RAILWAY COSTS AND PLANNING

By Stewart Joy*

INTRODUCTION
The live issue of railway policy in Britain in the first year of the Journal of Transport Economics and Policy was the request by the British Railways Board that the government should pay its track costs. This, it was thought, would avert the need for more detailed subsidies which would involve continuous supervision of British Rail by the government. The argument, crudely summarised, was that track costs were high (£92 million per annum) and fixed, and that the nation should be prepared to pay for the “capacity” which existed to run many more trains than the community was prepared to offer traffic for (Raymond, 1966; Barker, 1967). The debate was joined in this Journal (Joy, 1967), considered by the government-appointed group enquiring into the future of British Railways (Ministry of Transport, 1967), and finally rejected by that group. Twenty years later a similar sounding, but more subtle, debate is still under way, of equal importance in determining the long-run size and scope of Britain’s railway system. Now, instead of railway managers trying to sustain the size of their operations, the protagonists are more likely to be local groups trying to protect particular services. The debate in the sixties concerned the physical characteristics of a railway system, on a search for a subsidy basis which might give a generous government an opening to excuse the railway much of its annual losses. Now that sort of argument is limited to attempts to exploit particular costing conventions and subsidy decision rules for local benefit. British Rail, like most other rail systems, now has to survive tougher tests.

The modern justifications for Britain’s railway and its subsidy were traversed by the Serpell Committee in 1982. Its report — with the minority report by Alfred Goldstein — brings the discussion up to date (Serpell, 1983). Similar clear thinking about the railway problem is seen elsewhere. This article will survey the thought processes which have brought railway analysis from crude attempts to “whitewash the deficit” to a serious search for a stable and supportable role for the railway in the space age.

FINDING THE NEW ROLE FOR THE RAILWAY

Britain’s Transport Act of 1968, and the ensuing conventions on subsidy for “grant-aided passenger service”, created a fairly clear set of rules which ensured that the “social” passenger services would pay the costs which (because of the Government’s insistence on their operation) they imposed on the railway, and that the Intercity passenger and freight services would pay all the remainder of the costs. Where freight services could be fitted in on a grant-aided line without requiring additional facilities, there was explicit provision that they need only pay for rail wear. These conventions, with detailed service-by-service reporting, were dropped in 1975, in favour of a much broader-brush approach which obliged British Rail to continue to operate all its then present passenger services in return for an agreed global subsidy (Nash, 1985). In 1982, British Rail turned to “prime user costing” (Nash, 1985, pp. 116—117); this had the practical effect of returning it to the basis evolved with the original grant-aided service conventions under the 1968 Act, but without the service-by-service detail. But the requirement to examine the substitution of buses, and budgetary pressures generally, will now cause a much harder look at the justification for the “grant-aided railway” in the next decade. In turn, this may have repercussions for the associated freight services, which will be called upon to bear a larger amount of system cost in future.

The problem of redefining the role of the railway is not unique to Britain. Most continental systems, having avoided the agonies of Beeching plans and other restructurings in the sixties and seventies, now face a similar contraction of network and reduction in their labour forces, particularly in the context of deregulation in 1992. Similar problems exist in nearly every railway system of the free world — or, in other words, every railway system which is in open competition with the motor truck.

Each generation of railway managers seems fated to undergo the same debate about the role of rail, and to take some steps of varying length toward a stable solution for its own system; but budgetary pressures on governments mean that these steps must now be big enough to hold promise of success.

There is nothing new in this; it is a problem first faced by railway managers 150 years ago, when competition came from a canal or another railway. Railways have always measured themselves against three levels of success. First, as long as revenues in one year covered cash expenditures for operations in that year, a railway could carry on while its assets lasted. Second, if the revenues also provided enough to pay the interest on its debts, all was well. Third, if after payment of interest there was enough cash left to replace assets as they wore out, and to pay a dividend to shareholders, that was railway utopia. Railway managers knew well that level one was sufficient for short-term survival; survival on level two was limited by the lives of the assets; and all hoped that traffic would grow eventually to provide for level three. Of course, some of those ideas now belong in railway history. Except in North America and in a few special cases elsewhere, railway dividends are a thing of the past. Most railways live on subsidies from their governments, both for operations and (usually) in interest-free grants for renewals and other “investments”.

46
RAILWAY COSTS AND PLANNING

S. Joy

THE RAILWAY AS A "PRICE TAKER"

From the earliest days, railway practice was to accept the rates determined by the regulatory authority, or by the shippers' competitive options, and to seek all traffic available at those externally determined rates. The level of success depended on the traffic offered at those rates. Railways were thus practised "price takers" long before they faced competition from trucks. Because they were also common carriers, obliged to move all traffic offering, they had very limited control over their businesses (Milne and Laing, 1956). In the days of railway regulation, attempts were made to derive average costs in defence of proposed tariff charges. Those averages, in attempting to cover all three levels of cost, risked overstating the relevant costs for any particular decision.

When railways won the right to negotiate their own rates, and to choose whether or not to accept the traffic on offer, it became critically important to obtain appropriate estimates of the avoidable costs of alternative courses of action. It was at this stage that the understanding of railway cost causation was found to be sadly deficient.

THE "HIGH FIXED COST" ARGUMENT

Railway policy making was profoundly affected by the belief that a very high proportion of railway costs were fixed, and that, as changes in volume would affect only the small proportion of costs which varied with output, rates for new traffic need only cover those obviously variable costs (Acworth and Stephenson, 1924; Ripley, 1916; Wellington, 1889). This argument depended on three conditions:

(1) That the costs of a railway did include a high proportion of fixed costs;
(2) That only new traffic was priced to cover only the "low" variable costs;
(3) That the revenues of all the other tracitcs covered all the other costs.

If the third condition was not met, the railway would be unprofitable. If the second was not met, gradually all existing traffic would threaten to leave unless it, too, was classed as new traffic. If successful, this threat would then violate the third condition. But the most important condition is the first: the whole argument depended on the validity of the assertion that most railway costs would remain unchanged in the face of changes in traffic volumes. This type of argument has been used in different ways to influence railway policy in different countries.

In the United States, where the Interstate Commerce Commission imposed a "reasonableness" test (which was generally more reasonable for the railroad's competitors than for its clients), railroad managers argued that it was in order to compete with barge lines by quoting very low rates, much lower than they offered to existing clients who did not have a low-cost alternative. They failed to convince the Interstate Commerce Commission of the public benefit from this discriminatory pricing practice; nevertheless it had a strong and simple theoretical justification. In some countries, for example in Britain and Australia, as road competition became intense, railway freight managers adopted the same discriminatory
practice without having to offer public justification. In contrast, in other European countries, as long as road hauliers were restrained, railway managements were much less ready to quote low rates for marginal traffic. Nevertheless, the degree of protection offered to the railway systems in France and West Germany did not prevent substantial growth in the road haulage industry. If road rates are kept artificially high to protect the railways, the high road rates make possible high standards of service, which in turn attract traffic to the hauliers.

The fallacy in this approach to railway pricing by partial marginal cost is that eventually all traffic becomes marginal traffic. All traffic is quoted rates which cover the costs which it is thought might automatically be saved if the traffic failed to move, and very soon the total revenue of the railway concerned has fallen to a level which covers only those short-run costs (Joy, 1973, chapter 5). When that occurred in Britain, the railway tried to combine good economic theory with fallacious railway costing to argue that to cover only short-run marginal costs was in the national interest (Osborne, 1957).

The theory was right: welfare is maximised, and there is no misallocation of resources, as long as the outputs of an enterprise are priced as closely as possible to the marginal cost of the resources consumed in their production. What was wrong was the railways' estimation of the amount of their resources which were consumed with production, and which would not need to be consumed if production were reduced. This error, which first saw official light of day in 1956 (Transport Tribunal, 1956), persisted through the following decade (Raymond, 1966), despite some compelling evidence to the contrary — which incidentally was based on British Railways' own data (Joy, 1964). It was first laid to rest in 1971 (Joy, 1971), but was unfortunately replaced with another, equally dangerous, failure to recognise the causation of common costs, and that was only eliminated in 1982, when British Rail adopted the prime user concept (Nash, 1985; Serpell, 1982, paras 3.28, 9.26 to 9.33, and the Minority Report, para 5.17).

These misunderstandings about the variability and controllability of railway costs have not been limited to Britain. In the United States, the convictions of many railroad managements, that a railroad company, by its very nature, had a core of fixed costs which would not rise if the scale of the company's operations were enlarged, but could not fall if the scale were reduced, were a major basis for the merger movement in the sixties. This flew in the face of a number of academic studies of the United States railroads, which disputed the existence of the core of fixed costs, and argued that, if anything, railroads became less efficient with increasing size, rather than the reverse (see section on congestion costs below). The first United States evidence on this was available in Lorentz (1915), supported by the broader study by Clark (1923), and these were consolidated by the work of Borts (1960) and Healy (1961).

More recently, beginning with the work of Keeler (1974), and extending to the modern translog cost models of authors such as Caves, Christensen and Swanson (1981), has come the distinction between economies of density and economies of scale. Certainly, increase in traffic over a given infrastructure will be met by a less than proportionate increase in costs. However, when the long-run adjustment of infrastructure to the level of traffic is fully allowed for,
the cost elasticity is much higher than it was conventionally thought to be. When the scale of a railway as a whole — the length of its route network, number of terminals and so forth — is changed in proportion to traffic, then costs change roughly in the same proportion. It is thus when changes in the volume and scope of traffic handled are sufficient for discrete changes in the scale of the undertaking that the assumption of high fixed costs is most erroneous.

THE INFLUENCE OF EXCESS CAPACITY

In the face of these studies, railway managers could have such a fundamental misunderstanding of the process under their management only because they were confused about the effects of excess capacity, thinking that it represented something fundamental to, and unavoidable in, the railway process. Possibly this arose because, in the working lives of most railway managers, excess capacity had been the normal situation: there was always room for another train, or another tonne, or another passenger. They knew nothing different, and misread the symptom of a quite different problem. Of course, railway managers were not alone in making this error; for forty years their confusions had had the support of the writings of Sir William Acworth (1924) which had a profound effect on transport policy making in Britain and elsewhere.

The belief of railway managers that extra traffic could always be accepted with only a small increase in total cost was based on short-run cost evidence which could mean either (as they believed) sharply decreasing long-run average costs or excess capacity. The simple test devised by Borts (1954) for distinguishing these took fifteen or twenty years to infiltrate railway policy making. It showed that if, with increases in output, the decrease in average costs is greater in the short run than in the long run — that is, if the decrease in average cost would be less if operations or assets were to be changed — excess capacity is implied. A number of empirical studies in the United States confirmed this: for example, Meyer, Peck, Stenason and Zwick (1959); Stenason and Bandeen (1965).

The most important task now is to identify the social and institutional factors which caused, and in some cases still cause, railway managements to retain and operate “excess capacity”, that is, a plant and scale of operations which results in higher total (and thus average) costs than some smaller plant or pattern of operation. Of course, in most countries the most important factor so far has been that governments have been prepared to foot the bill. If the representatives of the people were willing to maintain excess capacity, why should managements disagree? But that era is ending.

The technical or managerial reason most commonly offered for the continuance, in the face of losses and of a lack of demand for its services, of a railway at its current size and scope of output is that railway capacity is variable only in discrete lumps, and that variations may be made only at long intervals, when major assets are due for replacement. But short-run effective plant capacity is determined by a combination of the physical assets and the chosen operating pattern.
Alternatively, it may be claimed, as it was with the “standby” issue in Britain (Joy, 1967), that the excess capacity is “provision for peak (or emergency)” traffic. When railways had an effective monopoly of inland transport, they may have felt a social, political or regulatory obligation to retain and bear the cost of excess capacity in return for the “benefits” of higher rates in the absence of near competitors. But after the arrival of the motor truck (and the barge line) no railway can be expected to make such a gift to the community it serves.

(Of course, some cases of short-run over-capacity are due to causes well beyond the control of the railway management. A notable instance was the Indian Railway’s traffic shortfall in 1972–73, when it had provided capacity based on the optimistic projections of the current Five Year Plan. When the plan outputs were not achieved, the railway incurred heavy losses.)

THE RIGIDITY OF RAILWAY COSTS

In considering railway costs, the analyst must admit to having a great advantage over most railway managers, in having evidence from many railways. This is important, because, even if a railway has experienced a wide variety of levels of output, it is most unlikely that satisfactory cost data have been collected. Often the range of outputs will have been experienced over such a long period that other changes — peacetime/wartime, steam/diesel, and so on — will be thought to invalidate cost comparisons over time.

Even if a railway has experienced a wide range of outputs over a period short enough to eliminate the possibility of any fundamental change in operating technique, the cost/output relationships thus measured will not necessarily provide even a satisfactory short-run cost function (that is, over a period in which plant capacity remained unchanged). This is because many railway costs can be deferred for varying periods without any immediate reduction in the quality of service.

Before exploring the effects of various time horizons on railway costs, we must note an important organisational limitation affecting them; this has important implications for both costing and control. Only a very small proportion of total railway cost varies autonomously with downward changes in net output. If the demand today is unexpectedly less than the demand yesterday, the only costs which will have fallen will be the usage of consignment forms and passenger tickets, and perhaps, with freight traffic, some saving in fuel. Virtually all other railway costs can vary downward in a step only after a manager has noticed the fall in demand and has decided that the change may be permanent.

The problem is not symmetrical with upward changes in demand. With an increase in demand, unless extra capacity is already being operated, the quality of output (service) will fall: passengers may have to stand, freight will be delayed, and so on. Costs will only rise with a moderate increase in demand (again excepting a minor increase in fuel consumption) when a manager acts to restore the previous quality of service. Of course, it is possible to envisage a sudden increase in demand large enough to cause a serious and immediate increase in
RAILWAY COSTS AND PLANNING

S. Joy

costs through its effects on congestion, but again this will almost certainly follow the cycle: decline in service quality — adjustment of output to restore service quality.

The main problem of rigidity in costs is with falling demand. A number of separate factors cause this unresponsiveness in railway costs. The actual identification and confirmation of the reduced demand is difficult enough, but more important is the problem of estimating the duration of the fall-off. There is a cost of redeploying resources away from, and then back to, a particular activity, and theoretically this determines the minimum period of shortfall in demand which will justify redeployment.

CONGESTION COSTS

The character of congestion costs has been already suggested in the discussion of minor variations of demand over normal capacity, which we define as the highest level of output at which the present size of plant offers lower total costs than any larger or smaller plant. Every combination of railway inputs (track, signalling, locomotives, rolling stock, number of shifts worked per day, and so on) has a normal capacity some way below its maximum capacity. It is thus possible to exceed normal capacity, but only at a cost (or quality of service) penalty which, if the higher demand were expected to continue, would justify investing in a larger plant with a larger normal capacity.

A special feature of the interaction of demand for, and cost of supplying, railway services is important here. A railway cannot set a price at which, according to some unchanging demand schedule, a unique level of output will be demanded. Its demands vary widely over time, and a railway must equip itself for most of them. The safety valve is that, in conditions of unpredictable “overload”, it can accommodate some of the extra traffic, even though with a diminished quality of service; and its clients can store some of the traffic till capacity is available, even in the form of deferred passenger journeys.

Up to the normal level of output, the price/quality (Kolsen, 1968) combination of rail usage does not vary with changes in output. The output level at which the rail/price quality is optimal for most users is likely to be significantly lower than the cost-minimising (that is, normal) output that the railway, at a given pattern of prices, might achieve. So, if a railway seeks to maximise its net revenue by operating as near to normal capacity as possible, it risks operating at a level of output at which quality, at the given structure of prices, is unacceptable to some users. On the other hand, when demand falls below the level at which the price/quality combination is optimal, the consequent raising of the standard of service is of only marginal benefit to customers.

The result is that the railway will attempt to provide a normal capacity somewhat short of the highest peak of traffic it may be required to move, but high enough for its most likely output, and thus quality of service, to be “commercially acceptable”, that is, not so low as to encourage users to consider shifting their custom elsewhere (Walters, 1958). This stochastic variation of actual railway
output means that the conventional economists' approach to the use of marginal cost is of limited usefulness, because it assumes a constant product (that is, quality) and a perfect knowledge of future demand. This analytical difficulty can be surmounted by probabilistic methods; but it is probably not significantly less accurate to use an average of the expected marginal costs, based on an average expected level of demand. At this stage of our discussion, it is sufficient to recognise that, with uncontrollable upward variations in rail traffic, average costs may rise significantly. Unfortunately, this is no longer a very common phenomenon on railways.

MODERN RAILWAY PLANNING

There are clear implications for railway planning from all these issues of cost variability. A sufficiently long view of the future must be taken to permit consideration of substantial changes in output. The book by Beesley and Kettle (1985) on the small railway on which I had my own early training — the former Victorian Railways, now the State Transport Authority — makes these points in a manner which has wide application, and the subsequent history of rail planning on that system has proved the validity of their approach. It contrasts with the hopeful "corporate plans" of British Rail through the seventies, of which the authors of the Serpell (1983) report were so critical.

Beesley and Kettle show that there must be simultaneous and realistic projection of all the controllable and uncontrollable variables impacting on a railway, and this must start with a description of the base from which operating and marketing results were projected. The authors kindly quote from a paper of mine which I had long since forgotten, which gives the rules of approach in the vernacular:

This sort of long-term planning is the only effective antidote to the "things can't get any worse" thinking of a management already suffering heavy deficits. The facts of railway life are that if things are already bad they are likely to get much worse unless something is done about them. In fact, it is often only when it is seen how much worse things can get that management and ministers are moved to act on their railway problems. Long-term planning is also the only way in which the employment and other social implications of the necessary changes can be handled in a way which is acceptable to users, workers, taxpayers and governments. (Joy, 1977.)

That was written shortly after I returned to Australia, fresh from my three years as chief economist of British Rail during the period in which the first "corporate plan" was prepared. My experiences then showed the beginnings of later problems of planning credibility for British Rail. Every corporate plan must have a "base system plan" and a "base year". Obviously, the only valid base year must be an already completed year, for which all results are known. For this the previous year was chosen, and the "base system plan" was projected from those results. The problem was that, by the time those preparing the plan had an acceptably optimistic set of projections, the traffic and financial results of
the current year were already seriously below the base system projections for that year. The first British Rail corporate plan was thus invalid before it was printed. The complaints in the Serpell report show that later BR plans shared this fault in varying degrees, till the planning system was reformed to ensure that line management was fully involved in preparation of the plan and committed to its achievement.

Over-optimistic planning is not wholly a fault of railway planners. Governments, railway trade unions, the railway press and other vocal groups often expect railway planners to produce a document which "proves" the reversal of a trend which has now been evident for over fifty years. Once one over-optimistic plan has been offered to, and accepted by, a government, an addictive and mutually dependent relationship sets in between railway and capital funder; subsequent plans must justify the faith already shown in past government approvals of investment. The inevitable disappointments from such a mutually destructive process usually lead to the replacement of the managers, and to a search for those who will claim, in defiance of the historic record, and with a nautical simile, to "turn the railway around". Such sloganeering misses the point. Most railways are already heading in the "right direction"; their problem is the accretion of barnacles they carry with them, impeding their progress towards a balance of social benefit and social cost. These hindrances to railways' search for a stable role take two forms: services which insufficient users wish to buy, and operating methods which cost more than the increasingly competitive market is prepared to pay for.

There is no alternative to the long hard look prescribed by Beesley and Kettle (1985), or taken by Serpell (1983) in the case of British Rail. We now know that every railway's output is variable over a very wide range, and that the future is manageable through substantial change. The generally long-lived nature of railway assets means that, provided progress is being achieved, this necessary process of adjustment can be spread over a very long period, making the best use of the human and physical capital invested in each system. No government (well, hardly any government) wants to close its railway by a peremptory cutting off of funds; but increasingly governments are demanding progress toward a sustainable equilibrium as the price of interim financing.

The explicit injection of non-commercial goals and constraints into this process of rail planning is readily achievable. In a review of the Tasmanian Railways, my colleagues and I (Joy, Hicks and Kershaw, 1977) were able to frame a corporate plan for a small but grossly over-producing railway system which matched the necessary run-down to the career expectations of the existing workforce. Alternatively, the trade-offs between the cost of redundancy payments and the net cost of a lower decline in output can be measured and presented to sponsoring ministers.

There is nothing about a railway's future which cannot be predicted and presented to those ultimately responsible for meeting the cost, so that decisions made can be optimal in the total community interest. The theory and practice of railway operation has long been available and understood; all that is needed is the will of governments and managements to apply it in an open and realistic manner.
REFERENCES


Barker, A. V. (1967): “Costs of Social Service and Stand-by Capacity”. Modern Railways, 23(221), 64–65 (February).


