Regulating Rail for Growth: Alternative regulatory paradigms

Nick Wills-Johnson, Fred Affleck
Planning and Transport Research Centre, Curtin University, Perth, Australia

1. Introduction

Infrastructure and the effects of regulation on its provision has been capturing the headlines around Australia recently. A Prime Ministerial Taskforce has delivered its findings on the debate (EIT, 2005) and Federal Parliament is examining a bill to alter Part IIIA of the Trade Practices Act to make it more conducive to investment. This paper charts some of the ways in which investment in the rail industry can be influenced by economic regulation and examines some of the consequences of measures intended to enhance investment.

There are, broadly speaking, two ways in which economic regulation of rail infrastructure can influence growth and investment; through regulations which determine industry structure, and through pricing. This paper examines both, and forms the first in a series of papers in a research project examining the impacts of National Competition Policy on the provision of rail infrastructure in Australia. Section One provides a brief overview of the underpinnings of the economic regulation of prices and structure. Section Two outlines the ways prices are regulated in various jurisdictions and some of the structural reforms implemented in Australia and overseas. Section Three examines how some jurisdictions pursue alternate policies to enhance railway investment. Section Four concludes.

2. The role of regulation

Economic regulation is aimed at ensuring that, absent of the competitive forces which characterise other industries, natural monopolists do not adversely impact the allocative efficiency of resource allocation in the wider economy, through the appropriation of rents via their control of certain bottleneck infrastructure assets.

Rail infrastructure is considered a natural monopoly because of its high sunk costs which means that it is generally not economically feasible for a new entrant to the rail market to duplicate existing infrastructure. In contrast to other natural monopolies however, the amount of pricing power in rail is often slight. Most obviously, rail pricing is constrained by other modes such as trucks, shipping and pipelines which are its substitutes. This means that may be more appropriate to class only certain assets in a system as natural monopoly assets rather than the whole system of infrastructure.

2.1 Prices

In the regulation of prices of natural monopolies, a regulator aims to emulate the perfectly competitive situation, where prices are set equal to marginal costs, and allocative efficiency is guaranteed. In principle, this is not difficult; all the regulator need do is ascertain the cost function of the firm and take its derivative. There are, however, two key issues which render this task, simple in principle, much more difficult in practice.

The first is the presence of large sunk costs in a natural monopoly industry. Whilst from a static perspective, sunk costs are not part of efficient marginal cost pricing, if a firm believes that sunk costs incurred now cannot be recovered in future prices due to regulation, it will not make such investments. In a more competitive industry, with smaller sunk costs, price levels
may fall below those necessary to recover sunk costs from time to time. This, however, is very different from a regulator requiring at all times that prices remain at these low levels, as in the competitive industry, prices recover to render ongoing investment sustainable.

The second issue is that the regulator can never have full knowledge of a regulated firm’s cost structure, which establishes an informational asymmetry between the regulator and the regulated firm. Laffont & Tirole (1993) formalise a modelling framework which highlights the dilemma and shows that a regulator cannot force firms with different cost structures to price at their marginal cost where an informational asymmetry exists. Myerson (1979) recognises that, whilst the regulator cannot know the cost structure perfectly, it might be able to ascertain a distribution of price and output pairs, which it can then present to the regulated firm(s) as a menu of prices. This is known as ‘incentive compatible’ regulation. Smith & Tsur (1997) present an example in rural water pricing and Kenyon & Wills–Johnson (2003) summarise some of the subsequent literature on informational asymmetries in a dynamic setting where the regulator is able to observe past performance, and formalise the notion of a dual informational asymmetry; the regulated firm never knows with certainty the price cap a regulator will impose, which may limit ambit claims in certain circumstances.

Thus, whilst price regulation is necessary to prevent abuse of monopoly power, a trade-off also exists between the exactitude with which the regulator can model the cost structure of a regulated firm and the degree to which regulation can become inimical to business efficiency.

2.2 Structure

In general, prices are regulated on an annual basis, and closely monitored by regulatory agencies. Structural regulation, however, is a once-off policy decision and is the outcome of a fundamental trade-off which exists in naturally monopolistic industries. Where the natural monopoly components of an industry are vertically integrated with competitive components (for example, the ‘below rail’ track and signalling component of rail and the ‘above rail’ rolling stock and train operations component) there are often efficiency gains. Pittman (2005) suggests that rail may be unique in the size of these gains. However, the gains accrue to the monopolistic rail provider, not its customers. More importantly, from the perspective of economic efficiency, a monopolist produces less, in general, than a competitive firm. The relevant trade-off is whether the output gains produced by the greater efficiencies of vertical integration are greater than the output losses associated with monopolistic production.

2.3 Regulating rail prices and structure: Common practices

Despite many differences in approach around the world, pricing regimes have one common thread: all try to set prices which in some way reflect the costs of the regulated railroad concerned. The intent is generally to be ‘neutral’ in effect: to not provide any obvious advantage or disadvantage to the regulated rail firm. However, in application, the effects of these attempts at ‘neutrality’ can be substantial and potentially difficult to detect.

2.3.1 Price Regulation

In general prices are not set, but rather capped at a level determined by a regulator to be the maximum that a hypothetical ‘optimised’ competitor would charge. There is some debate as to the merits of this approach to price capping compared to more traditional North American rate of return regulation (see Burns, Turvey & Weyman-Jones, 1998), but in its favour, price capping, applied appropriately, it is less susceptible to the Averch-Johnson Effect (1962) of rate of return regulation. However, price capping, as generally practised is far removed from
the CPI-X mechanism proposed by Littlechild (1983).\(^1\) It is an inward-focussed and complex process, incapable of falsification and with little theoretical backing (particularly in regards to asset values), which ensures that regulatory determinations can continue for months or even years. For these reasons, the Productivity Commission (2001) recommended substantial simplification of the regime and limitations to its scope, a call echoed by EIT (2005) and subsequently incorporated into recent changes to Australia’s *Trade Practices Act*.

Price caps exhibit two asymmetries in respect to their impact on investment. Firstly, if prices are capped too low, insufficient revenues are earned and investment will be curtailed. However, if price caps are generous, there is no guarantee that additional revenues will be put into greater investment rather than being returned to shareholders. Secondly, a price cap may restrict the type of investments made, with riskier assets being eschewed as the price cap may preclude returns sufficient to be commensurate with their risk. In particular, greenfield investment, where demand is uncertain, may be adversely affected by a regulatory price cap. It is partly for this reason that the ACCC employs different principles when assessing greenfield infrastructure projects (ACCC, 2002) and was a key reason for investors in the Tarcoola-Darwin Railway insisting on an effective regulatory holiday from the terms of the *Trade Practices Act* for a long period before they would commit funds.

Australia’s economic regulation systems for rail are complex.\(^2\) All jurisdictions aim to be ‘light-handed’ in the application of economic regulation, employing revenue floors and ceilings (based on some measure of the incremental and full costs of each line segment, slightly different in each State) and allowing a ‘negotiate-arbitrate’ model to operate between access seekers and access providers in the typically wide band of potential prices which exists between the revenue floor and ceiling. Some jurisdictions, such as WA calculate floors and ceilings on an ex-ante basis, whilst others, such as NSW, Victoria (currently under review) and SA operate on an ex-post basis, with substantial regulatory involvement only in cases of dispute. Queensland and the ARTC regime also provide indicative reference tariffs to assist negotiations. Revenue caps provide greater flexibility with respect to pricing, but they also influence the amount of revenue a firm may earn on each line segment. In practice, however, very few lines approach their revenue caps, as such caps are based upon an optimised model of throughput and very few Australian rail lines operate at optimal capacity. BTRE (2003) provides a detailed overview of the pricing regimes around Australia.

**Short Run vs Long Run Marginal Cost**

Economic efficiency requires pricing at marginal cost. However, in industries with high fixed and sunk costs, if short run marginal cost pricing is employed, pricing is highly volatile: low whilst spare capacity exists and spiking as capacity is reached. In some cases, such as electricity generation, this provides a strong signal for capacity investment, but in rail, with less peaky demand, it may provide poor incentives, particularly given that no rail jurisdiction has effectively managed to auction rail capacity slots to effectively allocate demand for peak period capacity. In Victoria, a key criticism of the existing regulatory regime (currently under review) is that, by refusing to allow the incumbent to recover its sunk costs, the regulator is requiring pricing at short run marginal cost and restricting incentives for new investment. In some cases, however, short-run marginal cost pricing may be appropriate. For example, in New South Wales, Pacific National’s (PN’s) contract to operate the grain line network expires in 2007 and it is not clear what will occur after that date (EIT, 2005). In setting regulatory prices, asset values have been set to zero to reflect this uncertain future.

---


There are three extensions to short run marginal cost pricing. The first is to include some portion of short run fixed costs in prices and hence price above short run marginal cost. The following section outlines three commonly used methodologies in this regard. The second is to price at short run marginal cost and then cover fixed costs with transfers from government. This preserves efficient prices, but at a high cost in government subsidies. The third approach is that used by most regulators, to price at long run rather than short run marginal cost. This allows future efficient fixed costs to be recovered and smoothes (although not completely) the spikes which occur in short run marginal cost pricing regimes. However, an implicit assumption of long run marginal cost pricing is that the regulated firm can in fact expand output to the optimal level, but there is no guarantee that it will be able to do so, as demand for rail is a derived demand which depends on other sectors of the economy. This is exacerbated in the case of a new rail link, where the initial customer can expand demand only to a limited portion of the total output required to be at the optimal price, and future customers are uncertain. As such, whilst long run marginal cost caps in principal are better for investment than short run marginal cost caps, a failure to satisfy the implicit assumption that output can be expanded could render them injurious to investment promotion.

Cost allocation mechanisms, non-linear pricing and Ramsey Pricing
Most jurisdictions have some mechanism for recovering fixed and common costs, which are large in the rail industry. There are essentially three ways in which this can occur, which are not necessarily mutually exclusive:
- Fixed costs can be shared, according to some administrative rule.
- Ramsey pricing can be used.
- Non-linear pricing can be used.

Each of these methods of incorporating fixed costs into prices has issues for the efficiency of pricing and hence for investment, particularly insofar as they influence demand and hence the ability to earn returns to satisfy it into the future. In Australia, fixed costs are allocated according to administrative rules by regulators when used in the setting of price/revenue floors and ceilings. They are not used by rail firms (necessarily) to price actual services, which are determined by negotiation. For this reason, the criticisms below may be muted in the Australian context, particularly where the relevant cap is on revenue, not price.

As noted by Breautigam (1980) and Kessides & Willig (1995,1998), all administrative methods of allocating common costs raise efficiency concerns. Breautigam formalises the reasons why such inefficiencies occur, and makes the following key criticisms of administrative rules for cost allocation:
- There is no relationship to marginal costs and hence to economic efficiency.
- There is no uniquely acceptable allocation rule and all have an arbitrary element.
- It is not necessarily optimal to recover all of the costs attributable to an item of output from that item of output, as allocative methods require.
- The determination of what constitutes fully distributed costs is an arbitrary administrative decision and hence there is no economic basis for concluding a service is not meeting its costs if revenues from that service are less than its fully distributed costs.
- Allocative rules are anticompetitive as firms cannot offer tariff lower than fully distributed costs, even if such tariffs would be above marginal cost.
- There is a path dependency whereby new tariffs are based on revenue or output shares from previous periods, which are themselves demand responses to an original set of administratively allocated cost prices.

When efficiency is negatively impacted by cost allocation mechanisms, it can have important flow-on effects to investment. Perhaps most importantly, if the rail infrastructure provider is unable to optimise existing infrastructure, it may be required to build new infrastructure, to accommodate the inefficiency. Paradoxically, in Australia, where both regulators and
commentators alike cite the construction of new infrastructure as evidence that the regulatory system is working appropriately, inefficiencies such as this may not be perceived as such. This highlights a key problem for regulation; it is difficult to observe the counterfactual and thus determine whether investment being produced is allocatively efficient.

Economists prefer Ramsey pricing, which rations fixed costs according to demand elasticities and is thus an optimal second-best solution. However, from an implementation perspective, there are two potential flaws, particularly in its use by regulators. Firstly, customers dislike Ramsey pricing which discriminates against them and may seek to use the regulatory process to overturn it. In Australia, since the repeal of Section 49 of the Trade Practices Act (which contained a per-se prohibition of price discrimination) there has been no legal restriction to Ramsey pricing at a Federal level. Section 46 of the Act, which contains the relevant pricing prohibitions, requires proof that the purpose of the pricing regime (not its effect) was to damage competition, which is a difficult hurdle in prosecution. The various state-based regimes have stricter requirements that prices reflect underlying costs, but without explicitly prohibiting Ramsey pricing. Whilst there may be no legal issue with Ramsey Pricing, ‘good’ price discrimination can be difficult to distinguish from price gouging, and the regulated firm may be unwilling to take the risk that the regulator will form an unfavourable impression. In Australia, ARTC has undertaken not to Ramsey price, whilst RIC has Ramsey prices for some of its services, which suggests that each (regulated by different regulators and different Acts) may have a different impression of how complaints about its pricing might be treated, or different impressions about how likely its customers are to complain. In the case of integrated service providers, Ramsey pricing might be used as a means of covering cross subsidisation between the integrated service provider’s downstream train operators and its competitors, which may make a vertically integrated access provider more wary of being seen to Ramsey price.

Secondly, Ramsey pricing is represents a huge informational cost for the regulator. Transport is a derived demand, and hence the regulator would need to calculate (and update) hundreds of demand elasticities for the different products carried. Thus, whilst Ramsey pricing might be the best way to price, it may be the worst way to regulate prices.

A final means of covering fixed costs is to charge an up-front access fee, as well as a variable charge for use. In Australia, ARTC prices its rail services in this non-linear manner, with the flagfall charge equalling some 20-40 percent of total charges. Whilst non-linear prices can ensure that fixed costs are recovered, they can also have effects on demand. They may, for example, represent a barrier to entry for smaller shippers, excluding them from the market and thus decreasing demand. Also, they may change shipper behaviour; if the fee is on a per-train basis, there is an incentive to run fewer, longer trains whilst, if the fee is an annual charge, there is an incentive to maximise track use. If fewer, longer trains are used, passing loops may need to be extended, along with terminals, platforms and shunting yards. If more, smaller trains are used, network congestion may increase, meaning network expansion must be brought forward sooner than would otherwise be the case. Thus, as with administrative cost allocation mechanisms, one may see an increase in investment, but it is unclear whether that investment is in fact warranted by the underlying economics of the industry, or whether it is simply a by-product of a component of regulation.

In summation
If prices are correctly set at long run marginal cost and no issues exist in relation to expansion of output to meet the optimal point on the long run marginal cost curve, then there is no reason to expect caps to unduly influence investment. The ARA (2005) suggests that only coal lines price to meet their caps, whilst lines used mainly for inter-modal freight barely

\(^3\) Only one rail industry case, Freightcorp vs Austrac has been brought to court, and no breach of Section 46 was found.
reach halfway to their caps, and grain lines recover even less revenue. This also suggests that caps may not be limiting investment substantially in Australia. However, departures from marginal cost pricing can influence demand behaviour, rendering it difficult to determine whether investment subsequent to the introduction of regulation is an efficient response to the economics of the industry, or merely an artefact of the regulatory regime.

2.3.2 Regulation of Structure

Structural regulation refers to the way in which the rail industry is restructured into the above and below rail components. There is a continuum of alternate structures possible, with fully vertically integrated operations at one end, and above rail and below rail owned by independent companies at the other end. Indeed, a one-dimensional continuum can become two-dimensional by adding ownership to the mix, and considering whether the above and/or below rail components should be held in public or private hands. This added complexity is beyond the scope of this paper, except to note that it is not ownership per se, but rather the incentives provided and the ability to contract to overcome principal-agent issues which are the key factors. Bartel & Harisson (2005) provide a good general review of the literature surrounding public versus private sector management of business.

A 'perfect' model of structural separation, which will provide for benefits in any jurisdiction in which it is employed does not exist. However, it is useful to consider different regimes and the historical context in which they have developed. The discussion below does so, proceeding from vertical integration through to full vertical separation.

The 'vertical integration' end of the spectrum is typified by Canada and the US, where history that has been a key factor market structure. In both of these countries, land grants and generous subsidies (discussed further below) during the Nineteenth century resulted in competing trans-continental lines. Thus, in the US and Canada, one has the potential for competition in the market between different rail companies serving the same origin and destination pairs. As such, there is little reason to attempt to engender competition by a wholesale vertical separation as has occurred in Europe. Also, the private ownership of both track and rolling stock means that any attempt at vertical separation would be much more expensive and difficult than it has been in Europe, where such firms are often state-owned.

However, even in the US and Canada, it is not possible to rely on competition in the market alone, as bottleneck infrastructure assets still exist. In the US, the Surface Transport Board has the power to mandate access to track. Canada relies on 'inter-switching', whereby the regulator can mandate the right of shippers to switch railway providers provided freight originates within 30 kilometres of the point at which two competing lines meet, and to mandate a price at which this occurs (CTARP, 2001). Whilst the approaches to enhance competition used in the US and Canada may be replicable elsewhere (and, indeed, preferable to regulating a whole rail network, rather than targeted assets within it), competing trans-continental lines would be prohibitively expensive to replicate elsewhere. As such, competition in the market has only limited applicability outside North America.

Western Australia and Queensland represent two cases whereby above and below rail operations form distinct and legally ring-fenced parts of the same organisation. In

---

4 CTARP (2001) estimate that almost two-fifths of Canadian rail traffic has the choice of at least two rail links and cites research by the Brookings Institute suggesting that only a fifth of US shippers are 'captive', in the sense that they are served by only one railroad and have no competing freight modes which can serve them in a cost competitive way.

5 Kwoka & White (1999) provide a summary of criticisms of mandated trackage rights, in an assessment of the Union Pacific and Southern Pacific merger. Canada also has mandated trackage rights, but these are rarely used (CTARP, 2001).
Queensland, this is government owned (Queensland Rail, or QR), whilst in WA, the lines have been transferred (via a 50 year) lease to a private sector organisation, the Australian Railroad Group (ARG). Ring fencing represents a slight departure along the continuum from full vertical integration. WA and Queensland are hardly unique, but the reasons why ring fencing might be successful in these cases are illuminating. The freight task in both states is dominated by haulage of minerals; primarily coal in the case of Queensland and bauxite, coal, grain, alumina, iron ore, nickel ore and woodchips in the case of WA. This minerals freight task has only a small number of shippers, each with large contracts. Entry to the freight market has the potential to occur only rarely, when haulage tenders are renewed and is unlikely to be sufficient to warrant costly full divestment between above and below rail operations. Essentially, ring fencing (for all the problems associated with ensuring business units in the same firm do not communicate) represents a form of contestability to the industry which may suit the nature of the industry in Western Australia and Queensland better than full separation. Moreover, there are some indications that it is working; several of the contracts for shipping minerals in WA have been renewed and, although the incumbent has won each time, industry sources suggest that prices have decreased substantially through competitive tendering and there is no indication that the incumbent can necessarily expect to win future contract renewals. In Queensland, Xstra, wishing to have a new spur line built for its coal operations recently, put the contract out to tender. QR won, but at a price much reduced from that which would have occurred before such contestability was possible.

Japan has a different type of competitive environment, and thus pursued a different type of reform, involving not vertical, but horizontal separation (see OECD, 1998 for details). Although a digression, it is worthy of some comment. Japan, like most OECD nations engaged in rail reform, separated passenger from freight rail. However, unlike other nations, where passenger rail was separated because it is unprofitable without government subsidy, in Japan, it has become a viable business. Freight rail is less important in Japan owing to its mountainous terrain and coastal cities, which mean that shipping is a competitor to rail for freight in a way which it is not elsewhere. Japan split its passenger business into six geographic monopolies which allows for benchmarking competition.

Mexico undertook separation at the next stage along the continuum from ring fencing. Like Japan, it established its rail concession as geographic, vertically integrated monopolies subject to benchmark competition. Unlike Japan, these concessions were franchised to private sector firms, providing competition for the market. Also unlike Japan, certain key elements of infrastructure used by all (chiefly the main terminal in Mexico City) were held in common by all and with part government ownership as well. This is because the nature of competition in Mexico is different again. It does not have many instances where multiple railways service both ends of a transport task, as occurs in the US and Canada, so it cannot support competition in the market. By the same token, it is a relatively large and diverse market, without the rare entry which characterises WA and Queensland. What it does have is origin or destination competition (see Winston & Grimm, 2000). In this case, by careful choice of the nature of the geographic concessions which are provided, government can ensure that the different rail links servicing the key industrial areas are owned by different companies. Of course, it is not possible to provide a perfect match for all rail links and all industrial areas, without sacrificing certain network efficiencies within the rail industry itself. This is the reason behind communal ownership of certain key bottleneck assets. This gives the system a great deal of flexibility, without incurring the sometimes burdensome costs associated with regulating an entire rail link.

---

6 Belgium, Greece, Ireland, Spain, Switzerland and Luxembourg have also created infrastructure and operations as different divisions of the same company and Germany and Italy have separated them into different companies, but kept common holding companies.

7 Melbourne, the UK and Sweden franchise passenger trains and Argentina franchised freight.
The Australian ‘Defined Interstate Rail Network’ (DIRN), and European networks present a different type of competition. Much of the traffic on these lines is inter-modal, an industry where barriers to entry are relatively low compared to those in the minerals industry. Also, traffic levels are high, and there is limited origin and/or destination competition.\(^8\) In Australia, ARTC owns (in some cases, leases) the track infrastructure and has no role in shipping. Network Rail in the UK is the same. Both represent cases of full vertical separation.

Thus, it is not the case that vertical integration is necessarily better or worse than vertical separation. Rather, the historical characteristics of the rail infrastructure’s development, the nature of the freight task, the nature of bottlenecks in the network and the nature of potential competition on the network all contribute towards the decision as to what represents the ‘best’ form of integration. However, vertical separation seems, in principal, to be limited in its applicability. Given that it is costly to implement, it is useful to explore its historical record at providing an incentive for competition, and some of the dynamics which throw light on why such competition might prove beneficial.

**The record of vertical separation**

The historical record is still relatively short, but is thus far not particularly bright. In Australia, immediately following the introduction of open access in the mid Nineties, SCT, Toll and Patrick Rail entered, with Toll and SCT taking up to a quarter of the market on the East West freight rail link. Toll was subsequently absorbed by Pacific National, and SCT now has a market share on the East West link of less than five percent. Other new entrants include:\(^9\)

- South Spur Rail, in WA, which has a small fleet of locomotives and undertakes hook and pull services for other rail companies.
- CRT, in Victoria and NSW which owns wagons, some terminals and has brought Sprinter trains into Australia.\(^10\)
- Lachlan Valley, which owns wagons and moves container freight in NSW.
- Chicago Freight Car Leasing, which leases wagons and locomotives in Australia, filling a key role for new market entrants.

The market share of these new entrants is negligible. Most of the competition in Australia comes from the five rail companies formed during the reform process or given concessions as part of it (ARTC, PN, ARG, QR and RIC)\(^11\) some of whom are now actively competing in each other’s markets, as well as the two companies formed with the completion of the Tarcoola-Darwin railway (Asia-Pacific Transport and Freightlink). In the future, it is likely that there will be four major above rail providers (QR, PN, Freightlink and ARG) operating in all major markets in Australia, supported by their own infrastructure (in the case of ARG and QR) and using the infrastructure managed by ARTC, Asia-Pacific Transport, and, to a much lesser extent, RIC. Each city maintains a monopoly over commuter transport, except Melbourne, which franchised its passenger rail task to two private-sector firms, Connex and the National Express Group in 1999, aiming for some competition for the market. However, National Express withdrew from its concession in December 2002 as it was not profitable and its franchise was taken over by Connex.

From a legislative perspective, the EU has gone furthest in vertical separation, establishing a framework which both mandates separation of above and below rail and provides openings

---

\(^8\) Europe, despite the efforts of the EU to build international freight routes, still largely follows national borders with respect to rail.

\(^9\) Asia-Pacific Transport, which built the Tarcoola-Darwin railway and manages it under lease from the governments of South Australia and the Northern Territory and Freightlink, the company formed to manage freight on the line, are not included, as they entered the market as a result of a particular government tender process, not due to the access regime.

\(^10\) CRT was recently bought out by QR.

\(^11\) Most of the track owned by RIC is now leased to or managed by ARTC
for competitors to establish themselves in each market. Moreover, Europe has much larger freight markets than Australia. Here too, however, progress has been disappointing, apart from incumbents in each country obtaining access to neighbouring markets on a reciprocal basis. Table One summarises new market entry in Europe.

### Table 1 New Entrants in European Rail Markets (2004)

<table>
<thead>
<tr>
<th>Country</th>
<th>Number of New Entrants</th>
<th>Share of market</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>10 (only one operating on Belgian network)</td>
<td>negligible</td>
</tr>
<tr>
<td>Belgium</td>
<td>4 (two of which are the transport arm of coal mines)</td>
<td>Not provided, but market regarded as effectively closed.</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>7 (plus small network concessions for regional passenger networks)</td>
<td>Two percent of freight market in tkm.</td>
</tr>
<tr>
<td>Denmark</td>
<td>6 (two passenger and two freight)</td>
<td>Less than five percent, for both freight and passenger transport.</td>
</tr>
<tr>
<td>Estonia</td>
<td>None (state railway has legal monopoly)</td>
<td>zero</td>
</tr>
<tr>
<td>Finland</td>
<td>None (two small operators scheduled to start in 2004 for local services)</td>
<td>Approximately 10 percent in freight and passenger, 100 percent (incumbent dissolved in rail reforms)</td>
</tr>
<tr>
<td>Germany</td>
<td>160 (including small network concessions for regional passenger networks)</td>
<td>Zero</td>
</tr>
<tr>
<td>Great Britain</td>
<td>28 (25 from concessions, 3 new) and four freight</td>
<td>Three percent in freight.</td>
</tr>
<tr>
<td>Greece</td>
<td>None (two small operators scheduled to start in 2004 for local services)</td>
<td>Three percent in freight.</td>
</tr>
<tr>
<td>Hungary</td>
<td>None (two incumbents share a duopoly)</td>
<td>Zero</td>
</tr>
<tr>
<td>Ireland</td>
<td>None (two incumbents share a duopoly)</td>
<td>Zero</td>
</tr>
<tr>
<td>Italy</td>
<td>33 (12 active on incumbent's network, remainder active before reform on regional networks)</td>
<td>Two percent in freight and less than one percent in passenger markets</td>
</tr>
<tr>
<td>Latvia</td>
<td>Two new entrants in freight, none in passenger rail</td>
<td>Two percent in freight market.</td>
</tr>
<tr>
<td>Lithuania</td>
<td>Four entrants hold licences but none are active</td>
<td>Zero</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>None (two small operators scheduled to start in 2004 for local services)</td>
<td>Zero</td>
</tr>
<tr>
<td>Netherlands</td>
<td>15 (six in freight)</td>
<td>n/a</td>
</tr>
<tr>
<td>Norway</td>
<td>None (two small operators scheduled to start in 2004 for local services)</td>
<td>Zero</td>
</tr>
<tr>
<td>Poland</td>
<td>21 in freight, zero in passenger</td>
<td>Three percent in freight.</td>
</tr>
<tr>
<td>Portugal</td>
<td>None (two small operators scheduled to start in 2004 for local services)</td>
<td>Zero</td>
</tr>
<tr>
<td>Slovakia</td>
<td>18 (only three active)</td>
<td>n/a</td>
</tr>
<tr>
<td>Slovenia</td>
<td>None (but 90 percent of the passenger and freight market is international thru-traffic)</td>
<td>Zero</td>
</tr>
<tr>
<td>Spain</td>
<td>Two state owned companies share the market as a duopoly.</td>
<td>Almost half of the freight market and a third of the passenger market (including regional franchises) in 2002. The passenger market is still closed.</td>
</tr>
<tr>
<td>Sweden</td>
<td>16 (ten active in freight, seven in passenger rail)</td>
<td>BLS Cargo has a 12 percent share of the freight market. The passenger market is still closed.</td>
</tr>
<tr>
<td>Switzerland</td>
<td>10, but only one (BLS Cargo) is a result of liberalisation.</td>
<td>BLS Cargo has a 12 percent share of the freight market. The passenger market is still closed.</td>
</tr>
</tbody>
</table>

Source: IBM Consulting Services, 2004. Note that figures for new entrants do not include reciprocal cabotage rights between incumbents in different countries, but rather to the number of new entrants which are domiciled in each country as a rail transport provider. These may have as their parent companies, incumbent operators from other countries.

---

**When is vertical separation likely to be beneficial?**

There are essentially three issues associated with whether competition in a separated above-rail industry is likely to be beneficial or not:

- Whether the relevant above-rail sector is in fact super-additive.
- Whether the technical efficiency losses of integration outweigh the gains in consumer surplus from a reduction in monopoly power being projected into above rail from the below rail natural monopoly.
- Whether the rents possible from above and below rail integration are substantially different from those possible with separation.

These three issues can be tested, and, indeed, will be in later parts of the project of which this paper is the first component. Bitzhan (2000) examines the subadditivity of above rail operations for Class One railways in the US from 1983-1997, utilising a methodology adapted from similar studies in telephony, and finds that 95 percent of simulations show lower costs for a monopoly than for two firms. Moreover, he finds that subadditivity conditions are met for 60 percent of observations in 1997, and that super-additivity is met for none. He estimates cost increases of between 3.8 and 15.5 percent if above rail operations...
are divested from below rail, but notes that these figures should be treated with some caution, as quasi-cost, rather than true cost functions were used. No similar studies have been undertaken for Australia as yet, but will form a future part of this project. However, if Australia, like the US, has a high incidence of above-rail sub-additivity, there seems limited scope for beneficial competition to arise above rail.

The technical efficiency gains from merging above and below rail are simple to list, but can be difficult to quantify. They include:

- The trade-off between rail grinding and wheel and bearing maintenance.
- Train scheduling and en-route train management.
- Maintenance scheduling.
- Train size and passing loop investment decisions.
- Coordinated planning of capacity expansion, with rail service providers responding to customer demand and infrastructure providers undertaking capital expansion.

Biggar (2004) cites US research suggesting efficiency losses of between 20 and 40 percent with separation of rail freight from infrastructure. He also cites the UK experience in apportioning penalty payments according to an assessment of whether the relevant fault was above or below the rail interface, which resulted in a total penalty payment of between £150 and £200 million in the three years to 2002/03 and employed some 300 people. A recent UK government White Paper (UKDFT, 2004) suggests that, whilst the franchising had improved understanding and transparency of rail costs, these costs had increased due to “complex commercial and bureaucratic relationships, the lack of clarity for responsibilities and the misaligned incentives between each part of the industry” (Ibid, p21). The White Paper suggested wide-ranging reforms in industry organisation, but stopped short of reintegration, despite the wishes of some stakeholders that this be considered. A comparison between the technical efficiency losses and the gains in output from vertical separation would be a useful means of assessing the beneficial nature of above-rail competition. To our knowledge, this has not occurred in Australia and was certainly not part of the assessment underpinning the reform policy (Travers-Morgan, 1995).

The sharing of rents can be examined from the perspective of game theory. If above and below rail are integrated, the two parties are essentially playing a co-operative game whilst, if they are separated, they are playing a non-co-operative game. In both cases, in game theory, the solution sets can be very large. However, the interest is not in the sets per-se, but rather in the differences between the two, which may be much smaller, and which represent a ‘coordination loss’ which is separate and distinct from the technical efficiency losses described above. We are not aware of any application of game theory such as this to the rail industry, but mean to make such an application as part of future research in this project. However, if the two sets are similarly sized, this provides support for the beneficial nature of above rail competition with vertical separation.

In Summation: Structural Separation and Investment
Structural separation may have substantial impacts on investment in rail infrastructure, depending upon the nature of the network and market in which it operates. This is because structural separation changes the way in which the ‘game’ is played between above and below rail in terms of investment incentives. If the above rail market is super-additive (meaning there is scope for additional competitors to arise), and the below rail company is separated from the above rail company, then one might reasonably expect more investment to occur with separation. This is because, with the potential for competition above rail and the increase in traffic which would eventuate as the above rail market expands, the below rail provider has a strong incentive to endeavour to expand capacity to meet new demand. On the other hand, if the above rail network is super-additive and the rail provider remains vertically integrated, one might expect less investment to occur. This is partially because a monopolist will produce less output than a competitive market, but may also be because the
level of output in a competitive above rail sector is in fact sufficiently large that a single organisation attempting to service it would suffer diseconomies of scale.

If the above rail market is subadditive, and competition unlikely to eventuate, then splitting the above and below rail components will result in no increase in output because above rail will remain a monopolist. Moreover, any increase in investment will be wasteful, eventuating because the co-ordination previously possible between above and below rail has now evaporated and hence the below rail (and potentially the above rail too) may need to undertake investment to account for changed incentives under the split regime. It may be hard to perceive wasted investment as such if policymakers base assessment on gross investment figures rather than their composition. By the same token, if vertical integration is retained and the above rail market is subadditive, one would likewise see no increase in investment (absent of an exogenous demand shock), as there would essentially be no change from the previous regime. However, it would actually be appropriate for there to be no investment in this case, in preference to the wasted investment referred to above.

Thus, whether structural reform will result in positive or negative impacts on new investment, and whether that new investment is actually useful or not, depends rather crucially on the nature of the above rail market. The mere presence or absence of an increase in investment is not a signal that the relevant reforms have been effective, which can only be assessed in light of the sub or super-additivity of the above-rail industry concerned.

3. Regulating rail for growth: Alternative approaches

Pricing at some measure close to marginal cost and engaging in structural reform are policies intended to be essentially neutral in respect to rail investment. However, there may be instances where an efficient market, due to market failure, cannot provide the socially optimal amount of investment and neutrality is thus not appropriate. In particular, whilst road is usually funded via public funds and assessed on an economic or cost benefit basis, rail (in Australia and North America at least) is usually funded via private investment funds and assessed on the basis of commercial returns. Compared with road, rail has fewer externalities such as pollution, accidents, noise and visual amenity. However, private rail owners are unable to capitalise on these externalities and so, potentially, an efficient market will provide less rail and more road than is socially optimal. There are three approaches government might take to redress this balance:

- Price and fund road and rail to include social costs and benefits and render the two modes competitively neutral (the example of Sweden is discussed).
- Provide targeted government subsidies to cases where rail is the socially efficient choice but through market failure it is not chosen (an example in Western Australia is provided).
- Provide additional revenue streams to railway owners which provide an incentive for more rail to be built (the example of 19th Century land grant railways is discussed).

3.1 Pricing road like rail: The case of Sweden

Sweden, over recent decades, has made substantial progress with a methodology incorporating ‘competitive neutrality’ between road and rail (see Nilsson, 1995 and Carlson, 2004 for more detail). There are two sides to competitive neutrality. On the demand side, users of road and rail should face the full social costs of their modal choice at the margin to provide appropriate demand signals for modal choice. On the supply side, funding for infrastructure should take into account the social costs of each mode to ensure efficient resource allocation between them. Sweden has arguably been more successful with the latter than the former.
The push for competitive neutrality in road and rail investment in Sweden had its origins in the *Transport Policy Act* of 1979, which also devolved control of regional rail networks to regional authorities, paving the way for future competitive tendering for these services. From 1935, rail in Sweden had been tightly controlled by government and the national railway Statens Järnvagar (SJ) began to gain autonomy for its pricing (particularly for freight) through reforms in the Sixties, but line closures remained a parliamentary prerogative and was a key cause for the increasing subsidies necessary to keep SJ operating. The 1979 Act was one of three (the others being the similarly named Acts of 1963 and 1988) which took progressive steps towards the deregulation of the Swedish rail industry, culminating in the vertical separation of SJ’s operations from its infrastructure, which was vested in a new state-owned enterprise called Banverket (BV) in January 1989 in a process which provided the blueprint for the European Union rail reform directives of 1991. The process was driven by concerns for the financial viability of SJ. Competitive neutrality of investment funding was first suggested in government planning reports leading to the *Railway Act* of 1985, as a way to realise the lower social and environmental costs of rail compared with road for some transport tasks.

Under the supply side of the Swedish scheme, road authorities and BV provide investment plans which show the net present value of the social returns to each investment. These returns include both monetary returns to the owners of the infrastructure (the Swedish rail infrastructure remains in state hands) and also externalities. Investment funds are allocated according to the lowest social cost in each transport corridor.

Under the demand side of the system, charges for use of the infrastructure, road or rail (also ports and airports) are also based on marginal social costs. Initially, rail access charges were two-part tariffs, incorporating a fixed annual charge per locomotive and wagon and a variable charge comprising maintenance, accident costs, marshalling yard fees and emissions. In 1998, under the *Transport Policy for Sustainable Development Act*, the fixed component was dropped, and the access charges comprise only the marginal social cost-based variable component, which is (Carlson, 2004):

- Track wear and tear: SEK 0.0028 per gtk.
- Marshalling yard fees: SEK 4 per wagon.
- Accidents: SEK 0.55 per km per freight & 1.10 per km per passenger train.
- Emissions: SEK 0.31 per litre of fuel consumed.

For road users, access charges are also based on marginal social costs, but these are not levied on a mass-distance basis as for rail users. Rather, they comprise a fuel tax, a vehicle licensing fee (based on weight) and a heavy vehicle fee.

There are five key issues which are immediately when examining the Swedish system and its wider applicability from the perspective of providing an incentive for more rail infrastructure provision. Firstly, and perhaps most obviously, it relies upon the calculation of values for externalities, which opens up a very wide field of debate as to what each of these values should be. The BTE (1999) has provided detailed comparative studies of the appropriate values of externalities for road and rail transport in Australia, but the question is far from settled. Laird (2005) provides an overview of some recent Australian debate in this regard.

Secondly, the supply side works in Sweden because the funding provider for both road and rail is the same and is the state. Since the funding provider for both modes is the same, it

---

12 In 1935, the Swedish government began nationalising financially distressed private railways, a process which was complete in 1950, when the whole network became privately owned.

13 In 2000, above rail operations were separated into Green Cargo (freight) and SJ (train operations).

14 Sweden is a Eurovignette partner and levies a fee on heavy vehicles to access the road system in Sweden. The charge incorporates road damage and environmental costs (Kaupilla & Eriksson, 2001)
can jointly optimise infrastructure provision between them. Separate ownership precludes such co-ordination, and may result in inappropriate modal choices or duplication. State ownership ensures that, provided political controls are strong and transparent, additional revenues earned by including social costs in prices can be returned to infrastructure provision. There is no such guarantee if ownership is in private hands, where a manager has an obligation to shareholders to maximise returns which may conflict with the desire of the wider community for more rail infrastructure. Both of these issues can be resolved by more government involvement in the planning of infrastructure but this raises property rights issues, where government directs industry, and blurs the line between the private and public sector.

Thirdly, the Swedish system provides for uniform prices amongst rail customers of a given class. There is no reason to expect that the freight customers’ elasticities of demand or the social costs of their freight will be homogenous (compare nuclear waste with pine logs). There is thus a concern as to the efficiency of a pricing regime which imposes homogeneity in this manner, by which investment suffers in essentially the same manner as discussed previously in respect to cost allocation mechanisms and flagfall tariffs.

Fourthly, from the demand side, Swedish road charges are not mass-distance charges and hence only poorly reflect the marginal costs of road use. It is thus difficult to perceive how the Swedish system could provide suitable inter-modal demand signals.

Finally, from the perspective of delivering returns for increased rail investment is quite simply that the Swedish system does not; prior to the passage of the *Transport Policy for Sustainable Development Bill* in 1998, when fixed costs were also part of the price, access charges recovered approximately one third of the total costs of the system. Now that fixed charges have been removed, they return only between 10 and 20 percent (Carlson, 2004). Moreover, the Swedish government, not BV, receives these access revenues, so BV receives few demand signals from rail users. That does not mean there is something inherent in social marginal cost pricing which exposes it to this low level of cost recovery. However, some evidence of social marginal cost pricing approaching cost recovery would be a more useful endorsement of the Swedish system’s ability to provide incentives (rather than simple subsidies) for rail infrastructure provision.

In summary, pricing road like rail, if road users are in fact charged the full social cost of their road use and they are charged at the margin as rail users are, then the correct pricing signals to users of road and rail will be sent and thus there is at least the potential for the socially optimal amount of rail infrastructure to be provided which. However, it does not guarantee the supply of such infrastructure. In Sweden, the supply-side of this model works because the State which can fund infrastructure on an economic basis is the owner of both sets of infrastructure and is able to ensure that optimisation across modes occurs, and that the extra revenues generated by charging at marginal social costs are in fact diverted into more infrastructure of the appropriate type. With differing providers, the former is not guaranteed and with private providers, the latter is not guaranteed either, short of substantial compulsion being applied by government in both cases.

### 3.2 Funding rail infrastructure like roads: The case of the Picton-Bunbury line (WA)

The Swedish system is successful on the supply side of infrastructure provision, but not on the demand side. An alternative to the Swedish system is to sever the direct link to demand, and focus on the funding government can supply for specific projects, subjected to a rigorous form of independent cost benefit analysis to ensure that the social benefits involved outweigh the subsidy costs. There are many examples of regimes which endeavour to deliver such targeted subsidies. PATREC (2004) reviews dozens from around Australia, Europe and
North America, and discusses elements of each which are successful. To highlight useful elements of such schemes, one relatively simple example from Western Australia is provided.

In Western Australia, the State Government recently decided to assist private industry in the funding of infrastructure on the Picton-Bunbury line, with a subsidy of approximately $10 million. The decision was based on two key considerations:

- Government policies in regards to old growth forests had restricted the output of the woodchip industry to the point where it no longer generated sufficient freight to warrant the upgrading of a rail link to service its demands.
- Trucks, which are currently being used for the freight task, transit a number of small towns (chiefly Bridgetown) producing substantial local externalities in the form of pollution, noise, visual amenity and accident risk.

The decision was based upon a detailed cost benefit analysis. Moreover, the subsidy was targeted quite specifically at the infrastructure involved, to ensure that it did not become a general subsidy which could be used for other purposes by the infrastructure manager. Targeting in this manner ensured that funds could only be used for the sunk-cost component of the business, which would be returned to Western Australia (which retains ownership of the rail infrastructure) and could not be appropriated.

There are three key issues associated with targeted subsidy schemes. Firstly, it is difficult to ensure that the funding decision process remains free of politicisation. Politically-driven subsidies to rail and the soft-budget constraints they engender were a key issue which drove the global reform process in rail in the Nineties, which suggests some prudence should be exercised before any substantial expansion of targeted funding programmes.

Secondly, targeted subsidies require detailed case-by-case assessment, which can be time consuming and expensive for government (to say nothing of the potential for such rules to be gamed). It may be possible to ameliorate these costs, however, by requiring the costs of government assessment to be included in the social cost benefit analysis (or better yet, including it in the government contribution to any joint funding programme) to limit the attractiveness of such programmes to projects with only marginal social benefits.

Thirdly, and perhaps most fundamentally, widespread application of targeted government funding blurs the line between public and private sector responsibility and may impinge upon or affect the property rights of asset owners. The economic and legal consequences of this may be diverse, and difficult to forecast. However, this does not prevent targeted funding programmes from being a highly effective niche application, which may in fact be all that is necessary to bridge the gap between the amount of rail infrastructure that is commercially attractive and the amount which is socially optimal.

### 3.3 Handing over the monopoly rents: Land grant railways in the 19th Century

Given the lack of suitable, idle, low-cost land available in most developed countries, land-grant railways are more of historical interest than being actually applicable in the modern world. However, the incentive mechanisms operational in land grant railways are worthy of some consideration. In essence, land grant railway owners were given a certain acreage of land for every mile of track built, which they could then on-sell to customers (usually farmers) in exchange for meeting the relevant deadlines in their construction contract with government. These grants improved the risk profile of the nascent rail venture, easing its access to capital on financial markets.

In Canada, almost three percent of its land mass was given out as land grants in the Nineteenth Century, but even when land grants were substantial, few railways were viable.
without some form of cash subsidy from government to assist in providing sufficient security to allow the railways to become viable investments for the private sector, given long time horizons between the start of construction and the realisation of revenue streams. Subsidies were often large. For example, in dollars of 1879 when it was completed, the Canadian Pacific railway cost $25 million in private capital stock, $25 million in cash subsidies and required 25 million acres in land grants, as well as the transfer, for free, of $14 million of existing government-built rail links (Affleck, 1971). Subsidies were seldom based on the likely consumer surpluses available from a given railway. In fact, when nation-building became a consuming passion of government, 'subsidy-wars' occurred, and money was wasted. McLean (1902) provides a summary of how the winding paths of some tracks reflected the availability of municipal subsidies in Canada rather than topography or underlying economics. Cochrane (1950) provides similar evidence for the US.

Land grant railways may also represent the outcome of a trade-off between coordination, speed and income distribution. Absent of land grants, railway owners and landowners had to bargain with each other, as land under agriculture and the railways which serviced it expanded in step. Such bargains could be complex, particularly when the landowners had yet to purchase their land, and the process was slow (see Cochrane, 1950). Government expedited the process via co-ordinating it through land grants. However, their actions were not necessarily costless.

Consider the situation absent of a land grant railway scheme, where farmers bargain with railway builders on rates that will be applied if and when the railway is built and the farmers purchase land (from government) adjacent to it. Before the farmers commit to purchasing the land, they have a strong bargaining position with a railway baron. This is enhanced by the free rider problem. A railway, to be useful, must be continuous, and cannot usually bypass non-paying farms easily. Moreover, once the track is built, the marginal costs of serving a farm adjacent to it are low. The total amount which is available to the railway owner and the farmer is the economic profits available in selling farm produce to a market. The farmer, realising his ability to free-ride, will bargain for low rates, whereby the bulk of any profits available in agriculture accrue to him. Thus, absent of land grants, the distribution of income may favour farmers if bargaining occurs before he commits to a land purchase.

If government provides land grants to railways before the land is sold to farmers, however, the bargaining position changes. The railways have land adjacent to the rail track in their possession, and will be unwilling to part with it for the same low price as existed in the counterfactual above. Rather, they will offer take-it-or-leave-it prices to farmers whereby the bulk of the surplus available to agriculture accrues to the railway owner in the price of the land sold. The farmers will accept this, provided they can earn their opportunity cost of capital. There need be no net change in the amount of output produced, but the distribution of income will be different, favouring railway owners rather than farmers. Thus, land grants represent a trade-off between the speed and co-ordination of development and resulting income distribution.

\[15\] This assumes that there are not multiple rail routes or multiple owners of trackside land. If there are, then each rail owner may not be in such a strong bargaining position. In the 19th Century US, this may have been a key factor, as numerous land grant railways were looking for farmers at the same time. In the US, conflicts between rail users and the ‘rail barons’ led to the formation of the Interstate Commerce Commission and the regulation of railway rates. Thus, the distribution of incomes resulting from land grant railways in the US may indeed have been an issue. Not all rail lines in the US were developed with land grants. This presents an opportunity to empirically test the hypothesis above, by comparing income distributions along rail lines built under land grants with those which were not, and by examining the speed of development in each case. To our knowledge, no such empirical investigation has been undertaken.
4. Conclusions

There is, sadly, no ‘magic bullet’ which will ensure that more resources are devoted to rail infrastructure, short of simply providing government subsidy on demand, which is precisely the root of the problems which have traditionally beset the rail industry around the world. Moreover, there is no sure way of ensuring that the ‘right amount’ of rail infrastructure is provided. Indeed, the issue is not in the methodology used to support infrastructure in the rail industry, but rather the way in which that methodology is applied. Marginal cost pricing regimes, all of which are roughly similar, have the scope to result in ‘appropriate’ amounts of investment being made, if they are implemented correctly. Structural separation need not be an issue, provided co-ordination between the disparate elements is able to be maintained. Additional measures, involving targeted government subsidies or better demand signalling via ‘competitive neutrality’ across modes can ensure that, where there are differences between the economic and financial benefits of rail infrastructure provision, sufficient is provided that the economic benefits are realised.

However, all of these methodologies have flaws, and none, applied individually, would necessarily provide the ‘right’ amount of infrastructure. Similarly, each, if misapplied, could have substantial consequences for the efficient allocation of resources in the rail sector. Misapplication comes from not understanding the sometimes subtle ways in which each of the methodologies influences a rail industry, and the (inevitably heterogeneous) characteristics of each rail industry which interact with the regulation which is applied to them. The purpose of this paper, and of the research programme which is to follow it, is to begin to throw some light on these issues.
Bibliography


Exports and Infrastructure Taskforce (EIT) (2005) *Report to the Prime Minister by the Exports and Infrastructure Taskforce* Canberra, AGPS


Littlechild, S C (1983) *Regulation of British Telecommunication Profitability* UK Dept of Industry, HMSO


Myerson, R B (1979) Incentive Compatibility and the Bargaining Problem *Econometrica* 47(1) 61-74


Smith, B W & Tsur, Y (1997) Asymmetric Information and the Prices of Natural Resources: the case of unmetered water *Land Economics* 77(3) 392-403


