CHAPTER 2

PRIVATE ENTERPRISE VERSUS PUBLIC ENTERPRISE THEORY

2.1 Introduction

In this chapter the theory that underlies the separation between the private and public sector enterprises is investigated. The assumption this chapter makes is that commuter rail will be concessioned to private sector operators. There should, therefore, be compelling arguments for the operation of rail to be shifted to the private sector. The arguments for and against the private sector on the one hand, and the public sector enterprises on the other, provide what is widely believed to be the differences between these two sectors.

This chapter begins by investigating private enterprise theory and public enterprise theory in terms of the principal-agent approach. The principal-agent approach is based on the work of Vickers and Yarrow and is extensively referred to in both section 2.2 and 2.3. The traditional debates on the two sectors are explored in section 2.4. The arguments look mainly at some empirical studies and the productive efficiency of rail systems. Section 2.5 then goes on to investigate the question of lack of competition in the market (lack of competition in the provision of rail services) in which public enterprises in general operate. In economic terms, the lack of competition in the market stems mainly from the monopolistic characteristics of public enterprises. The study investigates the relevant monopolistic characteristics and some empirical studies concerning economies of scale in the rail industry are mentioned. The chapter proceeds with the study of the X-efficiency theory in section 2.6 and the contestable market theory in the following section. Conclusions are given in section 2.7.

2.2 Private enterprise theory

The superiority of private over public enterprise is found to be based on the main assumption that the decision-makers involved aim to maximise profit (Vickers & Yarrow, 1989: 9). According to this source, if one investigates the relationship between the management of a private enterprise and the ultimate recipients of profits, one would end up adopting a principal-agent approach. In terms of the principal-agent approach, the shareholders are the principals, while the managers of the enterprise are the agents. It is therefore useful to consider some of the general features underlying the principal-agent approach and their implication for the behaviour of a typical enterprise that is transferred to the private sector. This is done by looking at the shareholders and the problems they experience in monitoring the activities of the managers of the enterprise; and the takeover of the private enterprise that serves as a constraint in the activities of the managers; and the possibility of a private enterprise going bankrupt, which could also serve as a useful constraint in the managers' activities.

2.2.1 Shareholders and monitoring problem

The assumption that is made in the case of private sector enterprise is that shareholders will seek to maximise their profit from the activities of the enterprise. The justification for this assumption could be found in the situation where shareholders hold a diversified portfolio of assets. In such a situation it is reasonable to assume that shareholders will be risk-neutral in respect of the return they receive from any individual firm. There are, however, possible objections to the assumption that shareholders are seeking to maximise profits from private sector enterprise. The first possible objection emerges as a result of asymmetry of information. For instance, even if it was true that each shareholder seeks to maximise profit, there would still be a lack of agreement by shareholders on their rankings of the alternative strategies that the managers put forward to realise maximum profit from the operations of the enterprise (Vickers & Yarrow, 1989: 11).

The second possible objection to the assumption that shareholders seek to maximise profits from their investment in an enterprise is, according to Vickers &

Yarrow (1989), if shareholders are also the users of the services and products that are provided by the relevant enterprise. In such a situation, the shareholders' interests will not just be limited to the managerial actions, which have a substantial impact on their returns. For example, if a high price was to be charged by the enterprise for the services rendered, it would increase the welfare of the shareholders in terms of the profits that they will earn. The same high price would, however, have the negative effect of reducing the shareholders' welfare, as they are also the consumers of the services produced by the relevant enterprise (1989: 12).

The third possible objection to this assumption arises from the implication of dispersed share ownership, which impacts on the effectiveness of shareholders in monitoring the activities of the managers. In a situation in which many investors hold shares in the enterprise, which is a characteristic of many modern enterprises, the activity of specifying and enforcing the contract between the shareholders and management is very difficult. As an example, consider a situation in which one shareholder engages in the activity of specifying the contract with management. Such a shareholder will end up paying all the costs and receiving only a small proportion of the profits, while the very same contract benefits other shareholders as well. This could result in a contract not being drawn up between shareholders and managers. In such a situation, managers may be said to have the discretion to pursue their own objectives, which may not necessarily be in line with the interests of shareholders. As a result, it may not be appropriate to base the analysis of private enterprise behaviour on the profit maximisation assumption (1989: 15).

Vickers and Yarrow conclude that it would be premature to accept other private enterprise behaviour that differs from the profit maximisation hypothesis, especially if shareholding is dispersed. Uncritical acceptance of the notion that managers of private enterprises will act in the best interest of shareholders to maximise profits is equally not acceptable (Vickers & Yarrow, 1989: 15).

2.2.2 Takeovers

Vickers and Yarrow (1989: 15) mention that the management of a newly privatised entity faces a number of shareholders who seek to introduce an incentive structure that aims at maximising profit. As a result of the size and distribution of shareholding in the entity, ownership can change quickly because of investors buying and selling decisions. At any time an individual or institution can seek to purchase all the shares by making a takeover bid for an enterprise. If successful, the bid would concentrate ownership and eliminate multiple holding. Consequent to the marketability of shares, dispersion of shareholding can be argued not to be a factor to be considered. In such a situation, if the management of the enterprise fails to act in ways that are consistent with the shareholder's interest of profit maximisation, the share price might fall and the costs of purchasing such shares would decline as well. In this type of scenario, existing management will at some point become vulnerable to a takeover raid and ultimately be replaced by new management. The threat of takeover acts as a mechanism that deters management of a private entity from pursuing policies that differ substantially from the interests of its shareholders. It is, however, assumed firstly that the objective of the acquiring entity is the expected profit maximisation and, secondly, that takeovers are triggered by management behaviour that differs from that of its shareholders. These assumptions are open to question. In the first place, raids may be motivated by a desire on the part of management to increase its utility rather than the attainment of the objectives of its shareholders. The takeovers should therefore be seen as a potential instrument for maximising managerial utility as well as a constraint on such behaviour. Furthermore, it is by no means clear whether, in the light of the high level of takeover activity in the private sector, this is in the interests of shareholders. In the second place, the fact that opens these assumptions to question is that even if the raiders are profit seekers, takeovers may be motivated by factors such as increased market influence. Limitations to the takeover constraint may come into question where the target firm is very large (Vickers & Yarrow, 1989: 15–24).

2.2.3 Bankruptcy

The possibility of bankruptcy for private enterprise may be seen as another means by which managers (agents) may lose control of the entity. Bankruptcy can therefore be viewed as an alternative of a takeover constraint. In discussing the bankruptcy constraint, one needs to take account of its special features like the loss of control that occurs through the circumstances under which bankruptcy is likely to occur; the fact that proceedings may be initiated by a different group of economic agents such as creditors; and the legal and regulatory framework governing the bankruptcy process. Bankruptcy can be assumed to occur when the market value of the firm's assets falls below the value of outstanding liabilities (Vickers & Yarrow, 1989: 24).

These authors mention that there are two limitations to the bankruptcy constraint. The first is when management believes that there is a chance of an enterprise going out of business. Under such circumstances management will conclude that it should be given more managerial discretion by the shareholders. If such discretion is granted, there will always be a limit to the debt of the enterprise. At the debt limit level, the usefulness of the bankruptcy constraint as a control mechanism will have been exploited to the fullest possible extent. Further improvement to the efficiency of the firm will not be feasible if reliance is placed on the bankruptcy mechanism (1989: 26).

The second limitation to the bankruptcy constraint is that, in practice, the determination of the enterprise's debt is a decision that is usually delegated to management (agent). As a result, management can ease this constraint and increase its own utility by choosing a lower debt level than that which shareholders wish to see. It is as in the presence of factors such as depressed demand or intense product market competition that bankruptcy will play an important role. Consequently, it is unlikely that the bankruptcy control mechanism will have much effect on the incentives for internal efficiency, especially in a privately owned monopoly (Vickers & Yarrow, 1989: 26).

2.3 Public enterprise theory

According to Vickers and Yarrow (1989: 27), public enterprise theory is based on the assumption that government seeks to maximise the economic welfare of a country. In terms of these authors, the rationale for this approach is that bodies such as public enterprises are themselves agents of the government and should therefore act in the best interest of the broader public. As in the case of private enterprises, an examination of public enterprise in terms of the principal-agent approach is necessary to assess whether the assumption that public enterprises act in the interest of the wider public provides a sound basis for analysing its behaviour.

In a public enterprise situation, there are two distinct groups of public officials that are involved in monitoring the activity of public enterprise. The first group of public officials is the politicians and the second group is the civil servants. The full monitoring hierarchy, however, consists of the general public, the elected political representatives, non-elected civil servants and managers of publicly owned enterprises (Vickers & Yarrow, 1989: 30).

Vickers and Yarrow (1989) mention that, in considering the relationship between the general public and its elected representatives, it is unlikely that the preferences of the politicians can accurately be captured by a simple and general objective function. One feature of the problem that is prominent in this relationship, particularly for the politician, is the relative insecurity of tenure of office that is involved. Politicians of a given party have a common interest in electoral success. Given this situation, it is likely that promotion and demotion of a relevant minister during tenure of office will depend upon that minister's contribution to the electoral prospects of his own party. If it is assumed that the utility of the minister is much higher in office than out of it and that the effects on the utility of changes in other variables are smaller (such as the improvement in the efficiency of the relevant public enterprise), it would suggest that a useful starting point for analysing political behaviour is the hypothesis that decisions are taken with a view to maximising the probability of electoral success. Assuming the hypothesis that political decisions are taken with a view to maximising electoral success, it can be maintained that politicians will seek to achieve economic efficiency with respect to public enterprise. This is because if resources are not efficiently utilised in the public enterprise, there will be scope for improvement in the economic welfare of some sections of the public without making others worse off. The improvement of efficiency in the public enterprise concerned would have positive effects on the prospects of the political party that is in power. The authors cited mention that the prospects of re-election would, however, depend on the extent to which voters are informed about decisions that were taken on their behalf. In practice, however, there are considerable information asymmetries between politicians and the electorate. This therefore implies that the minister responsible for monitoring the relevant state-owned enterprise (agency) may no longer derive electoral benefits from the improvement in the economic efficiency of the entity concerned (Vickers & Yarrow, 1989: 30–31).

Turning to the point of non-elected civil servants, Vickers and Yarrow mention that these officials support their minister by undertaking detailed monitoring of public enterprises. In the case of officials, insecurity of tenure of office does not play a central role. Their activities are in any case monitored by the minister in charge of their department. The monitoring of civil servants by ministers is, however, limited by the asymmetry of information between civil servants and the minister who may stay a relatively brief period in office; incentives for ministers to search for performance improvements that are generally weak since the pay-offs that result are unlikely to have much impact on electoral prospects; and factors that increase the economic welfare of civil serva nts are likely, all things being equal, to have positive effects on ministerial welfare (1989 :33).

Vickers and Yarrow (1989) conclude their analysis of public enterprise in terms of the principal-agent approach by identifying four potential sources of sub-optimality in the control of state-owned enterprises. The first is the displacement of social objectives by political objectives; the second is a tendency for direct political intervention in managerial decisions; the third is the internal inefficiencies in the bureaucratic arrangements and, lastly, the inefficient levels of bureaucratic activity (1989: 34).

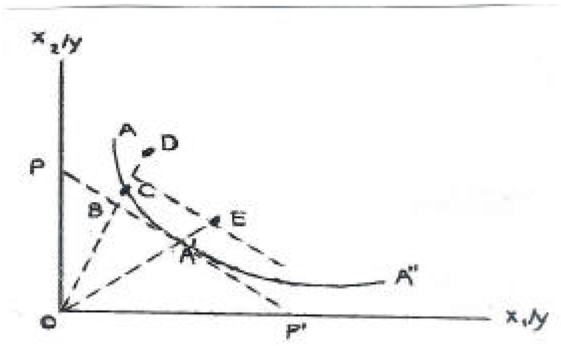
2.4 Traditional private versus public enterprise debate

In terms of Domberger and Piggott (1994: 35), the case for private ownership rests on the incentives and the constraints that the market provides to promote efficiency within the firm. The type of efficiency achieved is described as technical or productive efficiency and is achieved through cost minimisation for a given level of output. In the case of public enterprise economics, however, the approach has assumed that productive efficiency will be satisfied irrespective of ownership or competition.

Domberger and Piggott (1994) assert that productive efficiency can be thought of as having two distinct requirements. The first requirement has two related subsets. The first subset is that when all input levels are fixed, the minimum quantity of any input will be used in the production of a given level of output. The second subset is that inputs must be used in cost-minimising combinations that can only be determined by reference to the relative factor prices. Public enterprises fail to fulfil the first requirement because of overstaffing and overcapitalisation in the absence of incentives to minimise the costs of inputs. Excessive expenditure on labour within public enterprises arises because workers are in most cases underutilised and are given generous fringe benefits over and above the competitive wage. This is known as feather-bedding. Domberger and Piggott (1994) believe that the generous fringe benefits represent a redistribution of income from the consumers to public enterprise employees. The second requirement of productive efficiency is rapid adjustment of capital-labour ratios to changes in the relative factor prices. This introduces dynamism to the enterprise. In as far as the second requirement is concerned, public enterprises face obstacles to such adjustments, especially from the strong employee unions that are operative in the field of the relevant state-owned enterprise (1994: 36).

In investigating the concept of productive efficiency, Oum, Waters II and Yu (1999: 11) mention that productive efficiency consists of technical and allocative efficiency and illustrate the concept with the following figure.

Figure 2.1: Technical and allocative efficiency



Source: Oum, Waters II & Yu (1999: 12)

In Figure 2.1 above, the firm produces a single output y by using two inputs, x_1 and x_2. It is assumed that a firm has a constant return to scale and the production frontier is $1 = f(x_1/y,\ x_2/y)$. In Figure 2.1, an isoquant curve is represented by ACA'A". Isoquant is defined as "the set of input combination that can be used to produce a given level of output" (Train, 1991: 23). The firm that produces above the isoquant uses more than is necessary of at least one input. Suppose that the available budget is represented by curve PP' which is the isocost. The isocost curve is tangent to ACA'A" at point A'. Oum et al. mention that a firm is technically efficient if it chooses an input mix on the isoquant and it is allocatively efficient, that is price efficient, if the marginal rate of substitution between inputs is equal to the corresponding input price ratio (Oum et al., 1999: 11). According to Oum et al., technical inefficiency results from excessive use of inputs to produce a given level of output, and allocative inefficiency results from using inputs in wrong combinations. Productive efficiency in terms of these authors requires joint satisfaction of technical and allocative efficiency conditions, which are obtained at point A' in the above diagram. Firm A, C and A" are technically efficient as they use an input combination that falls on the isoquant, but allocatively they are inefficient in the above figure. Further technical efficiency is equivalent to 1 if the firm is on the isoquant. In the case of D in the diagram, technical efficiency is less than 1. In general, therefore, 0 ≤ technical efficiency ≤ 1 (Oum et al., 1999: 11-12).

Domberger and Piggott (1994: 36) further mention that there are three reasons why we might expect state-owned enterprise not to strive for cost minimisation. The first reason is the lack of a clear-cut profit objective, which is the overriding goal of private enterprise. The second reason arises from the fact that state-owned enterprises are often assigned a number of (conflicting) objectives among which cost minimisation takes low priority, such as when the government directs the management of public enterprise to pursue non-commercial objective for political reasons. The third, according to these authors, is that state-owned enterprise management's incentives are not compatible with the pursuit of efficiency since neither management earnings nor tenure is directly related to any measure of efficiency. Advocates of private ownership argue that management incentives promote efficiency. In terms of these authors, the threat of bankruptcy is perhaps the most important and they argue that private ownership frees the enterprise from the prospects of political intervention in managerial decision-making and this should count strongly in its favour. With private ownership there is a clearly defined profitability objective and clearly observable indicators of performance, like the stock market price or share price. It should, however, be borne in mind that private ownership does not always guarantee that maximum productive efficiency will be attained. Under private enterprise management, incentives and constraints on managerial behaviour may be blunted by the lack of competition in the product market and it may combine with the divorce of ownership control, where control is delegated to management. Seen from the perspective of the principal-agent approach, one characteristic of large modern corporations is that ownership is dispersed among many shareholders, while control of the enterprise rests with a small management group in the enterprise (Domberger & Piggott, 1994: 38).

Still investigating the issue of productive efficiency, Oum and Yu (1994) undertook an empirical study to determine the economic efficiency of 19 railway systems of Organisation for Economic Co-operation and Development (OECD) countries. The data used for most of these rail systems was of the period 1978-89. Since the focus was mainly on passenger rail systems, Oum and Yu chose the systems in the OECD countries that had a high percentage of business in passenger activity. These authors compared and reconciled the results of the productive efficiency obtained with two distinct concepts of output, that is, revenue outputs and available outputs. The sample featured a variety of institutional regulatory frameworks, and market and operating environments and the management of some of the rail systems enjoyed substantial freedom from the regulatory authorities, especially when making strategic and operating decisions. Some systems were subject to strict government control (Oum & Yu, 1994: 121–125). In providing conclusions, Oum and Yu mention, among other things that, firstly, railway systems with high dependence on public subsidies were significantly less efficient than similar railways with less dependence on subsidies, and secondly, railways with a high degree of management autonomy from the regulatory authorities tend to achieve high efficiency. These two concluding remarks imply that the productive efficiency of railway systems may be significantly improved by an institutional and regulatory framework that provides more freedom for managerial decision-making and that the institutional and regulatory framework for the rail industry must address the question of managerial autonomy. Oum and Yu (1994) point out that subsidy policy must encourage railways to use normal market mechanisms and to improve cost recovery while using subsidies to improve the services. Furthermore, the empirical results show that efficiency measures may not be meaningfully compared across rail systems without controlling the effects of the differences in the operating and market environments. Because of the limited information, the study examined the effects of government intervention in the rail systems and subsidisation in a broad sense. As a result, one needs to be careful as there are many alternatives and complementary ways that the government can use to intervene in railways, such as partial or full ownership of rail, operation of rail system by government branches, regulation of prices and service performance. These different types of intervention by government in the railways have a distinct effect on the efficiency of firms. The important point is that it is not only the amount of subsidy that affects management incentives in improving efficiency but also the method of subsidisation. As a result, a predetermined subsidy amount is considered more effective than subsidising the entire deficit (Oum & Yu, 1994: 137).

Oum, et al. (1999: 9) provide a comprehensive survey of the methodologies that are used in measuring and comparing productivity in the rail industry. These authors mention a number of empirical studies undertaken in terms of the different methodologies, such as the index number procedure, which includes partial factor productivity, total factor productivity; and conventional econometrix methods (1999: 25), as well as frontier econometrix methods (1999:30). In summarising and concluding their study, Oum et al. (1999: 34) quote Dogson (1985) and Hooper (1987) that the important aim of productivity studies is to compare efficiency between the private and government owned enterprises and to compare the efficiency of firms that operate in the regulated and unregulated environments. The conclusion is that in most of the rail productivity and efficiency studies that they have surveyed, virtually all the studies have concluded that increased competition through regulatory liberalisation and deregulation has improved productive efficiency. Oum et al. (1999) mention the Canadian railways, which achieved higher productivity than the US during the period 1960 and 1970 as a result of Canada's liberalisation of its rail pricing regulation. According to these authors, US rail productivity improved after the implementation of the Staggers Act of the 1980s, which introduced significant deregulation (Oum et al., 1999: 34).

Oum et al (1999) mention that many studies on European railways have Oum et al (1999) mention that many studies on European railways have investigated the effect of the managerial autonomy of government owned or mixed owned rail enterprises on efficiency. To a large extent such studies have concluded that efficiency is positively influenced by management freedom (Oum et al., 1999: 34).

Ramanadham (1988: 12) argues for private ownership. The argument is concerned with firstly, the ability of the private enterprise to reduce the financial deficits sustained by the state-owned enterprises; secondly, the potential of private ownership to reduce the needs of state-owned enterprise for funds (subsidy) from the government; thirdly, the view that distributional outcomes of state-owned enterprise operations are not always conclusively of the desirable kind; and fourthly, the relief that private ownership offers to civil servants who are overburdened with their involvement in the state-owned enterprise matters. Related to the latter argument, Ramanadham noticed that there is duplication and in the end often amateurish application of civil servants' time and energy to approvals or disapprovals of investments in public enterprises. In some cases, senior civil servants sit on the boards of directors of many public enterprises and, as a result, several management controls, whose exercise should ideally rest with the relevant enterprise, get externalised at the level of the government department (1988: 16).

Authors like Galal, Jones, Tandon, & Volgelsang (1994: 10) ask whether ownership matters. The authors mention that there is a huge amount of theoretical literature on public and private differences that draws on property rights, transactional costs, contract theory and public choice tradition. According these authors, hypothesised differences between public and private enterprise can be assigned, firstly, to differences in objectives, and secondly, to differences in constraints. In considering the differences in objectives, Galal et al. (1994) mention that private enterprise pursues the profit objective while the state-owned enterprise may pursue whatever the government may want to provide and is able to finance. On the one hand, this may mean that the public enterprise can promote consumer welfare by not exploiting its monopoly position. On the other hand it may mean that public enterprises can instead promote the welfare of the politicians by hiring a large number of redundant workers.

In considering the constraints faced by the enterprises operative in the two sectors, Galal et al. (1994: 10) mention that even if the objectives were the same for both sectors, private and public enterprises would face different constraints in the pursuit of their objectives. The stories of public enterprises that are forced to operate with insufficient autonomy are legion. The constraints in the private sector are less commonly mentioned, but are manifest in the frequent claim, especially in developing countries, that large, capital-intensive projects require public enterprise because smaller private sector enterprise in developing countries does not have sufficient access to capital. Galal et al. (1994) mention that an important constraint when property rights are held by the state is that of incentives, because no individual is given the incentive to exert effort to see that resources are used efficiently. A lack of sufficient application of management effort therefore results in high cost production by the state.

Galal et al. (1994:11) maintains that the most striking characteristic of the oldest and largest body of empirical literature that compares public and private enterprise is its almost laughable diversity of results. There are two broad sets of conclusions that have emerged from the empirical literature. The first set found private enterprise to be clearly superior and quotes Boardman and Vining (1989: 17) that there "is (robust) evidence that state enterprise and mixed enterprises are less profitable and less efficient than private enterprise". Galal et al. (1994) quote Boycko, Schleifer and Vishny (1993: 1) that "there is virtually universal consensus that privatisation improves efficiency". Still related to the empirical literature that found private enterprise to be superior, Galal et al. quote from Bennett and Johnson (1979: 59) that "without exception, the empirical findings indicate that the same level of output could be provided at substantially lower cost if output were produced by the private rather than the public sector". The second body of empirical literature draws a rather different conclusion. The most influential formulation is the work by Caves and Christensen (1980: 974) and Galal et al. quote from this work: “contrary to what is predicted in the property rights literature, we find no evidence of inferior efficiency performance by government owned railroads ... public ownership is not inherently less efficient than private ownership. The oft-noted inefficiency of government enterprises stems from their isolation from effective competition rather than their public ownership per se”.

Galal et al (1994: 13), in reconciling the two sets of empirical literature, find that there are three sets of factors that are involved. In the case of the first set, a few of the empirical studies find private enterprise to be superior for illegitimate reasons in that they compare competitive enterprises with monopolistic enterprises. Furthermore, these studies find private enterprise to be superior for legitimate reasons because such studies compared reasonably competitive enterprises. Also, when public and private monopolies are compared, the results would always be over the map. This result would be such because of the fact that the two enterprises will be pursuing different objectives. The second set that reconciles the differences in the conclusions of the empirical studies involves the variables used to measure performance. The profit measure will always favour the private firms because the state-owned enterprise operates in a different environment, that is, one of promoting economic welfare. A further problem with performance measurement is that, on the efficiency side, most of the studies measure the changes in total factor productivity (TFP). According to these authors, if this is similar in public and private sector enterprises, it does not mean that the two sectors are equally efficient but rather that whatever differences may exist, such differences are neither widening nor narrowing. The third reconciliatory factor for the different conclusions reached by empirical studies is the small number’s problem. In the case of large monopolies, a country may have say one rail enterprise and say one steel mill. If these enterprises were in the public sector in one country and the private sector in another country, it would be difficult to ascribe the difference in performance to ownership rather than the differences in the economies and cultures of the respective countries. The problem here is mainly of comparing like with like, with the result that it would be difficult to prove convincingly the existence of public and private differentials in the case of large monopolistic enterprises. These are therefore the reasons for the contradictory and ambiguous results of the empirical literature on public and private enterprise (Galal et al., 1994: 14).

Laffont and Tirole (1993: 640) have identified various factors associated with the conventional wisdom for differences in public and private ownership. These authors assign costs or benefits to the various factors that are identified. Benefits of public ownership are attributed to the social welfare objective and centralised control. The cost or disbenefits of public enterprise are an absence of capital monitoring, soft budget constraints, expropriation of investments, a lack of precise objectives and lobbying.

2.5 Competition and public enterprises

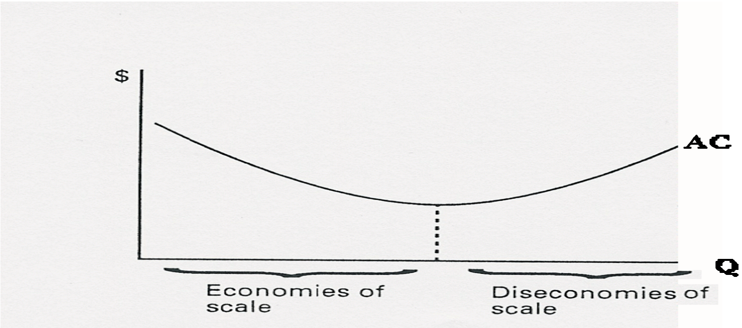
An economic analysis of state-owned enterprises is traditionally concerned with pricing and investment policy, which have a welfare economic orientation. This is as a result of a lack of competition and allocative efficiency. The conditions needed for allocative efficiency to be attained include the well-known marginal cost pricing (Domberger & Piggott, 1994: 33). If a firm is operating in a perfectly competitive environment, it will be a price taker because it cannot influence the price of the services that are offered (Mohr, Fourie & Associates, 1995: 353).

The natural monopoly is a characteristic of the environment in which many state-owned enterprises operates. The market failure of various types is considered to be so endemic to the sectors in which state-owned enterprises have traditionally operated that government intervention was necessary. However, government intervention, which was supposed to tackle the problem of market failure, brought with it the challenge of regulatory failure with the result that the choice of policy in this sector involves the balancing of risks, the resultant market failure and regulatory failure. In the situation where a natural monopoly therefore precludes the introduction of competition within the market, it is possible and desirable to introduce competition for the market through, say, a concessioning mechanism. In a natural monopoly situation, economies of scale or scope may warrant a single firm to serve the market at lower unit costs than if several firms serve the market. A natural monopoly occurs in the industries that involve extensive distribution networks, like the railroads (Thompson, 1988: 41–43).

The question that arises in connection with the assertion that a natural monopoly is a characteristic of the environment in which state utilities operate is what are the features of a natural monopoly. There are two features of a natural monopoly and they are the economies of scale and economies of scope (Train, 1991: 5).

Figure 2.2 below shows the economies of scale as well as the diseconomies of scale.

Figure 2.2: Economies and diseconomies of scale



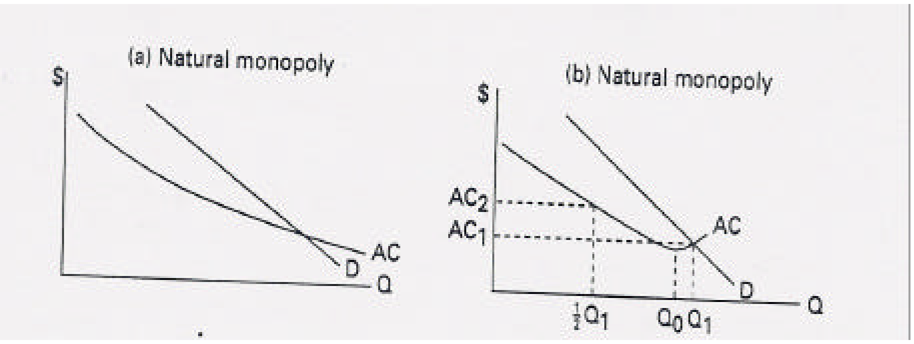
Source: Train (1991: 6)

In terms of Figure 2.2, economies of scale exist when the average cost curve slopes downward, indicating that average costs fall as output increases. According to Train (1991: 6), economies of scale are caused by fixed costs. Fixed costs are the costs that must be incurred no matter how many units of output are produced. When output expands, such costs are spread over more units so that average costs decline. An important issue to remember, according to Train, is that economies of scale can exist over certain ranges of output, but not other ranges.

At low levels of output, scale economies may be present; at larger levels of output, diseconomies of scale may be present; and at larger levels of output, diseconomies of scale may set in. This gives rise to the U-shaped average cost curve as shown in Figure 2.2.

According to Train, the prevalence of a natural monopoly depends on the range of economies of scale relative to market demand (1991: 6). The prevalence of a natural monopoly can be illustrated by using the following figures.

Figure 2.3: Relation of average costs to demand



Source: Train (1991: 7)

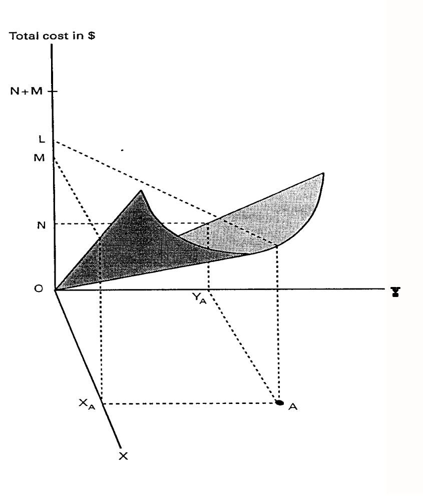
Figure 2.3(a) depicts a standard situation in that average costs decline over all levels of output that can be demanded at any price. A natural monopoly exists under such a situation. A natural monopoly can also exist with economies of scale over a smaller range of output, such as in Figure 2.3(b). In terms of (b), economies of scale continue up to the Q_0 level of output after which diseconomies of scale follow. One firm can produce quantity Q_1 with average costs at AC_1. If two firms were to produce this output, each would incur an average cost of $AC_{2,}$ which is greater than AC_1. The high cost of AC_2 applies in Figure 2.3(b) if the two firms share the market equally. If they were to split the market unequally, their average costs would differ, but their total costs would exceed the costs of one firm (Train, 1991: 7). This means that it would be uneconomical to have the two firms serve the market as one could do so more economically from a cost perspective. This explains the existence of a natural monopoly.

Friedlaender, Berndt, Chiang, Showalter and Vellturo (1993), in discussing the nature and extent of economies of scale specifically with regard to US Class 1 railroads, mention that the concept of economies of scale relates to "change in the firm's level of cost to change in its level of output" (1993: 138). According to these authors, this refers mainly to the elasticity of cost (E_{cy}) which "reflects the percentage change in cost relative to percentage change in output (dC/C)/(dy/y)" (1993: 138). According to these authors, diseconomies of scale exist if elasticity of cost is greater than 1, while economies of scale exist if elasticity of cost is less than 1, and constant return to scale exists when elasticity of cost is equal to 1. It is usually said that a firm is subject to increasing return to scale if economies of scale are greater than 1. If economies of scale are less than 1 then the firm is experiencing a decreasing return to scale, and a constant return to scale if economies of scale are equal to 1 (1993: 141). Friedlaender et al. mention that in discussing economies of scale in the rail industry it is important to differentiate between output-related economies of scale that emerge as a result of changes in output and size-related economies that emerge from changes in the technological environment like change in the network size in which the rail operates. This distinction is essential because the output-related economies of scale are conditional on existing capital stock and size-related economies are conditional on the optimal capital stock, given the level of output. Friedlaender et al. focus on return to scale associated with output, particularly tonnage, and mention that this is typically referred to as economies of density (1993: 141).

McGeehan (1993: 19) also undertook the empirical studies of railway costs and productivity growth. McGeehan mentions that "as observed by Keeler (1974), if the network configuration is held fixed, then economies of scale resulting from increased traffic volume is defined as economies of density. Economies of density measure the relationship between unit costs and the intensity of utilisation of capacity" (1993: 23). According to this author, the economies of density are given by the equation $ED = 1-(\partial 1nC/\partial 1nY+\partial 1nC/\partial 1nQ)$. Where ED is greater than zero, positive economies of density exist and if ED is smaller than zero, diseconomies exist. McGeehan's study found the elasticity of costs with respect to output to be 0.246. This indicated that the rail network studied, namely Æorus Iompair Éireann (CIE) rail, had positive economies of density. Thus, given fixed network capacity, as output in terms of ton miles and passenger miles increased, unit costs of production increased less than proportionally (1993: 28). This author mainly discusses the first feature of natural monopoly, which is the economies of scale.

The second feature of a natural monopoly is economies of scope. When more than one good is produced, a natural monopoly can arise from economies of scope as well as economies of scale. In the production of several goods, there are sometimes either shared common facilities or shared equipment that makes it less expensive to produce the relevant goods together than when they are produced separately. As a result, economies of scope exist "if a given quantity of each of the two or more goods can be produced by one firm at lower total cost than if each good were produced separately by different firms" (Train, 1991: 8). To give more meaning to the definition of economies of scope, suppose that the total cost to a firm to produce two goods in the quantities of X and Y can be represented as f(X,Y). The cost of producing X only is therefore f(X,0). By the same token, if Y only is produced, the cost is f(0,Y). The definition says that economies of scope exist if f(X,Y) is smaller than f(X,0) + f(0,Y). This can be represented by the following diagram.

Figure 2.4: Economies of scope



Source: Train (1991: 9)

In Figure 2.4, the cost function is shown as the shaded surface. This gives the cost of producing any combination of the two goods. Point A in the figure represents a combination of producing X_A and Y_A. Costs at point A are $f(X_A,Y_A\ )$ for producing both goods. The cost of producing both goods is OL on the cost axis in Figure 2.4. If the relevant firm produced quantity X_A only and no Y, then the resultant cost would have been $f(X_A,0)$, which is the distance OM in the diagram. If another firm produced Y_A only and no X, its costs would have been $f(0,Y_A)$ which is the distance ON in Figure 2.4 above. The combined cost of the two firms would thus have been ON + OM. Since this is higher than L, it would be more costly to have separate firms produce each quantity than for one firm to produce the quantity of both goods (Train, 1991: 8). Train further mentions that, like economies of scale, it is possible for economies of scope to exist at some levels of output of the goods and not other levels. From a cost perspective, whether to have one firm produce depends on how economies and diseconomies of scope relate to the demand for the two goods. The existence or relevance of economies of scope often depends on how goods are defined (1991: 10). This may have some implications for the current South African passenger rail arrangements especially with regard to commuter and long distance services. How this is answered requires an empirical study.

Further economies of scope can exist with or without economies of scale and vice versa. As a result, it is possible for joint facilities to be used in the production of two goods and yet, by expanding production of both, raise costs more than proportionally. Whether a natural monopoly exists depends on the overall cost situation taking into account economies or diseconomies of scope and/or scale. Economists use the term "subadditivity" for this purpose. "A cost curve is said to exhibit subadditivity at a given level of one or more outputs if the cost of producing the output is lower with one firm than with more than one firm, regardless of how the output might be divided among the multiple firms" (1991: 11). Cost subadditivity essentially means that a natural monopoly exists. This throws more light on understanding the economies of scope as a feature of natural monopoly.

Returning to public enterprise and competition, the International Encyclopaedia of Business and Management (1996, s.v. "Privatisation and regulation") mentions that the

> … trend towards privatisation has found its reflection inevitably in economic theory. The previous era, which witnessed significant degrees of nationalization and state ownership through the Organization for Economic Cooperation and Development (OECD) economies, was generally interpreted in mainstream economics in terms of various forms of 'market failure' requiring state intervention. The mainstream response to privatization has therefore been to interpret the changing industrial ownership structure in terms of 'government failure' or 'public failure' which in turn must have come to outweigh the original market failure.

This quote can also be said to be applicable in our situation, especially if one takes into account the history of railways in South Africa, as well as the statements made in the White Paper about the role of government, which was quoted in the introductory chapter. The prevalence of market failure was generally cited as a justification for state ownership of various enterprises like railways. According to this quotation, the situation has changed, however. Anonymous (1994: 27) has noted that while the majority of railways are still public enterprises, restructuring and some other forms of privatisation have taken place or are currently ongoing. This trend, as noted by the International Encyclopaedia, is presently being interpreted as failure of the government or public failure to ensure the delivery of effective and efficient rail services.

Crampes and Estache (1997: 3) mention that in considering monopoly, both economists and politicians are against private monopolies, but they differ in their reasons for being against them. Economists do not like private monopolies because they do not provide enough goods or services to the market, which results in a dead-weight loss of surplus. The politician sees private monopolies as bad because their prices are too high and they appropriate profits from the consumers that buy their services.

Crampes and Estache argue that when the economist exerts pressure on the monopolist to increase output, the politician usually urges the firm to cut prices. Here at least both the economist and the politician would agree. It is therefore the only area where the economist's and the politician's views converge for the purpose of achieving efficiency. According to Crampes and Estache, politicians have little trouble with price reductions without any change in output. The economist will, however, disagree since it will create the rationing of output without any improvement in the social surplus. The economist can resolve the problem of the inefficiency of a monopoly by allowing it to fix perfect discriminatory prices. Applying price discrimination will, however, be rejected by the politician on the grounds that everyone should pay the same price and that this will be too much in the favour of the monopolist (Crampes & Estache, 1997: 4).

In as far as the regulation of a monopoly is concerned, Crampes and Estache mention that both the politician and the economist agree on the need for regulation, but differ on how to do it. According to the authors, the economist would favour any solution that promotes efficiency if it is feasible, while the politician would be interested in household welfare in so far as it improves the prospects of re-election. As a result, political office bearers prefer solutions that are inexpensive for the public budget and that protect social equilibrium. According to Crampes and Estache, politicians therefore can choose market solutions that enable them to keep control of the operations and the development of the firm. They argue that these compromises may explain why a concession regime is increasingly of interest to researchers in the provision of public services (1997: 4).

2.6 Allocative efficiency versus X-efficiency

"At the core of economics is the concept of efficiency. Micro-economic theory is concerned with allocative efficiency" (Leibenstein, 1989: 3). Domberger and Piggott (1994: 33) mention that the conditions for allocative efficiency to be attained include marginal cost pricing, which requires output to be priced at the marginal cost of production. Oum et al. (1999) have illustrated in terms of Figure 2.1 the difference between technical efficiency and allocative efficiency. Leibenstein, however, argues that empirical evidence suggests that the problem of allocative efficiency is small. The empirical studies that Leibenstein used found that allocative inefficiency is frequently no more than one tenth of one percent. This suggests that X-inefficiency exists and that it is frequently much more than allocative inefficiency, which is associated with a monopoly.

Leibenstein (1989: 17) specifies three elements in determining what the X-efficiency is. The first is the intra-plant motivational efficiency, which relates to the efforts of managers and employees. The second is the external motivational efficiency and the third is the non-market input efficiency. According to Leibenstein, neither the individual nor the firm work hard. The individual and the firm are not searching for information as effectively as they should, "the importance of motivation and its association with the degree of effort and search arises because the relation between input and output is not a determinate one" (1989: 18). This author gives four reasons why the given input cannot be transformed into the predetermined output. The first is that the contract for labour is incomplete; secondly, not all the factors of production are marketed; thirdly, the production function is not completely specified or known; and lastly, the interdependence and uncertainty lead competing firms to cooperate tacitly with each other in some respects, and to imitate each other with respect to technique to some degree.

In summary, Leibenstein's X-efficiency theory points out how X-inefficiency may emerge in selected contexts and mentions that firstly, considerable X-inefficiency may arise as a result of low pressure (competition) for performance from the environment, and monopoly is the case in point; secondly, even under competition, the pressure may be limited in the sense that entrepreneurs capable of entering the industry are few and not competent enough to organise the firm to produce at lower costs than those already in the industry; thirdly, some firms may be sheltered either by a system of government regulation e.g. price regulation guaranteeing a fair return, or a situation in which firms operate on a cost-plus contract basis; fourthly, low pressure on firms may exist because of considerable inability on the part of buyers of the service to understand the nature of the product; fifthly, despite a reasonable degree of pressure from the environment, some firms may suffer from the organisational inefficiencies in that the transmission of signal for performance or the transmission of incentives become blurred or lost as they make their way through the hierarchy and X-inefficiency results; and sixthly, there is no reason to expect minimal cost equilibrium, even in the case of a large number of sellers, if the environment creates relatively weak motivation for entrepreneurs to enter the market.(1989: 57).

Leibenstein's selected context for X-inefficiency, can be said to be applicable mainly to monopolistic enterprises. There is also a striking similarity between the X-efficiency of Leibenstein and the technical efficiency discussed earlier in section 2.4 of this chapter. In that section, Domberger and Piggott (1994: 36) argued about technical efficiency and mention that it is synonymous with cost minimisation for the given level of output. Technical efficiency is, therefore, a necessary condition for allocative efficiency to be achieved. In the context of concessioning, the operator will strive to achieve technical efficiency as he will know that if minimum cost of production is not achieved it will be inefficient and therefore the possibility of losing the concession to a more efficient operator at the renewal stage will be increased.

2.7 Contestable market theory

The contestable market theory was first developed by Baumol. This market theory provides a generalisation of the concept of the perfect competitive market and is applicable to the various ranges of industry structures including monopoly and oligopoly (Baumol, 1982: 2). Contestable market theory provides a broader ideal benchmark than the perfect competition theory against which various applications including rail can be measured. According to Baumol, a contestable monopoly "offers us some presumption, but no guarantee, of behaviour consistent with a second best optimum, subject to the constraint that the firm be viable financially despite the presence of scale economies which render marginal cost pricing financially infeasible. That is, a contestable monopoly has some reason to adopt the Ramsey optimal price-output vector" (1982: 2). The application of marginal cost pricing in the case of a natural monopoly will result in the firm losing money, as the relevant marginal costs, depending on demand, will be less than the average cost. As a result, the first-best pricing solution, which is marginal cost pricing, cannot be relied on in the case of a natural monopoly (Train, 1991: 89). If such a firm adopts the Ramsey prices, it must set prices sufficiently above the marginal cost to enable it to break even, in other words to earn zero profit. The dependence on pricing above marginal cost can therefore only provide the second best solution (Train, 1991: 117).

The underlying assumptions to the contestable market theory were further analysed by Shires, Preston, Nash and Wardman (1994: 14), especially with regard to rail franchising in the United Kingdom. For the market to be contestable, three key assumptions have to be fulfilled. These assumptions are:

i that the entrant and the incumbent must be symmetrically placed. This assumption means that both the entrant and the incumbent must be subject to the same regulations, they must possess similar knowledge and produce output perceived as being of the same quality and at the same cost.

ii there must be an absence of entry barriers like sunk costs, so that entry and exit from the market is costless e.g. any assets accumulated can be used for the production of either goods or services and their value can be recouped in the second-hand market. According to Train (1991: 303), entry into the market must be "free" and exit "costless". Free entry into the market, however, does not mean that an entrant need not incur any cost, but rather that a new firm does not have to incur costs that are not also incurred by the incumbent firm. In other words, the entrant should not be at a cost disadvantage. In the case of the costless exit, it means that when the firm leaves its operations it must be able to recoup all the costs that it expended when it entered.

iii hit and run entry must be possible. The hit and run assumption means that either consumer reaction time to price differences must be quicker than the incumbent or it is possible for the entrant to enter into secret supply contract and negotiate with consumers to secure a period of profitable entry (Shires et al., 1994: 15). Train quotes Baumol (1982) and he mention that the crucial feature of the contestable market is the hit and run entry. The entrant can go into the market and before the incumbent can adjust the price, can collect his gains and leave the market. Train mention that, because of the threat of entry, actual entry may never occur and this threat will keep the incumbent monopolist at zero profit with efficient production (Train, 1991: 303)

Should all the assumptions be fulfilled any remaining profit opportunity will be exploited by the possible entrant. Consequently, the incumbent is constrained to set a price that is equal to the average costs. Should the incumbent charge prices that are above the average cost, it would be facing the threat of entry from a potential competitor (Shires et al., 1994: 15).

Apart from entry barriers to the market such as sunk costs, there are other entry barriers like the strategic, predatory and innocent barriers. The innocent barriers mainly include economies of experience like the management and staff knowledge of the activities of the relevant firm. The existence of innocent barriers can give rise to strategic and predatory behaviour towards entry. The strategic barriers occur mainly before entry by the new entrant into the relevant market and are usually in the form of price and service matching promises. The predatory barriers can take the form of unsustainable fare cuts by the incumbent before and during entry into the market. The success of predatory barriers will depend on the financial muscle of the incumbent and the market segment that the incumbent and the entrant are competing for (Shires et al., 1994: 17).

> The theory of contestable markets largely provides a yardstick against which the contestability of the various industries can be judged. Baumol notes, contestable markets may share at most one attribute with perfect competition. Their firm need not be small or numerous or independent in their decision making or produce homogenous products. In short, a perfect competitive market is necessarily perfectly contestable, but not vice versa (1982: 4).

Train (1991: 305) identified a limitation to the contestable market theory. The limitation is the assumption of hit-and-run entry, because such an assumption rests on the idea that an entrant can enter the market and earn profit before the incumbent firm can retaliate by reducing prices. In reality it is much easier and quicker for the incumbent to reduce prices than the potential entrant to purchase equipment and other production facilities, including the hiring of labour, and to notify customers about its entrance. In such a situation, the incumbent monopolist can maintain high prices indefinitely. Train mentions that

> When the incumbent observes that a new firm is starting to establish operations, it lowers its price before the new firm can actually offer service. After the new firm is run out of the market, the incumbent simply raises its prices again. In fact, because the potential entrant knows that the incumbent will do this, the potential entrant will not enter even though the incumbent is earning positive profit and/or producing inefficiently (1991: 303).

This behaviour is what Shires et al. (1994: 17) describe as strategic and predatory behaviour to prevent entry.

Train (1991) further points out that there are two ways in which the contestability theorist can support his argument. In the first place, the potential entrant can sign long-term contracts with the targeted customers before actually establishing the operations. In this situation, the potential entrant would lock-in customers to its service after the establishment of the operations. Train argues that the entrant would be able to bind customers to its service after the actual establishment of operations if the entrant lowered its prices more than the existing competitor. Where the entrant prices are lower than the incumbent's, the customers will be willing to sign contracts with the potential entrant even if they know that the incumbent will reduce prices when the new firm enters the market. Given this situation, therefore, customers will sign long-term contracts with the entrant because they know that if they do not sign, the existing incumbent will raise prices again after the entrant has left the market. As a result, the only hope of customers for long-term price reductions is to sign the agreement with the entrant. Train further argues that the incumbent could also attempt to sign up customers on terms equal to or better than those offered by the new entrant. If the incumbent succeeds in preventing customers from signing with the new entrant, prices will come down in the long-term contract. As the potential entrant can start to sign customers before the establishment of the operations, the entrant need not expend any cost on entry until profitable entry is a certainty (1991: 305). Although this is not applicable in a situation where on-route competition is not allowed, it nevertheless throws more light on the strategies that can be employed by the entrant and incumbent where on-route competition is permitted.

According to Train, the second way to maintain the contestability is by a mandate of the regulator. The regulator may be required to prohibit predatory behaviour by the incumbent, especially where there is entry by another firm. Despite the prohibition on predatory behaviour, the incumbent may choose a low price that prevents entry (1991: 306). In the concessioning environment, Shires et al. mention that contestability depends on the effectiveness of the regulator. Referring to the situation of "on-the-route-competition" (open access) in the UK, Shires et al. further mention that the regulator would need to differentiate predatory behaviour from genuine competition and, to accomplish such task, would need reliable and realistic information about the costs and the revenue of the incumbent firm (1994: 18). Consequently, this means that the predatory behaviour in as far as commuter rail concessioning in South Africa is concerned would be greatly reduced because the current policy does not allow open access competition (on-the-route competition).

2.8 Conclusion

This chapter studied the theory underlying the private and public sector enterprises in terms of the principal-agent approach. Assuming that commuter rail services will be concessioned to the private sector, it is essential to have a broad understanding of the objectives pursued and the problems that plague both sectors. The underlying principal-agent approach throws more light on this.

The main accepted objective of private sector enterprise is the maximisation of profit. Maximisation of profit enables shareholders to receive a return on their invested capital and the firm to sustain its operations. The discussion on private enterprise behaviour according to the principal-agent approach shows some problems with regard to the accepted objectives of this sector. The principal (shareholders) problem is the lack of sufficient information concerning the activities of the agent (management) of the enterprise. The constraints that confront management of the private sector enterprise in aligning their activities (efforts) with the objectives of shareholders are not sufficient to constrain members of management from furthering their careers at the expense of the shareholders. It is all very well to criticise and reject the accepted objective of private sector enterprise, but it has to be replaced by an acceptable alternative objective. This therefore leaves private sector with the profit maximisation objective.

It is accepted that the major objective of state-owned enterprises is the improvement of the economic welfare of the country. The principal-agent approach reveals problems experienced here as well. The major problem again is that of information and this is compounded by the large hierarchy that is used to monitor the activities of the agents (state-owned enterprises). In the absence of sufficient information about the activities of public enterprises, it is also difficult to say that they will strive for the economic welfare of the country. Without an accepted alternative objective for public sector enterprise, the promotion of the economic welfare objective therefore remains the main objective for the public sector enterprise.

This chapter also looked at the traditional arguments concerning private and public enterprises. It was argued that private enterprise provides constraints through competition to promote productive efficiency, which public enterprises fail to achieve. The said is in no exception to the publicly owned rail enterprises. Productive efficiency was discussed in terms of Figure 2.1 and reasons why state-owned enterprises may not strive for cost minimisation were given. The empirical studies undertaken to determine the economic efficiency of railway systems in OECD countries were discussed. These studies concluded mainly that the efficiency of rail can be significantly enhanced by the regulatory framework that gives autonomy to the management of the rail enterprise. Furthermore, the study that comprehensively surveyed the methodologies used in measuring and comparing productivity in the rail industry was reported on. This study concluded that increased competition through regulatory liberalisation improved rail efficiency.

Furthermore, there is large amount of literature that debates the superiority of private enterprise over public enterprise. The literature is largely theoretical and the conclusion that is made is based on the conviction of the authors. It is reported that the principal-agent approach and the traditional arguments are not essentially different in the content of their arguments but differ mainly with regard to the approach used.

This chapter also investigated the question of lack of competition in the markets, particularly those served by public enterprises. The lack of competition is derived from the monopolistic characteristics of public enterprises, which are exhibited by economies of scale and scope. In rail transport a distinction is made between economies of density and economies of size. Empirical studies regarding economies of scale in the rail industry shows that a rail network can have increasing, decreasing or constant return to scale.

There is an increasing trend to shift services traditionally provided by public utilities to private enterprises and the provision of rail services is no exception. This trend is seen as "public failure" to be a low cost producer in the relevant services. The shifting of services that are traditionally provided by government-owned enterprises to the private sector is accomplished by various mechanisms, such as concessioning in the case of rail services. It should, however, not be interpreted as the only mechanism as there are other ways that can be used especially with regard to rail.

The X-efficiency theory asserts that this type of efficiency is large when it is achieved relative to the allocative efficiency. However, should X-efficiency not be achieved, the resultant X-inefficiency is equally large relative to the allocative inefficiency. Major areas where X-inefficiency could be experienced include monopolistic enterprises. X-efficiency theory needs to be understood in terms of the technical efficiency that is part of the productive efficiency discussed in this chapter.

The theory of the contestable market provides a yardstick against which a monopolistic industry can be made to be more contestable. There are, however, assumptions that need to be realised in practice when restructuring a monopolistic enterprise to make the market in which such an enterprise operates more contestable.