company, some of which it may be better suited to than others. As firm size increases, however, then the company fleet can be increased not only in size but also in scope, hence more suitable vehicles can be used for more suitable tasks. This should lead to higher productivity.

Thirdly, it may make more sense for larger companies to carry a larger number of spares and maintenance facilities, hence downtime of capital equipment as such should be reduced and consequently higher productivity achieved. It should be noted that even with 'just-in-time' production methods and the contracting out of maintenance and servicing facilities (as has happened with many bus companies), certain advantages with regard to this issue may still be expected. It may be anticipated for example that larger companies will have far more influence with the suppliers of such services that results in quicker response rates to service requests. Capital downtime, therefore, may be expected to be lower.

Indivisibilities – the standard example of an indivisibility is a telephone line. When setting up in business, a company will need to install and rent a phone line. With small expansions in size, there will probably be no need to install a second phone line, hence this 'input' is spread over a larger output. There will obviously come a point where a second phone line will have to be installed, but this would only be done if it was advantageous to do so, e.g. if it improved 'productivity'. Another example of an indivisibility of course is our sole trader's van. As firm size increases, the van's overall utilisation would improve.

All of these sources of increasing returns to scale should be reasonably clear, but what may cause this increase in productivity to tail-off, i.e. sources of decreasing returns to scale? This is particularly prevalent in the long run as there is no upper constraint in the form of a fixed input.

Sources of decreasing returns to scale

Loss of control – as firm size increases, there is a loss of control over the whole organisation. This loss of control decreases overall productivity. Under this heading it is worth highlighting the concept of 'X-inefficiency', originally devised by Leibenstein (1966). In simple terms, X-inefficiency relates to general management slack, and large (publicly owned) organisations are said to be particularly vulnerable to this concept. In less academic terms, in larger organisations there may be a loss of the 'sense of the individual' and more opportunities for general slack working practices, hence leading to lower productivity levels.

Geographical location – particularly prevalent in the bus industry, but true of other transport industries, when a firm initially sets up in business it will probably be on or near to the optimal location. Increasing size in the longer term means building other production facilities, such as depots, and these will not necessarily be at the best location. This can result in fairly long distances between the depot and the market served, hence a significant proportion of time is spent in driving vehicles between the two and not actually providing transport services. As a result, productivity decreases.

Administration procedures – large firms often require many more layers of middle and upper management, plus administration procedures, in order to control costs and processes within the organisation. This is commonly known as 'bureaucracy'; however, in this situation this should not be confused with 'red tape', as that is considered in the next section. More specifically, this refers to the time dimension that such 'form filling' requires and hence the opportunity cost of this form filling is the distraction of employees from the production process. When measured in terms of overall output, therefore, it requires on average a higher number of employees to produce a higher level of output.

Average and marginal costs in the long run

Having considered production in the long run, it is possible to see how costs would be expected to behave in the long run. These are graphed in Figure 5.7.

As can be seen from Figure 5.7, average costs at first fall as firm size (as measured by output) increases. This continues up to the point where average costs are minimised at the optimum level of production, known as the minimum efficiency scale (MES). After this point the trend is reversed and average costs rise as firm size increases. Along the part of the curve where the average cost is falling the firm would be said to be experiencing economies of scale. Very often these are incorrectly termed increasing returns to scale; however, returns to scale relates to production output while economies of scale relates to production costs. Along the part of the curve where average costs are rising the firm would be said to be experiencing diseconomies of scale, or again often incorrectly referred to as decreasing returns to scale.

This does once again pose the question as to why long run average costs should first fall and then rise as output (and hence firm size) increases. This is explained below.

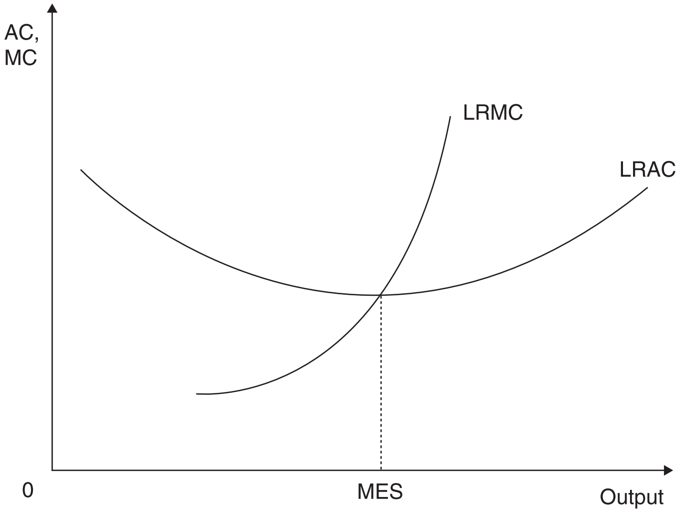

Figure 5.7 *Long run average and marginal costs*

Sources of economies of scale

Increasing returns to scale – as seen above, at first as firm size increases it encounters increasing returns to scale i.e. ever higher levels of productivity. This increased productivity therefore means that relatively lower levels of the inputs need to be employed to produce higher levels of output, hence the average costs per unit of output falls.

Bulk buying – larger firms can normally obtain some form of discount for buying capital equipment and supplies in larger numbers, and hence average costs would be expected to be lower for larger firms. Bus companies are a good example of this, where large group holding companies can negotiate discounts on fuel and tyres due to the sheer volume that the company will use in the course of its normal operations. Due to this volume, suppliers can afford to concede larger discounts and still return acceptable profits.

High cost inputs – these have already been examined under returns to scale, hence most are associated with improving productivity rather than directly reducing average costs. It is worth highlighting however that advertising is also sometimes cited as a high cost input, but it is debatable if this really leads to economies of scale. In the Cola market, for example, it almost certainly does. By advertising extensively, firms seek to increase the size of the market and hence allow them to increase production in order to take advantage of other economies of scale.

Financial economies – larger firms are normally better placed to secure additional finance as they can offer greater security. Interest rates therefore may be lower as there is a lower risk involved to the finance company, hence average costs are lower.

Sources of diseconomies of scale

Decreasing returns to scale – as highlighted above, there are a number of sources of technical inefficiencies that lead to reduced productivity for larger firms, thus the average output per unit of input falls. In order to produce higher levels of output, therefore, relatively higher numbers of inputs need to be employed, and this adds to costs causing diseconomies of scale.

Red tape – as noted above, larger firms often require many more layers of middle and upper management, plus administration procedures in order to control costs and processes within the organisation. This is 'red tape' and the actual cost of this added administration burden will add to costs, hence increasing the average cost per unit. Note however if we are being absolutely correct, the extra staff results in decreasing returns to scale (as proportionally less staff are directly involved in production), whilst the added costs associated with red tape such as stationery costs, the need for greater office space and more IT facilities and so on result in diseconomies of scale.

The coverage of costs in the long run is completed by examining economies of scale in railway operations and the impact that this can have upon how services are actually provided to the market.

Case study 5.3 Economies of scale and reform in railway operations

The general view of economies of scale within the rail industry is that, due to a high capital requirement in the provision of rail services, economies of scale are significant and hence company size needs to be large in order to capture all of these effects. In the past this was one of the main reasons which led to the nationalisation of railway industries across Europe (the first being Switzerland in 1901 and the last Britain in 1948), where most of the main-line railways were taken under the control of a single company so that economies of scale could be achieved. There were also a number of other important reasons for nationalisation; however, here we only concentrate on the economies of scale argument. This particular view of railway economics has become known as the traditional view (Preston, 1994), in which infrastructure and services are part of an integrated system and economies of scale in both are significant. Organisationally, therefore, services and infrastructure should be part of the same (large) company.

In recent years, however, virtually every European country has re-organised their railway systems, with many separating both organisationally and financially infrastructure from services, i.e. one company owns and operates the infrastructure whilst a different company owns and operates the rolling stock. This is known as a vertically separated railway. In Europe, Sweden was the first country to organise its rail system along these lines. All the infrastructure functions were separated out into the Swedish National Rail Administration (Banverket) whilst services remained within the Swedish State Operator (Statens Järnvägar). In contradiction to the traditional view outlined above, however, such a division is consistent with what has become known as the revisionist view of railway economics. Under this belief, the central premise is that economies of scale are only associated with the infrastructure and not with services. Therefore, scale effects will still be taken advantage of as long as infrastructure is retained as a single entity. As regards the size of operating companies, this is unimportant as there are no economies of scale associated with this activity. The advantage of such a system over the traditional railway is that different operators can operate on the network and hence some form of competition introduced into the provision of rail services.

Note that both the traditional and revisionist views are merely matters of opinion, and for a better understanding we need to examine the empirical research carried out on the topic. This would suggest that the theoretical U-shaped cost curve (and hence S-shaped production function) applies to railway operations. For example, Preston (1994) in a study of 15 Western European (integrated) railways found diseconomies of scale for larger train systems such as (West) Germany and Britain, and increasing returns for smaller rail systems such as Ireland, Switzerland and Belgium. In an updated study, Shires and Preston (1999) found that the MES (minimum efficiency scale) for integrated railways was around about the size of the Danish and Belgian rail networks. The implication, if only economies of scale are considered, is that countries with large networks such as Britain should organise their rail system into three to four integrated railways (perhaps on a regional basis) rather than as one single national operator. Such a structure would place each system close to the MES point and hence eradicate diseconomies of scale.

Within smaller railway systems, scale effects have been found to be even more substantial. Filippini and Maggio's (1992) study of the Swiss 'private' rail network revealed scale effects to be considerable, leading the authors to conclude that there were potentially substantial benefits to be gained from end-to-end and parallel mergers within the Swiss industry. Cowie (1999) also found scale effects to be significant in the Swiss industry. Switzerland is made up of a single national operator, CFF, on the main lines, and around 60 to 70 mainly publicly owned local railways on the local lines, varying in size from as small as 4 kilometres up to around 400 kilometres in length. Significant scale effects in such small systems would again be consistent with a U-shaped cost curve. This is because the gains that could be achieved in terms of lowering average cost for smaller systems by increasing output would be higher than for medium-sized operators (have a look at Figure 5.7 to confirm this).

To date, however, very little research has been carried out on the impact of separating operations from infrastructure on economies of scale. In one of the few such studies, Cowie (2002) examined the British train-operating companies and found that scale effects were significant, hence suggesting that the pure revisionist view, that there are no economies in operation, was not true. Size, therefore, with regard to how train companies are organised, is important. Scale effects however were found to be smaller than for integrated railways, with the MES point found to be around two thirds of the output level of Preston's earlier study based upon integrated railways. It nevertheless suggests that in Britain there should be around four to five train-operating companies rather than the current number of seventeen. The research also suggested by implication that there existed significant economies of scale in the provision of infrastructure.

Economies of scale have a major impact on the consideration of the best size of railway to produce rail services, as clearly if scale effects are considerable then costs can be significantly reduced by having a very large operator. Costs however are only one half of the profit equation, and hence where railways are heavy loss makers and thus heavily subsidised (as is the case in most of Europe), the attainment of economies of scale in the production of services whilst important is only one aspect amongst many other considerations when policy makers consider how 'best' to organise a rail system. For example there may be benefits in having a higher number of rail operators than the 'optimal' level due to an increase in competitive pressures, particularly for the market.

SHORT AND LONG RUN AVERAGE COSTS

We briefly end the chapter with a look at the relationship between the short and the long run average and marginal cost curves. This is shown in Figure 5.8, and is useful to further underline the relationship between all of these concepts.

In Figure 5.8, the long run average cost curve is a summation of a series of short run average cost curves. Beginning at an output level of Q_1, the firm is operating at point a on both the short run and long run average cost curves. Note that the short run average cost curve is tangential to the long run average cost curve at that level of output. Note also that at each point on the long run average cost curve there is a similar short run average cost curve which is tangential to it. If the firm was to increase production from Q_1 to Q_2 in the short run, however, i.e. where at least one of the inputs is fixed, average costs would increase to point c on the short run average cost curve. This would be considerably higher than if production was increased to Q_2 in the long run, in which case the average cost would be found at point b on the long run average cost curve. This is because in the short run the firm would encounter the law of diminishing marginal returns, and hence because one of the factors is fixed, say vehicles, the variable factor, labour, cannot be fully utilised. If however production is increased in the long run, then diminishing marginal returns will not be encountered and in this example, the firm will experience economies of scale as it approaches the

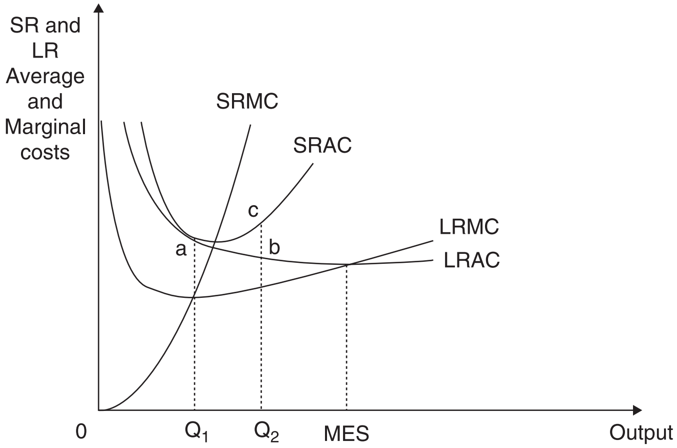

Figure 5.8 *Long and short run average and marginal cost curves*

MES point. An alternative way of viewing this is that short run costs are higher because when operating in the long run the firm effectively has a blank sheet of paper (in theory at least), and hence can plan for the lowest cost level of production. Once one of the factors is fixed, however, then it is more confined in what it can do and has to work around this constraint, which will incur a higher average (short run) cost. This relationship between costs in the short and long runs is important, particularly when examining transport markets. Very short increases in demand have to be met by a short run increase in supply, which in turn would incur higher costs. Given the peak nature of most public transport markets, the result is that firms can very seldom 'optimise' on the long run average cost curve, as supply needs to be flexible to meet these peaks in demand.

CHAPTER SUMMARY AND REFLECTION

In this chapter the focus has been on transport costs. This began with a definition of production, in which it was revealed that the definition of the short run is where at least one factor of production is fixed. We then examined the behaviour of costs both in the short and long runs, and found that both the short and long run cost curves were U shaped. In the case of the former this was due to the law of diminishing returns and the latter due to economies/diseconomies of scale. We finished the chapter by detailing the relationship between costs in the short and long run. A case study of the rail industry revealed that costs in practice would appear to follow the theoretical concepts, with studies indicating that average cost curves are U shaped in both vertically and integrated railway systems. Whilst not explicitly stated, it also highlighted the effect of costs on the structure of the industry, where for example if other things remained equal then it would be expected that more firms would operate in say the road haulage or bus markets than the passenger or freight rail markets primarily due to the differences in cost structures and the effects of economies of scale. What has also been highlighted is that costs are only one part of calculating the profit level and consequently only one factor in the planning of transport operations. Revenues, both in terms of that collected directly from the passenger and also sums paid by transport authorities in the form of transport subsidies, are also required since profit is revenue minus cost. The following chapter builds directly on the ideas outlined in this chapter and introduces revenue into the analysis in order to provide a complete insight into both sides of the transport market. Subsequent chapters will then consider the organisation of transport services and transport subsidy, both of which are heavily influenced by the structure of the costs of provision of transport services. From a public perspective, however, the financial costs, both in terms of capital and operating, need to be offset against not only the financial gains but also wider public benefits, and this topic is taken up further in Chapter 14 on transport appraisal

CHAPTER EXERCISES

Exercise 5.1 Technical, cost and allocative efficiency in bus operations

The following table gives some basic information relating to two small-scale bus operators:

	Company A	*Company B*
Number of buses:	3	5
Number of employees:	13	21
Average wage:	19000	16000
Vehicle kilometres run:	210000	300000
Bus cost per kilometre (including fuel costs):	0.36	0.30
Annual number of passengers carried:	370000	460000

a) Consider the following questions:

i Which of the two companies is more technically efficient in the production of bus services (note: you will need to separately calculate labour productivity and bus productivity and compare the two figures)?

ii Which of the two companies is more cost efficient in the production of bus services?

iii Which of the two companies is more allocatively efficient?

b) In your answer to part a. (i) you should have found that Company A had an advantage in both productivity ratios, hence by implication is more technically efficient. However, say this had not been the case. For example, if Company B had only 16 staff rather than 21 then it would have a superior labour productivity but an inferior bus productivity, hence there would be no clear answer (you should check this), as that would be dependent upon the balance of the two inputs used in the production process. In other words, such evaluations need some way of combining these ratios to come up with a single answer. One fairly simply way of doing this would be to weight each productivity ratio by that factor inputs share of costs. This would be a basic form of what is known as a Tornqvist productivity ratio. You should now calculate this index for both Company A and Company B using B's revised labour figure of 16 staff to determine which company is more technically efficient.

Exercise 5.2 Total, average and marginal products and costs

This exercise concerns the provision of rail services, and the task is comparatively straightforward if slightly involved. Quite simply, you have to fill in all the blanks, for which you will need the following information:

Fixed Costs:	100000
Price of a variable factor:	50000

You should round all figures to two decimal places.

Table 5.2a

Labour	*Output(000s)*			*Costs*				
Units	*TP*	*AP*	*MP*	*TFC*	*TVC*	*TC*	*ATC*	*MC*
0	0							
1	50							
2	110							
3	180							
4	260							
5	350							
6	420							
7	480							
8	530							
9	570							
10	590							

TP = *total product* AP = *average product* MP = *marginal product*
TFC = *total fixed costs* TVC = *total variable costs* TC = *total costs*
ATC = *average total costs* MC = *marginal cost*

Once you have completed this table, you should use your calculations to answer the following questions:

a) At what level of output should the firm operate at?

b) What is the most 'efficient' level of output in terms of:

   i Technical efficiency?
   ii Cost efficiency?
   iii In terms of measuring the firm's 'efficiency', which of these two measures should be used and why?

c) What units is the level of output measured in?

Exercise 5.3 Economies of scale in railway operation

Re-examine Case study 5.3 and answer the following questions:

1 Briefly outline your understanding of the traditional and revisionist views of railway economics.

2 List what you believe to be the main sources of economies of scale in the rail industry. Once you have produced this list, indicate which arise as a result of returns to scale and which are cost savings.

3 What on the other hand do you believe are the main sources of diseconomies of scale in larger integrated railways?

4 If you were a rail industry regulator in Britain today, what other factors apart from economies of scale would you take into account when deciding on the number of operators to have in the market?