Handbooks in Transport
Volume 6

Handbook of TRANSPORT STRATEGY, POLICY *and* INSTITUTIONS

Edited by
Kenneth J. Button and David A. Hensher

“Edited by two of the leading figures in transportation research and dissemination, these handbooks are likely to become the essential reference work in the field.”

Dr John Preston,
Director, Transport Studies Unit University of Oxford

HANDBOOK
OF
TRANSPORT STRATEGY, POLICY AND INSTITUTIONS

HANDBOOKS IN TRANSPORT

6

Series Editors

DAVID A. HENSHER

KENNETH J. BUTTON

HANDBOOK OF TRANSPORT STRATEGY, POLICY AND INSTITUTIONS

Edited by

KENNETH J. BUTTON
The School of Public Policy,
George Mason University

DAVID A. HENSHER
Institute of Transport Studies,
University of Sydney

ELSEVIER

2005

Amsterdam – Boston – Heidelberg – London – New York – Oxford
Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo

ELSEVIER B.V.
Sara Burgerhartstraat 25
P.O. Box 211
1000 AE Amsterdam
The Netherlands

ELSEVIER Inc.
525 B Street, Suite 1900
San Diego
CA 92101-4495
USA

ELSEVIER Ltd
The Boulevard
Langford Lane, Kidlington
Oxford OX5 1GB
UK

ELSEVIER Ltd
84 Theobalds Road
London WC1X 8RR
UK

First edition 2005

Library of Congress Cataloging in Publication Data
A catalog record is available from the Library of Congress.

British Library Cataloguing in Publication Data
A catalogue record is available from the British Library.

ISBN 0-08-044115-7
Series ISSN: 1472-7889

∞ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).
Printed in The Netherlands.

INTRODUCTION TO THE SERIES

Transportation and logistics research has now reached maturity, with a solid foundation of established methodology for professionals to turn to and for future researchers and practitioners to build on. Elsevier has marked this stage in the life of the subject by publishing a landmark series of reference works: *Elsevier's Handbooks in Transport*. Comprising specially commissioned chapters from the leading experts of their topics, each title in the series encapsulates the essential knowledge of a major area within transportation and logistics. To practitioners, researchers and students alike, these books are authoritative, accessible and invaluable.

David A. Hensher
Kenneth J. Button

This Page Intentionally Left Blank

CONTENTS

Chapter 1

INTRODUCTION

KENNETH J. BUTTON
George Mason University, Fairfax, VA

DAVID A. HENSHER
University of Sydney

1. Introduction

Transport strategy and policy is a wide-ranging subject. It embraces, among other things, the collection of data (and its transformation into information), the formation of policy objectives by government (at all its different levels), the establishment of institutional structures to carry through these goals, the creation of the resourcing for these institutions, the carrying through of actions, and the policing and monitoring of outcomes. It covers a wide range of disciplines – economics, politics, law, physical planning, psychology, engineering, etc. – and much of it is normative in its orientation.

From this cursory listing it also becomes clear that transport policy formulation is far from institutional-neutral. It is something that varies with time and place, and seldom rides on the back of a consensus view. It depends upon the underlying political and philosophical underpinnings of the society concerned, and often what scientists consider rational choices give way to beliefs, norms, and prejudices that are inherent in the institutions of the country.

In part the situation arises because transport policy is often remote from what are normally viewed as transportation considerations. Political image is often important, as with the propaganda in Germany accompanying the construction of the autobahns in the 1930s, and the Italian fascists on their ability to run trains on time; and transporting men to the moon was hardly a rational economic action. But even when more narrowly defined economic and social factors come into play there can be seriously divergent views on the appropriate transport policies to adopt.

As an example, the overall approach to transport policy varies considerably between countries. At one extreme is the Anglo-Saxon perspective that leaves much to the individual and market forces, with government simply being a

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

facilitator with minimal intervention when serious market failures emerge. At the other extreme is the Continental European philosophy of direct government planning and control, with markets treated cautiously and only allowed when they meet clear official objectives. The underlying view of where transport policy fits within the wider social and political systems also differs between the two schools. Anglo-Saxon policy seeks efficiency within transportation itself whereas the Continental approach sees transport as an input to a larger vision of macroeconomic efficiency and equity – for example spreading national welfare across all regions and fostering particular types of industry. Of course, these models are stereotypes, and reality hovers somewhere away from the extremes, but they do represent tendencies.

The context of transport strategy, policy and institutions changes with time. This is in part influenced by technology shifts; safety policies designed for horse-drawn transport moving goods at four miles per hour have only limited relevance in the automobile age. Social change and shifting political thinking also influence it. Higher real incomes mean that society often places more weight on equity considerations when thinking about transport strategies, and with this the insurance that the lower-income groups and the physically impaired have a reasonable level of mobility. New intellectual ideas, clarification of theory and enhanced knowledge can also lead to sea changes in policy, or at least offer support for reassessment of existing policies. For example, the move toward more liberal transport markets from the late 1970s (the misnamed "deregulation" that is often traced to the 1978 Airline Deregulation Act in the USA) was partly influenced by an enhanced understanding of how markets function, how public policy can be captured by vested interests, and by innovative statistical analysis of the regulated transport industries.

2. Changing themes in transport strategy, policy, and institutions

Historically, transport was important in shaping human geography, and in particular it was a facilitator of the development of urban society. Early cities needed food from their hinterlands, and organized transport systems were developed to provide this. Most revolved around rivers (which also conveniently irrigated crops), and with this came policies and strategies for port developments, and more locally for planned urban distribution systems (streets). But it also entailed securing rivers from attack, and this in turn led to their use for offensive military actions. Good transportation – and the availability to develop efficient logistics – was a core component for military strategy from pre-classical times.

The availability of the water routes of the Mediterranean and rivers initially focused transport policy on maritime matters in the West, but in parts of Asia long-distance road communications systems developed to support a succession of mega empires, most notably in China and India. In Europe, as the Roman Empire emerged it also moved inland and constructed a highly efficient metalled road network – an army could march from Rome to northern Spain in 27 days. The Roman Empire also created a vast system of aqueducts to move water to its cities – some enjoyed three times the per capita water supply that they have today. These actions, which were very highly directed from the center, largely involved standardized engineering techniques, and were financed from public funds, were deemed essential for the political and social cohesion of the empire. Economic assessment and the shadow price of construction (particularly in terms of human lives lost in construction) hardly entered the arithmetic. Social service was not at a premium.

The Dark Ages in part lacked illumination because of a breakdown in the transport system of Europe. The roads and ports of the Roman Empire went into decay as society focused more on local than on global matters. While there is evidence of considerable movements of clerics and scholars across Europe during this time, indeed some studies suggest on a par with the modern "frequent flyer professor," this was very much the exception rather than the rule. The empires of the East, in contrast, continued, despite various waves of conquest, to maintain good communications systems as part of imperial policy and as infrastructure for continuing trade.

The Renaissance in Europe brought with it a greater interest in transportation. As trade grew, and exploration took place, the mercantile city states on the Mediterranean and in Northern Europe consciously developed their ports and inland transport networks. There was public investment in shipping and the funding of expeditions of discovery as a clear strategy to build up national wealth. The prevailing economic orthodoxy helped. This was the beginning of the high-water of mercantilism. The belief was that the more bullion – gold and silver – a country had the stronger was its economy and the larger its political influence. Trade was vital to earn these precious metals, and new colonies would provide added wealth. Hence, a substantial merchant fleet with associated infrastructure was needed and maintained, often with state support.

The policy changed somewhat in the later part of the eighteenth century as the *laisser-faire* ideas of Adam Smith took hold in many countries, and the Agricultural and Industrial Revolutions unfolded. While the mercantilists viewed trade as desirable provided a county was in a balance of payments surplus – hence their fondness for import barriers of all kinds – Smith, Ricardo and others convinced the UK in particular that free trade *per se* enhanced national economic wellbeing. They also argued that markets were more efficient in themselves, and that the freeing of the "invisible hand" offered the best strategy for economic growth. Hence, transport should be left to market forces. Privately financed toll roads (known as turnpikes), shipping lines, canals and the railways thus expanded in the UK and also in many other parts of the world. This notion of free trade was supported by the creation of a near-global currency regime, the "gold standard," that facilitated easier transactions and commercial intercourse.

Within this apparent conquest of Europe by the Anglo-Saxon philosophy there were some countries such as France and Germany that retained a much more centralist approach to transport policy. But even among those that largely favored private sector supply and the flexibility of market forces there were numerous government interventions. In many cases, as with land acquisition for canals and railway construction, the strategy was one of coordination and facilitation of enhanced transport. In some cases, as for example the UK government's support of steam-turbine power for maritime transport, it was for a wider policy that formed part of the administration of a global empire. In yet other cases, as with regulations over railways, it had microeconomic underpinnings, in this case the containment of monopoly power, or social and safety connotations.

One should also not overlook the military strategies that became entwined in the shifting transportation policy landscape of the early industrial era. The railway was a powerful instrument of troop deployment in the US civil war, and this influenced transport network design in Europe – France, for example, after 1870 had a sparse rail (and also road) network in regions vulnerable to German incursions to prevent its use by attackers, and Britain's coaling stations for its merchant marine were also conveniently located for Royal Navy use. Transport policies and strategies in this way frequently transcend short-term economic and social considerations although the academic literature often forgets this.

The post First World War period saw in many countries a greater public sector role in regulating transportation. New modes, such as air transport, were controlled almost from their outset. This state control was partly for safety reasons, but also it was seen as an important national unifying force by the US federal government and by the UK administration as a way of governing its empire – hence subsidies were common. The US passenger air transport industry, for example, developed from the 1920s on the back of airmail subsidies. But even local forms of transport became more regulated, and the 1920s and 1930s saw a slew of laws passed to control public modes such as buses, taxis and trucks in many countries. One reason for this increased legislation was to help older forms of transport, especially the railways, whose fortunes had begun to wane, but there was also a concern that excessive competition was a problem if reliable and safe services were to be ensured. Stable service, it was argued, could not be assured if suppliers were forever changing their networks in pursuit of a commercial return.

The regulatory phase was complemented in some countries by the nationalization of parts of the transport system. Roads had generally come under state ownership gradually from the end of the nineteenth century as the railways had eroded their toll-based financial structures. But in the late 1930s, and more so after the Second World War, more widespread nationalization took place. The need to reconstruct after the war was one of rationalization; the private sector simply could not muster the necessary resources. There was also seen the need to rapidly invest in high-quality inter-urban road networks to meet the needs of the growing manufacturing sector, and to satisfy the needs of the rising number of car users. But in many cases there were ideological overtones, with transport being seen as a public service with all citizens having a right to high levels of mobility. This was supported by the prevailing idea that with scale economies, and the need to coordinate transport services, the state was best suited technically to control the transport sector.

This approach of, predominately, planning transport supply, frequently not only in terms of infrastructure but also operations, through regulation and state ownership prevailed largely unchanged until the late 1970s. It tended to be reinforced by the apparent economic success of the eastern and central European communist states, which had totally planned and coordinated transport systems, that were believed to be out-performing western Europe and the USA.

Downturns occurred in the economies of many countries in the late 1970s and early 1980s, and policy makers sought ways of combating the macroeconomic "stagflation" – combined inflation and high levels of unemployment – that was prevalent. Academic evidence that regulation was keeping prices high, stymieing economic growth and hampering technical development, led to the removal of many regulations over transport, and the privatization of large parts of transport systems, in many countries. This was a pattern that was witnessed almost universally, and took a dramatic twist with the collapse of the Soviet-style economies in 1989; it was recognized that the image of efficiency that had emanated from the communist-controlled states was largely just that, an image.

An added twist and complication to policy formation from the late 1970s was an increased concern with the environmental damage that accompanies transport. Initial public interest centered mainly on local pollution (notably lead emissions from cars) and noise nuisance, but subsequently extended to include acid rain and global warming gas emissions. As a consequence, while the economic regulation of transport (over prices and output) generically tended to become weaker, there was an upsurge of social regulation ("quality regulation" in European terminology) aimed at reducing the environmental impacts of transport. The issues surrounding this have been discussed in contributions to a companion volume, the *Handbook of Transport and the Environment*.

A further recent trend has been in terms of an enhanced focus on policies and the regulation regimes that govern international transport. International transport policy was a major concern after the Second World War, and the formation of the International Civil Aviation Organization and the International Maritime Organization under the umbrella of the United Nations provided agencies aimed at coordinating international policies in air transport and shipping. There have subsequently emerged a number of super-regional groups of nations involved in varying degrees with economic and political integrations (for example the EU, NAFTA and ASEAN). With free trade a key component of these accords and unions, there have come reductions in the regulations governing international transport. This trend has been compounded at the global level as various initiatives to open transport markets outside of these formal regional groups, either on a bilateral (e.g. the US Open Skies policy for international aviation) or a multilateral basis (e.g. through the World Trade Organization), have taken place.

3. The elements of transport strategies, policy, and institutions

Transport strategies, policies and institutions provide a wide-ranging topic. It is not only a matter of defining the terms but also of deciding where boundaries lay. Transport, for example, interacts with land use, and the direction of causality is seldom very clear. Indeed, the phrase "land use and transportation planning" is widely used when it comes to urban design. There are, therefore, issues of just how far one needs to look at land use and related policies when reviewing transport strategies. There are also issues of how far up the value chain of transport supply it is useful to go. This chain includes such things as vehicle production and design, and policies relating to hardware can certainly influence the ways that governments develop strategies for transport use. The UK's fostering of the steam turbine cited earlier is an example, but one may also add in, beside others, the European support for Airbus, and its influence on the air transport market.

The most conventional approach, largely because of its tractability, is to focus mainly on the final delivery stage of transport and to treat other interactions as implicit or peripheral. This is what is done in this volume. Those interested in matters relating to the interaction of transport and land use are referred to material in the *Handbook of Transport Geography and Spatial Systems.*

Initially, it is important to understand the larger institutional structures in which transport strategies are developed and refined. In particular, in the early part of the twenty-first century this involves the consideration of transport markets, market structures, market failures and the ways in which official policies seek to confront market forces. The challenges that confront governments include whether there is a need to regulate prices or output, to subsidize or tax, and whether to take over ownership or not. Some of the issues concern social and environmental considerations, but since, as noted above, these are covered elsewhere, they are not dealt with here. What are considered here are policy and strategy positions that governments take regarding the economy nature of transport industries.

What is covered within this volume, and does overlap to some extent with both land use and environmental matters, is transport planning. Although there has been a shift toward making greater use of the market, there is still extensive planning of transport, and especially of transport infrastructure investment. Even here, however, there have been changes. Some of these have been in terms of the modeling frameworks that are used to assess the implications of alternative plans (topics covered in the *Handbook of Transport Modelling* and not considered in detail here), but there have also been shifts in overall planning philosophies, the objectives being sought, and the formal institutional structures in which plans are carried through. Coupled with this have been developments in the ways that the implementation of plans and associated investments can be financed and in decision-making about administrative effectiveness.

There are also important matters concerning the forms that regulations over transport operation should take where there are market failures, or when transport is being used to achieve specific, non-economically efficient goals (e.g. providing social services). Combined with this is the matter of how various strategies and policy options can and are evaluated. Techniques of policy assessment have become considerably more sophisticated over time, and the strategy/policy analyst now has a large portfolio of tools available that extends beyond the traditional cost-effectiveness procedures. Some of the key issues in regulation, together with tools of policy evaluation, are looked at largely from a case study perspective.

Finally, there are significant differences in national approaches to transport policy, and to the regulation of transportation. These differences are partly due to variations in national geography and legacy effects, but also can reflect variations in priorities and in approaches. There are potential lessons to be learned from the experiences of different countries or unions of countries.

4. The Handbook

To put things into a broader context, the *Handbook of Transport Strategy, Policy and Institutions* is the sixth volume in the *Handbooks in Transport* series. The earlier volumes have been concerned with transport modeling, with transport logistics and supply chain management, with transport systems and traffic control, with environmental matters, and with transport geography. The coverage of the handbooks is meant to be neither comprehensive nor always excessively deep; they are neither textbooks nor are they research monographs. Each chapter in a volume is designed so that it can be read by those new to a transport field as well as by those who are already familiar with the area but wish for some updating. As such, each chapter has been written in a way that readers will gain an overview and useful insights. At the same time, those readers who are already practitioners in that field will gain knowledge through the presentation of advanced and/or updated tools and new materials and state-of-the-art developments.

The *Oxford Dictionary* definition of a handbook is that it is a "guidebook" or a "manual." In other words, it is a practical tool to help its readers carry through particular activities or operations. It is not a textbook. A textbook is, again deferring to the *Oxford Dictionary*, a "manual of instruction" because it goes beyond a simple pedagogic device. It is also not a monograph that offers a "separate treatise on a single subject or class of subjects." A handbook contains information and concepts that are useful and offered in a concise fashion – indeed, that is where the term derives from; it can be carried in the hand.

The current handbook may be useful in some contexts for instruction, but its main aim is to help those involved in transport with a focus on the environment to perform their tasks effectively and efficiently. As anyone using a handbook knows, there is always the problem of level. The key thing about a handbook is that it should be accessible to those that need to consult it.

The coverage in this volume reflects what are generally seen as key subject areas. In this case, since the handbook is part of a more extensive series, some topics have been allocated to other volumes simply because they also have as much right to be there as in this book. When designing the framework for the volume it was decided to be as contemporary as possible and to bias the contributions toward articles that reflect the current state-of-the-art rather than to simply set down what is often the current practice. This has the limitation that the handbook moves a little away from being a strict "manual," but it does offer what we feel will be a more enduring body of work as a result.

Although any classification of themes is to some degree inevitably arbitrary, we have divided the subject into a series of topic areas that were broadly outlined in the previous section. This represents our effort to typify what researchers and practitioners see as the foci of any structured study in which strategy and policy is a central input. There has been no effort to completely standardize the different contributions. This would destroy the individuality of the authors, and also artificially disguise the fact that there is no consensus as to how policies and strategies are developed in transport or how they are implemented or how effective they are. Subjects in the policy arena are inevitably fluid, and ideas and methodologies quite correctly change as new information emerges and new thoughts are stirred. Indeed, the material set out in this introduction is largely subjective, and there are others who would take a somewhat different approach to the subjects covered. The common denominator to the approaches adopted is the deployment of a mixture of a synthesis of methods and case studies.

Part 1

INSTITUTIONAL SETTINGS AND MARKETS

This Page Intentionally Left Blank

Chapter 2

MARKET AND GOVERNMENT FAILURES IN TRANSPORTATION

KENNETH J. BUTTON
George Mason University, Fairfax, VA

1. Introduction

Transport markets are subject to a wide variety of regulations and controls. This seems to have been the case from the earliest of times when community leaders controlled transport for military and governance purposes. Such controls continue to exist today, but there are now many more interventions by government for economic, environmental, and political motives. It is these latter forms of intervention that are examined here. Essentially the chapter looks at why in economies that are largely based on market structures, transport is so often regulated, and it considers whether such regulation is always needed or desirable.

There has been a general tendency since the publication of Adam Smith's *Wealth of Nations* in the eighteenth century for markets to be seen as largely serving the public interest. The invisible hand of individual, self-motivated actions is accepted as leading to efficiency in the productive system. This, what is often called the Anglo-Saxon approach to economics, does not preclude government intervention in very specific cases of "market failure," some of which were highlighted by Smith himself, but the onus is on proving that interventions will improve the situation. Market failures *per se* do not justify regulations and controls; they are only desirable if they demonstrably bring about improvement.

The Anglo-Saxon approach can be set against that of the Continental European philosophy. The latter sees transport, or indeed any other sector of the economy, as part of a larger process that requires manipulation to serve a variety of public interests. While the Anglo-Saxon approach seeks to make transport supply efficient in itself, the Continental approach is less concerned with transport efficiency but sees transport as an input into meeting regional development objectives, bringing about changes in income distribution, and fostering such things as trade. Markets do not necessarily achieve this, and so there is a need for government interventions to steer the transportation sector in the desired direction. Markets are not ignored, but the onus of proof is on showing that they help achieve these larger goals.

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

Of course these are extreme positions, and advocates of one or the other tend to look at issues much more in terms of the appropriate degree of intervention, and the nature of interventions, rather than advocating markets being completely abandoned or government involvement being entirely absent. But the divergent positions do impact on transportation policy debates, and, for example, one of the problems that existed for many years in the EU in its efforts to formulate a common transport policy was the dichotomous philosophy held by various groups of member states (Button, 2004a).

The approach here is to follow the Anglo-Saxon view that markets have many intrinsic merits, and that it is desirable to largely leave them to determine the nature of transportation systems. Interventions then have to be justified in terms of remedial actions to rectify situations where markets show deficiencies.[a] This is a useful pedagogic device because it does allow simplifications to be introduced into a very large subject arena. The literature on market failure, both generally and in the particular case of transportation, is very extensive and has a long pedigree. Combined with the theoretical discussions there is a large body of empirical literature that has sought to quantify market failures and to assess the implications of various policy actions to counter them. Indeed, progress has been considerable in the econometric and operations research fields over recent years, and a wide range of quantitative techniques are now available for assessing the scale market failures and the potential (and *ex post*, actual) affects of governments becoming involved. These techniques are not, however, discussed in any detail here.

2. Nature of market failures

Market failures take a variety of forms. They can also involve normative as well as positive economic issues. The aim here is to largely ignore normative matters, such as the distribution of costs and benefits that can emerge in markets, and to focus almost entirely on matters of economic efficiency. The rationale for this is not entirely one of space limitations but embraces the idea that an efficient system will always generate the maximum output – how this is shared out then becomes a political judgment; a point made by Robbins (1932) many years ago. Of course this is a gross simplification when incentive structures are brought into the analysis, and when there are transaction costs involved in carrying through transfers of costs and benefits, but it does provide a tractable framework for discussion.

[a] The distinction between market failures and government intervention failures is in fact a rather blurred one. In the sense that government has control over any society then any failure is in that sense its failure. The particular advantage of taking the market as a "strawman" and looking at how it may not function perfectly offers a way of examining some of the problems that transportation policy-makers encounter in reacting to these problems and how they may actually make them worse.

There has been something of a revolution over the past 30 years in the way in which market failures are viewed, and in the types of policies that have been advanced to tackle their worst distortions. These changes in thinking have been wide-ranging, from new ideas about the situations that bring about some types of failure to how failures when they do exist should be handled. In some instances, new theories of how markets function have questioned whether market failures are as common as once thought, and whether their implications for economic efficiency are of the form once thought dominant. A new battery of regulatory tools have been crafted to handle more precisely the types of market failures that can be serious. Indeed, looking across a range of countries it is clear that the perception of how transport markets work has changed quite dramatically, and with this has come a spate of privatization and deregulation (perhaps more accurately, liberalization) measures that have allowed market forces to function more completely. Even where traditional types of markets are absent, there have been efforts to initiate quasi-market situations such as the establishment of a Crown corporation in Canada to own the air traffic control system, and the initiation of road congestion charging in the UK in London.

The aim here is not be comprehensive in the coverage of market failures in transportation but rather to focus on those areas where there has been recent development in thinking, and also on several cases where this has actually led to some reformations in policy. There has been significant concentration on the reassessment of market failures since the 1970s, and, while not exclusively involving transportation markets, many of the more generic elements in the new understanding have emanated from issues arising in that sector.

2.1. Market power

Monopoly power in markets has traditionally been the subject of most concern among policy-makers, and has attracted the greatest attention of academics (Kahn, 1988; Train, 1991; Foster, 1992). Much of the legislation in Europe and the USA when railways were introduced in the nineteenth century aimed at limiting the exploitation of market power, and it has often been the motivating factor for state provision of transport infrastructure. On the operational side, there have been extended debates about how to handle cartel-style arrangements such as shipping conferences and airline alliances, and various forms of competition laws have emerged to control such practices as predatory pricing.

The traditional view of looking at market power from an economic efficiency perspective was in terms of the potential allocative efficiency losses that it can create. The situation is often depicted diagrammatically as in Figure 1, which also, for pedagogic reasons, makes the simplifying assumption of constant marginal costs. The optimal price in this case that maximizes social surplus is where

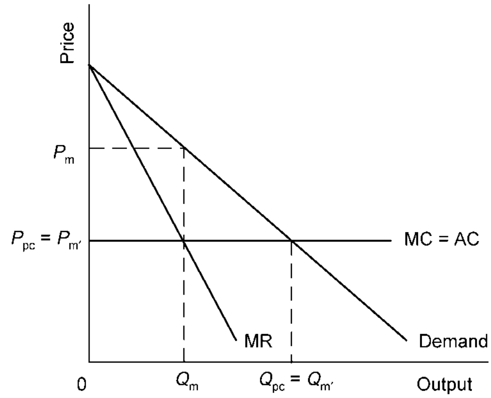


Figure 1. Monopoly power.

marginal cost (MC) is equated with demand P_{pc}, whereas a profit-maximizing monopolist will price at P_m, with a commensurate reduction in output and social surplus. What, however, became clear by the 1960s was that allocative inefficiency was not always the primary problem with monopoly power. There was also often the matter of X-inefficiency (Leibenstein, 1966). Monopoly power creates much less incentive for managers, and in particular those in undertakings where management is separated from ownership, to minimize their costs of production. The result is higher marginal costs that push up prices, limiting output even further than the traditional model implies. The lack of incentive is a particular feature of the principal-agent problem when ownership of an undertaking, mainly due to limited information and ineffective incentives, cannot ensure that managers will push for maximum profits.

The policy response to allocative inefficiency in monopoly had traditionally been one of command-and-control instruments to regulate price, output, or a combination of the two, coupled with competition policies to limit cartel-type actions. Rate-of-return regulations were, for example, widely used in the USA from the 1930s for rail, trucking and air transport regulation. Simply, transportation suppliers under such regimes were allowed to charge a price equal to costs plus a stipulated mark-up (intended to approximate for the risk-adjust normal profit). The difficulty was that regulators needed information from the transportation undertaking in order to estimate the costs of supply. This gave limited incentive for the regulated to minimize costs – any increases could simply be passed on in fares or cargo rates, as would any cost savings. The result was laxity in management, and higher prices, not because of allocative inefficiency but rather because of X-inefficiency. Classic studies indicating the effects of this on price include work on the US domestic airline industry where comparisons were made between unregulated intra-state services and comparable, but rate-of-return-regulated, inter-state services (Levine, 1965).

Added to these static issues was the potential effect that monopoly power may have on the dynamics of transportation industries. Monopoly had traditionally been seen to have the redeeming feature that it allowed for the super-normal profits that were earned to be used for investment in large-scale infrastructure and for conducting research and development. Transportation undertakings subject to high levels of competition were seen as largely incapable of amassing the resources to carry out these functions optimally. While the rent acquisition of monopoly power offers the potential for innovation, the motivational factors of managers of monopoly may however be inadequate to bring it about. Tullock (1965), for example, argued that monopoly power breeds pressures to use profits to fortify the monopoly position through lobbying, buying up competitors, etc., rather than to use it for productive investment. But even without this, when there is X-inefficiency there are not only potentially inadequate incentives to use current resources efficiency there is also the potential for dynamic X-inefficiency and a failure to exploit and develop new technologies.

Predatory behavior, whereby incumbent suppliers act to limit new market entrance, has both the potential for allocative and dynamic X-inefficiency (Dodgson et al., 1991). It occurs when an existing supplier confronted with a new market entrant acts by dropping its own price below marginal cost, or flooding the market with its own services, to make the position of the newcomer untenable. Once the new threat is forced from the market the incumbent raises prices and reduces output to earn super-normal returns. In allocative terms, this leads to higher than normal profits in the long term, and in X-efficiency terms can prevent new, innovative suppliers with lower costs from gaining critical market share.

Perhaps a more fundamental matter that has arisen is the issue of whether monopoly power is quite as pervasive as many policy-makers traditionally argued. If it is not, then the edifice of market interventions that has been used may not be needed, or it may need to take another form. The important development here was the refinement to contestability theory (Baumol et al., 1982). Traditional market analysis had almost entirely focused on the role of actual competition as the means to limit market power. Where this was missing, government intervention was seen as necessary in the public interest. Contestability theory takes actual competition as a special case, and looks at markets from a broader perspective that also embodies the role of potential competition.

The idea is that if a market has totally free market entry, and that exit involves no sunk costs (in the sense that leaving a market does not leave any stranded resources), then the threat of "hit-and-run" entry will force even a monopolist to refrain from market exploitation. Of course, in reality totally free market entry and exit do not exist, but advocates argue that even imperfectly contestable markets can be more efficient than government anti-trust measures. This type of argument has underpinned a range of transportation market liberalizations around the world. Evidence about the power of potential competition has, however, been somewhat mixed.

Perhaps the most studied market has been the US domestic airline industry, although there the original deregulation had largely been premised on the role of actual competition. Briefly, a number of pricing studies highlighted the imperfections of contestability, giving only support at best to the weak contestability hypothesis that potentially competitive elements do have some effect, but not enough to yield prices comparable to those generated by actual competition. Call and Keeler (1984), for example, found a positive link between market concentration and the profit of airlines, indicating that potential competition was not always generating normal profit levels, while for market entry they found that incumbents responded by cutting their fares; an action not consistent with them previously being under contestable pressures. Morrison and Winston (1987) did find that potential market entry could limit the actions of incumbents but only when the number of potential entrants exceeds three. Moore (1986) found that there had to be five or more actual carriers on a route for any downward pressure on fares.

2.2. *Externalities*

Externalities in the simplest terms are costs or benefits that affect people who are not parties to a transaction. They are external to the price mechanism and, as the Nobel Prize winner Ronald Coase (1960) highlighted, come about because of an inadequate allocation of property rights within markets. Externalities take a diversity of forms. The focus here is on technological externalities, rather than on pecuniary externalities. The former involve real resource costs, and ignoring them in decisions leads to societal costs. The emission of greenhouse gases is seldom considered by automobile drivers but is likely to have long-term detrimental effects for future generations. A pecuniary externality, in contrast, has no net resource implications but is a transfer of costs or benefits between groups.[a]

Technological externalities are widespread within transportation, and take a number of distinct forms. Traffic congestion, for example, is a particular type of externality related to the club good nature of roads. It is a user-on-user external cost because one person's use of a road affects the speed and travel costs of other road users without the initiator taking full account of these costs (Lindsey and Verhoef, 2000). In contrast the various forms of intrusion caused by mechanized

[a]This does not mean that policy-makers have no interest in pecuniary externalities. They are often as concerned with who gains and loses as a result of their actions as they are with the aggregate effect.

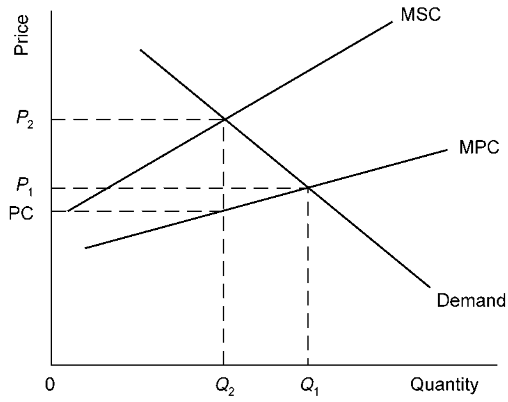


Figure 2. External costs. MSC, marginal social cost; MPC, marginal private cost.

transportation on the environment (e.g. air pollution, noise, soil contamination, and visual intrusion – see Hensher and Button, 2003) are largely forms of user-on-non-user technological externality in that those adversely affected are not those actually making trips. Both of these types of externality have the clear characteristic that they affect the overall welfare of society.

The situation of technological externalities and their implications is normally depicted as in Figure 2. For ease of exposition, and to avoid some of the peculiarities of traffic congestion externalities, we consider an environmental problem, the nitrogen dioxide (NO_2) air pollution that is often associated with cars.[a] The quantity of traffic on the road (usually measured as flow) increases, and this pushes up the time and money costs of trip making as congestion rises. The marginal potential road user will take this into account when deciding whether it is worth entering the traffic flow (the MPC curve), as well as the benefits of the trip (as proxied by the demand function). The resultant traffic flow is Q_1. If the car user were to take the effects of their NO_2 emissions (namely acid rain damaging agricultural products and polluting water) into account, then the reaction function would become the MSC, and the resultant flow would be at the lower level Q_2. By not taking due allowance of the external cost, there is an excess of traffic and a loss of social welfare (which can be measured in terms of lost consumer surplus) because someone ultimately bears the cost of the NO_2 emissions.

[a] The complexity with the congestion externality is that in some cases, such as "simple congestion," there is a technical link between the MSC and the MPC curves because they are derived from the same speed–flow relationship. This is not so with environmental externalities.

A range of policy options are available to optimize the level of external costs, and these are covered in detail elsewhere – see Hensher and Button (2003). Suffice to say here that the most efficient strategy for optimizing negative technological externalities depends on context. While strict theory may point toward a more complete allocation of property rights, in practice a portfolio of other measures may prove a more realistic approach. Externalities are also partial equilibrium concepts that depend upon the domain in which they are reviewed (Bonnafous, 1994). A particular market failure of concern in the twenty-first century concerns global warming that extends beyond the normal domain of external effects within a partial equilibrium framework. These are discussed in more detail below in the context of government intervention failures.

The discussion so far has been on negative externalities, but positive externalities may emerge in some instances. Positive externalities occur because their creators do not realize the full extent of the benefits their actions convey on others. The textbook example is the immunization of individuals against smallpox that not only protected them but also limited the spread of the disease to others. In consequence, activities involving positive externalities tend to be undersupplied if left to the market. But there are also factors that limit the situations where positive externalities exist in transportation. This is because there is a strong incentive for the perpetrator to internalize the externality and enjoy the benefits him- or herself. An example is the case of "grissers" (UK train spotters) that may be charged to spend time on railway stations to enjoy looking at trains whereas previously they were allowed to do so for nothing.

There is sometimes a tendency to incorrectly look at indirect benefits as external benefits, especially in situations where a transportation investment affects land values and brings about local economic development. Some of these latter gains are clear spatial transfers, one area gaining at the expense of others, but even when adjustments are made for this there is often a claim that the residual increase in land values reflects an external benefit of the transportation investment. In perfect markets this change in land values is, however, simply the shadow of enhanced transportation access that is captured in the gains to the transportation users in terms of time and money costs of travel. To include both the land value changes and the travel benefits is double counting. If the markets are imperfect, then there may be some external benefits associated with the transportation investment, but their calculation is not straightforward.

2.3. *Excessive competition*

Over the years there have been many policy initiatives that have sought to contain the problems of excessive competition. These include licensing systems in many cities to limit the number of taxicabs (and often to prevent jitney operations from

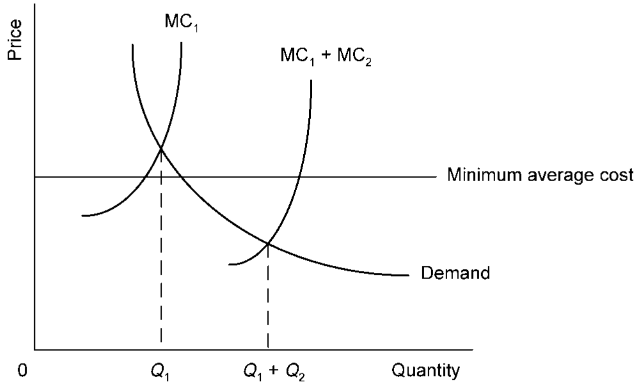


Figure 3. An example of an empty-core problem – the "Viner case."

competing with them), the licensing of bus and other scheduled services, and the relaxation of antitrust laws to allow incumbent transportation suppliers to protect their markets through cartel agreements (as with shipping conferences and airline alliances).

There are a variety of reasons why excessive competition may arise, but when it does it is seen as potentially destabilizing in the short-term, leading to under-supply in the market in the longer term, as well as acting to reduce the incentive for innovation. The particular economic issue is essentially one of whether excessive competition prevents the emergence of a "core." In technical terms, a core exists when there is one coalition of actors in a market that can dominate that market, and thus achieve a stable equilibrium. Put another way, in the case of excess competition the market will not produce some transportation services even if the public is willing to pay for them because unfettered competition will push fares/rates down to short-run marginal cost, leaving no margin to cover the full costs (Telser, 1994).

Much of the theory of markets relies upon assumptions of constant costs, or "U-shaped" cost curves, but in practice there may be a variety of factors that result in transport being supplied with decreasing costs. In some cases this involves the traditional notion of fixed costs that are spread over increasing units of transport, but it may also involve indivisibilities. A standard example is the "Viner case." Figure 3 shows a scheduled airline confronted with a downward-sloping demand curve offering a service between A and B. It has a marginal cost curve of MC_1, and a constant minimum long-run average cost curve. The negative slope of the marginal cost curve implies that there are some fixed costs. Assuming the carrier seeks to maximize profits, then it will price to at least cover its minimum average cost – i.e. offer an output less than Q_1 where the marginal cost is equivalent to the marginal revenue, MC = MR, with the concomitant price established by the demand schedule. A second carrier with identical costs entering the market in pursuit of some of the rent earned by the pioneer results in a combined marginal cost curve of $MC_1 + MC_2$. In the absence of any agreement, the carriers with a commitment to their schedules will try to fill their capacity, but in the absence of any market power they will compete fares down to marginal cost – in the case where $MC_1 + MC_2 =$ demand. This is an output and fare combination below that required to fully recover costs. Either one or both carriers will eventually leave the market. With full information (essentially rational expectations), the risk of not recovering costs will limit the market supply below the optimal level.

While the Viner case is the most discussed, there are prospects of empty cores emerging when there is a low elasticity of demand, when firms have wide variations in their cost curves leading to fragmentation of the market supply curve, and when there are serious fluctuations in demand. Whether this type of market failure is a serious one in transportation markets is debated, but there are arguments that it is a feature of many scheduled airline (Button, 2003) and shipping (Pirrong, 1992) markets.

2.4. *Other market failures*

The categories of market failure outlined above are not exhaustive, nor are they fully refined, and some comment is now offered on a number of other potential failures. Some of these are often cited as reasons for government interventions in markets, but they are less common than is sometimes argued by policy-makers and generally less severe in their intensity. Others have more of a normative connotation associated with them, and are touched upon mainly for completeness.

Adam Smith pointed to the potential failure of markets to deliver sufficient investment in larger projects because of the problems that the private sector has in bringing together the finance, and because of the potentially heavy losses that a failed mega-project may have associated with it.[a] A refinement of this argument is that some transportation projects or activities involve significant uncertainty and that in the absence of the ability to estimate actuarial risk this implies the market will under-invest. In these situations the government may be seen as the only agency with the administrative ability to marshal the necessary resources and the resources to cover the actuarial risks involved. The difficulty is often one of deciding on who should bear risk – the private sector, in seeking the highest potential return, will try to transfer as much risk as possible to the public sector. In practice, many very large schemes, such as the Channel Tunnel between the UK

[a] The large absolute possible losses in a situation of asymmetric risk assessment involving high levels of risk adversity may be too great to tempt any existing potential investor to participate.

and France, have been privately financed in recent years. There are also moves in some countries to unbundle activities so that those that pose genuine investment finance challenges are separated from those where the private sector can adequately decide.

An oft-cited notion is that many aspects of transportation infrastructure have the features of public goods, and will therefore be under-supplied if left to the market. While it is true that public goods, because they are non-excludable, and thus technically difficult to price, and non-rival, and thus would have a zero scarcity rent anyway, would be under-supplied in a market context, it is difficult to think of many types of transportation infrastructure that fall into this category. Certainly, roads – perhaps most often referred to as public goods – hardly meet the criteria. It is easy to limit the use of roads, and roads demonstrably often suffer from congestion. As Coase points out in his seminal paper, historically even lighthouses around the UK managed to extract monies from passing vessels. As with many things, the failure is often a matter of degree. For example, it may be possible to exclude users of transportation infrastructure, but for institutional reasons it may be more efficient for the government to do this.

Another frequently mentioned market failure is that transportation is a merit good that should be allocated according to "need" rather than effective demand. Often this argument is supplemented by the idea that a basic level of accessibility or mobility should be assured to all. This, however, is a normative judgment rather than a question of economic efficiency. It is true that markets are based on the notion of effective demand and, therefore, allow those with the highest income to consume more transportation services if they are so inclined. This may be seen as unfair, or it may be considered socially desirable for everyone to enjoy a minimum level of mobility if they wish irrespective of income. These are subjective positions. A more positively oriented issue is the consideration of the most efficient way of meeting these normative ambitions. This may be through actions in the transportation field, although these inevitably have negative overall resource implications, or it may be through other measures such as direct income transfers that allow lower income groups the option of using additional money on transportation or any other goods or services. One of the government intervention failures discussed below involves "capture" of decision-making processes, and there is a danger that attempts to meet social objectives through actions in particular sectors, such as transportation, can be captured by interest groups that effectively act to lobby for their own cause under the claim of meeting merit needs.

3. Government intervention failures

The traditional view under the Anglo-Saxon philosophy was that governments intervened to rectify market failures and thus to serve the public interest. It is this school of thought that we continue to focus on here. The emphasis would be somewhat different if the Continental philosophy were taken as the basis of discussion. It does mean, however, that the focus can remain primarily on matters of economic efficiency and leave aside what in practice may prove more politically important concerns with the incidence of alternative actions on different groups.

The idea that governmental interventions are sometimes needed in transportation markets has led to numerous debates about whether governments, even with the public interest at heart, are capable of improving on the market (Button, 1992). At the more micro-level, there are concerns about which types of policy best serve the public interest for any type of market failure. More recently, the issue has been broadened to a questioning of the motivations that underlie interventions and whether they are always initiated with the aim of serving the public interest.

3.1. Information issues

Modern economies are complex, and transportation services are complex to provide. It is also a very dynamic sector, both in terms of the types of services it provides and the types of service that potential users are seeking. To handle this complexity and dynamism optimally requires an immense amount of information together with mechanisms for making the best use of this information.

One, and often the main, reason that markets are favored is because of a lack of adequate information for a centralized policy to be efficient. A perfect market is assumed to have full information of all demand and supply factors, as is a perfect government. When there is less than full information, the traditional economic argument is that markets are preferable to government intervention because risks of mistakes are spread across a number of actors, each of which will make different use of the information that is available. Government tends to be narrower in its approach and to adopt a single course of action that may fail badly. Essentially, the market approach in these conditions is seen as more risk-averse.

But there is also a more practical matter: the collection of information by government can be costly and takes so much time that it is simply not efficient. A classic case of this was in the 1960s and 1970s when the European Community (as it then was) sought to regulate international trucking rates within its territories by imposing a forked tariff regime (Button, 1984). This involved establishing a maximum rate for each carriage when the market was buoyant and quasi-monopoly power a potential issue, and a minimum rate when the market was depressed and excessive competition was seen as likely. Calculating the forks proved virtually impossible, as was adjusting them to meet particular circumstances. They were extremely difficult to enforce.

Linked to the information issue, government intervention normally requires that the things being regulated are quantifiable; without that condition, qualitative controls are often ineffectual. The problem in some cases is that the output attributes of many transportation services are difficult to capture in numerical data (e.g. externalities such as visual intrusion or service qualities such as comfort). As a result, more easily quantifiable input measures are often used as proxies for outputs and are the subject of government controls. The links are often tenuous and often pose particular problems over the longer term if transportation suppliers have an incentive to reduce the coupling between the output and the input proxy. For example, a tax on petroleum to reflect its long-term shadow depletion cost may initially lead to a reduction in the sale of petroleum but in the longer term, market adjustments may result in higher sales of vehicles with smaller engines that negate the intended goal of the tax.

3.2. Regulatory capture

Tied to information issues is the matter of who actually has the information in any regulated situation (Posner, 1974). In most cases the regulators have to rely heavily on the information provided by the regulated transportation firms and industries. This is particularly so in the case of rate-of-return regulations that, for example, were widely used to regulate transportation industries in the USA for half a century from the 1930s. The transport firms recorded their costs, which they were then allowed to pass on to their customers together with a mark up to provide an acceptable rate of return. Monitoring of these cost figures was generally difficult, but even in the absence of any explicit efforts at fraud the system hardly provided an incentive for maximizing X-efficiency. The airlines, trucking companies, etc., had minimal reason to engage in rigorous efforts at cost minimization when costs could readily be passed on in fares and rates.

But government interventions may also be captured in other ways, most notably in the ways in which regulations and controls are defined. The "capture" of the market intervention by transportation suppliers may also be of a much less direct kind. In this case it is not the transportation supplier that is the major beneficiary but rather the suppliers of services further up the value chain. This may entail monopoly rents being enjoyed by unionized labor, the costs of which are passed on to transportation users via the airlines, trucking companies, bus operators, etc. One example of this involved bus subsidies in the UK; studies by the Transport Research Laboratory revealed half of the subsidies went to labor working in the industry as wages above those in comparable jobs elsewhere in the economy. Such indirect capture can also be through the providers of infrastructure or hardware used by the transportation companies (aircraft, air traffic control, airports, etc., in the case of air transportation – Button, 2004b). This type of situation would seem more common when parts of the transportation supply chain are left to competitive forces, and are thus price takers, while other elements further up the chain are publicly owned or enjoy institutionalized monopoly power.

This type of capture of regulation can also take other forms. As Stigler (1971) and others have pointed out, those actually responsible for the market interventions may not do so for the public interest, or, even if the initial intervention did aim at serving the public, regulations and controls may be subsequently manipulated to meet the ends of the regulators. There is an issue of how the public interest is defined in this context. Following the Anglo-Saxon philosophy then, simple efficiency criteria – basically output is determined at the point where full marginal cost equals demand – are the objective. But those initiating regulations are often concerned with their own political aims such as re-election; they may manipulate transportation policies to ensure the favor of the median voter. It is difficult to disentangle such policies that often involve a complex combination of transportation subsidies, taxes, and regulations from an espoused social policy that, for example, is argued for public interest distributional reasons.

Not only is there potential for those legislating regulation to capture interventions for their own ends but administrators of regulation may also seek to manipulate interventions to meet their objectives (Tullock, 1965; Posner, 1974). The issue here is often that of a principal-agent problem. Even if the *de jure* government interventions are aimed at minimizing market failures, those responsible for the *de facto* implementation and monitoring of the law have flexibility in interpretation, and can influence administrative structures. The private motivation of bureaucracies is often not the same as that of the electorate or of the policy-makers.

Modern theories of the firm, even those describing the behavior of firms in the private sector, note that factors such as quality of offices, size of staff, nature of health and pension plans, job security, etc., are often very important to management. These sorts of fringe benefits, and emblems of position, take an even more important position in the public sector where salaries are generally lower. There may also be cases where the regulators have a vested interest in assuring their future after they cease to be regulators. Often these things are closely related to the size of the agency, and especially the number of employees, giving the public servant a personal incentive to maximize rather than minimize regulation. The bureaucracy can exercise power to meet its particular objectives, as can the regulated in industries in many cases, through its control over information flows. It is generally the bureaucracy that collects, collates, and feeds information and data to the policy-makers, and to the public more generally. They also often control a large part of public policy research budgets, which influences the direction of the research agenda.

This view of government intervention in transportation is relatively new but has gained a considerable acceptance in many countries. The notion that regulations can be captured is certainly not unique to transportation, and the reactions to it have transcended transportation policy. The changes have included the movement toward widespread liberalization of transportation markets since the late 1970s, the privatization of transportation-supplying industries – including in some cases infrastructure – and reforms to the types of regulations that have been retained. In the last case, interventions continue but in ways that are less information-intensive and less prone to capture. An example of this is the control of the monopoly power of some items of transportation infrastructure. The traditional rate-of-return regulation, with the need for detailed cost information, has often given way to "price capping," which can be operated with less data, and involves a much smaller bureaucracy. Technically, price capping is also more able to deal with issues of static and dynamic X-inefficiency that have often been a major problem. BAA, which owns the main London airports is, for example, the subject of price caps.

3.3. International coordination

The increasing concern about the dangers of global warming are closely tied to that of the externalities that exist in transportation markets – car users, for example, do not take due account of their carbon dioxide emissions when deciding on their trip making.[a] In that sense the issue is one of market failure. But there is another dimension to the topic, namely: Why have governments been slow to intervene and react to this, and some other essentially global problems? The problem may be seen in this context as a failure of governments to agree on a mechanism for property right allocations because of the capture of the system by their own populations. An example of this is that many states have failed to adopt the Kyoto Protocol. The underlying challenge is that there are failures in the institutional structure to handle such international issues. The problem is illustrated in Figure 4.

The figure depicts the marginal domestic benefits (MBD) associated with a single country imposing instruments to reduce emissions of global warming gases. Conventional notions of diminishing returns suggest that the benefits of increasing levels of abatement diminish as abatement takes place. For simplicity it is assumed that there are no critical levels. The benefit function ignores the benefits to other nations of this single state acting to reduce its own carbon dioxide

[a]Transportation is one of the largest contributors to the emissions of global warming gas emissions, and is the fastest growing source.

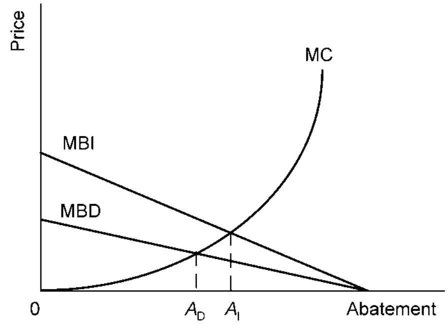


Figure 4. Government failures to internalize global effects.

emissions. If these benefits to other countries are added in, the resultant international marginal benefit (MBI) curve lies further out. The tendency is for domestic governments, however, to effectively become free riders in these circumstances. They may advocate global interventions to restrict the use of fossil fuels, but do not have any incentive to take their full share of responsibility for doing so. If the MC curve in the diagram is the marginal cost of abatement for the individual country, then without appropriate international cooperation the incentive for the single country is to abate to point A_D rather than the global optimum of A_I.

The problem here may be seen as one involving an inherent market failure, but in fact it is combined with an institutional structural failure that prevents effective governmental actions to overcome this. This type of situation generally arises in situations where there is no internationally agreed ownership of property rights such as the atmosphere or the oceans, and can lead to such things as oil spillage as well as excessive emissions of greenhouse gases.

4. Conclusions

Over the years, economists, lawyers, and policy-makers have expended considerable energies seeking out ways of deciding when transportation markets fail, and in designing structures that minimize these failures or at least minimize their adverse implications. More recently the resultant regulatory edifices of the past have come under scrutiny themselves, and questions have been asked about the effectiveness of market interventions and, in some cases, whether they have offered "cures" that have actually been worse than the diseases themselves. These debates continue.

One of the problems of making clear-cut decisions about the nature and scale of market and government failures in transport is that of complexity. Much of the discussion above has been couched in terms of partial equilibrium analysis, where it is implicitly assumed that only the market or government intervention is of interest while all other activities are assumed to meet standard Pareto optimality criteria – essentially marginal cost pricing. This is never the case with market failures, and second-best approaches are inevitable. Similarly, individual government interventions in markets are very much the exception rather than the rule, with policy normally involving the use of a portfolio of interacting measures. If one follows Tinbergen's dictate that there should be a specific instrument targeted at a specific objective then this may in itself be seen as a market failure, but at a more micro-level it often disguises the effects of individual policy tools.

References

Baumol, W.J., J.C. Panzar and R.D. Willig (1982) *Contestable markets and the theory of industry structure*. New York: Harcourt Brace Jovanovich.

Bonnafous, A. (1994) "Summary and conclusions," in: European Conference of Ministers of Transport, ed., *Internalising the social costs of transport*. Paris: ECMT.

Button, K.J. (1984) *Road haulage licensing and EC transport policy.* Aldershot: Gower.

Button, K.J. (1992) *Market and government failures in environmental management: the case of transport.* Paris: OECD.

Button, K.J. (2003) "Does the theory of the 'core' explain why airlines fail to cover their long-run costs of capital?" *Journal of Air Transport Management*, 9:5–14.

Button, K.J. (2004a) "Transport policy," in: A.M. El-Agraa, ed., *The European Union; history, institutions, economics and politics*, 7th edn. London: Prentice Hall.

Button, K.J.(2004b) *Towards an efficient European air transport system*. Aldershot: Ashgate.

Call, G. and T.E. Keeler (1984) "Airline deregulation, fares, and market behaviour: some empirical evidence," in: A.F. Daughety, ed., *Analytical studies in transport economics*. Cambridge: Cambridge University Press.

Coase, R. (1960) "The problem of social cost," *Journal of Law and Economics*, 3:1–44.

Dodgson, J., Y. Katsoulacous and R. Pryke (1991) *Predatory behaviour in aviation*. Luxembourg: EC.

Foster, C.D. (1992) *Privatization, public ownership and the regulation of natural monopoly.* Oxford: Blackwell.

Hensher, D.A. and K.J. Button, eds (2003) *Handbook of transport and the environment*. Oxford: Elsevier.

Kahn, A.E. (1988) *The economics of regulation: principles and institutions*. Cambridge: MIT Press.

Leibenstein, H. (1966) "Allocative efficiency vs X-efficiency," *American Economic Review,* 56: 392–415.

Levine, M.E. (1965) "Is regulation necessary? California air transportation and national regulatory policy," *Yale Law Journal*, 74:1416–47.

Lindsey, R. and E. Verhoef (2000) "Congestion modeling," in: D.A. Hensher and K.J. Button, *Handbook of transport modelling*. Oxford: Pergamon.

Moore, T.G. (1986) "US airline deregulation; its effects on passengers, capital and labor," *Journal of Law and Economics*, 29:1–28.

Morrison, S.A. and C. Winston (1987) "Empirical implications and tests of the contestability hypothesis," *Journal of Law and Economics*, 30:53–66.

Pirrong, S.C. (1992) "An application of core theory to the analysis of ocean shipping markets," *Journal of Law and Economics*, 35:131.

Posner, R. (1974) "Theories of economic regulation," *Bell Journal of Economics and Management Science,* 5:335–58.
Robbins, L. (1932) *The nature and significance of economic science*. London: Macmillan.
Stigler, G.J. (1971) "The theory of economic regulation," *Bell Journal of Economics and Management Science,* 11:3–21.
Telser, L.G. (1994) "The usefulness of core theory in economics," *Journal of Economic Perspectives*, 8: 51–64.
Train, K.E. (1991) *Optimal regulation: the economic theory of natural monopoly.* Cambridge: MIT Press.
Tullock, G. (1965) *The politics of bureaucracy.* Washington, DC: Public Affairs Press.

Chapter 3

REGULATORY TRANSITIONS

WILLIAM B. TYE
The Brattle Group, Cambridge, MA

1. Introduction

The change toward greater reliance on competition and less on economic regulation in the transportation industries created a special class of public policy problems associated with regulatory transitions (Tye, 1991). Chief among them was the transition from regulated private firms or state-owned enterprises to deregulated privately owned monopoly or competitive firms. While these "liberalization" developments are common around the world, the US railroad industry provides useful examples, and has inspired many of the conceptual developments in identifying and solving the resulting problems. Of course, one of the lessons from these experiences is that every transition has its own idiosyncrasies that must be taken into account. Nevertheless, the experience of the US rail industry can provide a guide to the types of problems, methods of analysis, and candidate solutions that are likely to emerge in any particular application to other modes of transportation and geographic circumstances. Different circumstances may require different policy responses.

2. Origins of the regulatory transition problem

2.1. The regulatory transition problem defined

A substantial amount of research has addressed the economics of regulation and the case for deregulation and privatization of state-owned monopoly (Friedlaender and Spady, 1981; Keeler, 1983; Winston et al., 1990; Gómez-Ibañez et al., 1999). However, there was little anticipation of the special problems that would arise during the transition to greater reliance upon competition.

However, the economic market structure and practices that emerged during the transition were not necessarily the same as those that would have arisen if regulation and state ownership had never occurred. Nor were they identical to the

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

long-run competitive equilibrium that would be achieved in the fullness of time under deregulation. While the problems caused by traditional economic regulation and state ownership may have abated, the original problems that regulation sought to solve in many cases did not simply go away. Furthermore, certain new problems were created by the transition itself.

2.2. Regulatory and antitrust issues in the transition to deregulation in the US rail industry

US railroads were recognized early on as a candidate for economic reforms because traditional comprehensive regulation simply was not working (Wilson, 1980; Joskow and Rose, 1989; Viscusi et al., 1995). Much of the literature focused on the belief that there was a "misallocation of freight" among the modes due to rate regulation and excessive capacity due to constraints on rail service abandonment. Also, an emerging body of economic literature showed that regulation itself was to blame for many of the financial problems of the industry. In addition, many economists questioned the assumption of large and ubiquitous scale economies in virtually all transportation modes and markets. This assumption had supported the notion that regulation was necessary because tendencies toward natural monopoly meant that competition would not work.

Comprehensive reform came in the USA with the Staggers Rail Act of 1980. The new Act contemplated a substantial dismantling of the regulatory institutions that had previously impeded market forces. The Staggers Rail Act addressed each of the most important issues posed by the regulatory transition. For our purposes, the most important reforms were the replacement of the publicly filed tariff system with confidential negotiated contracts (with disputes to be resolved in the courts, not before regulatory authorities); the elimination of maximum reasonable rate regulation except where the carrier exercised "market dominance"; and maximum feasible reliance on competition and market forces rather than regulatory fiat.

While regulatory reform in the USA involved simply relaxed regulation of private companies, elsewhere around the world it often also involved the difficult additional problem of privatization of public enterprise combined with the need for regulatory systems to guide the transition (Vickers and Yarrow, 1988; Armstrong et al., 1994). The transportation modes in the USA, however, have tended to go either to a more or less "cold-turkey" deregulation in circumstances where neither customers nor shippers have substantial sunk investments (i.e. investments that had little or no salvage value upon exit from a market) that commit them to prior competitive relationships (air and motor carriers), or a long-term period of transition (in the case of railroads). It is the latter cases that interest us here.

3. Issues in regulatory transitions

3.1 Importance of addressing explicitly the specific problems raised by regulatory transitions

The most obvious starting question is why we need a transition at all: why not just deregulate promptly for all the modes (Meyer and Tye, 1985, 1988)? First, the political compromises inherent in the deregulation legislation may have produced specific transitory regulatory mechanisms to deal with them. As a practical political matter, we may have no choice. Second, many old contracts, such as labor agreements, were hopelessly inconsistent with the new competitive regime but remained in effect. These legacies restrained the flexibility of buyers and sellers to respond to the new incentives of competition, often frustrated the achievement of the objectives of deregulation, and confused observers, who compared the actual immediate results of deregulation with the predictions of the long-run equilibrium in the economic literature. Third, sunk costs raised substantial equity and efficiency concerns. Where buyers of transportation services had incurred sunk costs that made them captive to only a few suppliers, these buyers might lose the regulatory protection they had relied upon against a "hold-up" (or "opportunistic behavior") when originally making their investment. The legacy of past sunk investments by buyers could create arbitrary differences in the benefits and costs of deregulation across customer market segments and motivate challenges to the process of deregulation. Conversely, investors in regulated firms may have sunk past investments under the prior regulatory regime with the expectation that they would be amortized with the aid of regulatory enforced pricing schemes and prohibitions against free entry. These mechanisms for cost recovery employed a "regulatory contract" rather than the long-term contracts with buyers that are typical in other industries (Goldberg, 1976). With an immediate switch to more competition, investors in incumbent firms might find that they are unable either to sign contracts with buyers or to price output in "spot markets" to recover their "stranded costs." In other situations, investors might receive substantial windfall gains when regulatory restraints on prices are lifted. Finally, botched deregulatory experiences in the US savings and loan industry and the energy industries have brought attention to the potential for very costly unsuccessful transitions.

In addition to the equity issues raised by the problem of sunk costs, there are efficiency problems in regulatory transitions as well. As a practical political matter, the efficiency gains from regulatory reform may be unrealizable unless the inequities are addressed. And bygones are not necessarily bygones when it comes to monopoly power created by sunk costs in the regulatory transition. A power plant suddenly held hostage to a sole rail provider could be required to pay monopoly prices for many years, with all the efficiency costs of monopolization, as a result of its prior reliance on regulated prices to protect it from monopoly power.

These difficulties point to a more general concern. Industries undergoing a transition to deregulation often do not have the contractual arrangements among buyers and sellers that are used to solve many of the problems previously resolved through the regulatory contract. Without some explicit regulatory intervention, the contracts that would be signed to fill this void would reflect the legacies of past regulatory policies, sunk costs, and contractual commitments. These new contracts would not reflect the conditions that historically would have existed in the absence of regulation or conditions that would exist after the legacies of regulation have expired. These difficulties create equity and efficiency consequences that could threaten a successful transition to deregulation.

To a large degree, threats to a successful transition can be traced to an improper conceptualization of the transition problem. For example, "spot markets," such as stock or commodity markets, became the implicit economic model for many regulatory policies, when in fact unregulated markets tend to address these issues with long-term contracts. Another illustration was the belief in an implicit model of industry structure where each firm is a stand-alone competitor despite the complex vertical (i.e. end-to-end) economic relationships and horizontal (i.e. joint) ventures among carriers that usually characterize the transportation industries. However, one of the common reasons that many of these transportation industries were regulated in the first place is that they involved complex networks and a high degree of cooperation among competitors to serve demand. As we shall see more fully below, these misconceptions led to misguided models of how to price services during the transition and how to address issues such as agreements among carriers to provide services jointly (the interconnection or competitive access problem).

3.2. Objectives for the regulatory transition

A successful transition to deregulation must develop policies and economic models during the transition to deregulation that maximize the success of the greatest ultimate reliance – where justified – upon competitive markets and traditional legal remedies in the new long-run equilibrium and conversely minimize the need for intervention during the transition via traditional regulatory methods. One must begin with a clear understanding of the existing industry structure, market practices, and regulations on the eve of deregulation, particularly the legacy of sunk costs and contracts that must be overcome in moving to greater reliance upon competition. Fashioning the transition means developing models of economic and political analysis to steer the industry from the prior regulatory system to competitive prices, outputs, and market mechanisms (including contracts). This steering should give due regard to regulatory and transaction costs and the need to maximize the chance of success.

Regulatory intervention in the transition to further this objective should not be considered backsliding from the goal of deregulation but a necessary cost of achieving the economic benefits that motivated the need for regulatory reform (Meyer and Tye, 1985).

3.3. Three insights into the structure of transition problems

Experience in posing and resolving these issues of regulatory transitions in the US rail industry points to the value of three key insights that have proved useful in understanding and solving these problems. The approach of this chapter will be first to identify these insights without elaborate explanation and then seek to illustrate their relevance in the discussion of specific problems that follows. The place to begin is the insight that public policy toward the special problems created by regulatory transitions cannot assume a clean slate and simply apply policies that would make sense in a regime that had never experienced the rules, incentives, and commitments of economic regulation. Many of these problems arise from the existence of a legacy of sunk costs from the prior regime of regulation and state-owned enterprise (Meyer and Tye, 1988).

The second insight is that many of the problems which regulation is designed to solve, such as idiosyncratic investments subject to opportunistic behavior by customers, suppliers, or joint venturers, are solved by contracts or vertical integration (i.e. common ownership of successive steps in the supply chain) in unregulated industries. Accordingly, the objective of the transition is to encourage customers and suppliers to sign the same kind of contracts that they would have signed had there been no sunk cost legacy of regulation (dubbed the "contractual equilibrium"), so that the regulatory transition will atrophy and regulatory institutions will wither away.

The third insight begins with the notion that, unless specifically justified to the contrary, transition regimes should be "competitively neutral," i.e. eliminate both the artificial benefits and costs of incumbency so that all firms can compete on a "level playing field." The concept of competitive neutrality is particularly useful in resolving the issues of regulatory transitions for networks requiring interconnection of service among competitors (e.g. where two railroads seek to interchange joint-line traffic). Here, we must distinguish two desirable dimensions of competitive neutrality. The first is the property of an interconnection scheme that the least-cost competitor (measured in terms of incremental costs going forward, irrespective of sunk costs) will always find it possible in a "winner take all" competition (i.e. all market share goes to the firm offering the lowest price) to underbid profitably the lowest possible price of a less efficient competitor. This property of achieving static economic efficiency, labeled "weak competitive neutrality," is obviously a desirable minimum condition in a transition.

Fortunately, our third insight is that weak competitive neutrality is achieved by a wide variety of interconnection pricing schemes. Rather, the choice among alternative interconnection schemes must be based on more discriminating criteria, which comprise "strong competitive neutrality" (which among other conditions requires that equally efficient competitors have an equal opportunity to recover their sunk costs). The distinction between types of competitive neutrality can be used to resolve many of the disputes over important issues in the transition, such as pricing of access to network facilities and recovery of sunk costs incurred in the prior regulatory regime. As explained more fully below, the distinction also reveals that many of the competitive regimes proposed and implemented in regulatory transitions involving network industries are driven by the erroneous belief that these regimes uniquely achieve static economic efficiency, while they ironically nevertheless fail to achieve the more important tests for strong competitive neutrality.

4. Ramsey pricing and maximum reasonable rates

Many rail customers have sunk costs in idiosyncratic investments that leave them "captive" to a single carrier, and demand protection against a "hold-up" in the regulatory transition as a condition to supporting deregulation. Conversely, in other cases carriers may have sunk costs in specialized investments that left have them dependent upon regulated tariffs for the recovery of their costs. The "contractual equilibrium" as described above looks to the pricing results from long-term contracts (negotiated without the influence of sunk costs on either side) as the benchmark for whether captive shippers or carriers were being "held up" during the transition. The power of these ideas to resolve the problem of shipper captivity during the transition may be best seen by comparing them with the regimes of price discrimination endorsed by some regulatory authorities and many economists. This extreme price discrimination bears little resemblance to the contractual model of a competitive regime (where no firm would ever sink costs under such conditions without contractual protection from such a "hold-up") (Klein et al., 1978).

Many economists (most forcefully represented by Baumol, 1979; Baumol and Willig, 1983; Bailey and Baumol, 1984; Kessides and Willig, 1998) have nevertheless urged that regulatory institutions in the transition permit incumbent firms to exploit the sunk costs of captive customers incurred in the prior regulatory regime via Ramsey pricing (or, more crudely, "charging what the market will bear") in order to solve the overhang of excess costs incurred as a result of the incentives created by the prior regulatory regime. The model of the transition that relies on price discontinuation views temporary captivity of shippers during the transition as a blessing to be exploited, not an obstacle to successful deregulation that must be mitigated. However, such a model of transition encourages firms to seek additional elements of monopoly power through bottlenecks (i.e. monopolized links in the network) and mergers and creates a vested interest in preserving the transition and delaying the day of the onset of true competition (after the costs incurred in the prior regime have been amortized). It also turns out that in practice Ramsey pricing often incorporates the worst of both worlds: statistically insignificant differences in demand elasticities drive large shifts in rates, but large errors in measuring demand elasticity deprive those rate differences of their justification in terms of efficiency gains as measured in consumers' surplus (Tye and Leonard, 1983; Gómez-Ibañez, 1999). Interestingly, the US Interstate Commerce Commission (1985a) (and its successor agency, the Surface Transportation Board) considered and rejected the application of Ramsey pricing to the US rail industry because it was administratively infeasible (due to inadequate data).

In addition to the wisdom and feasibility of applying Ramsey pricing to facilitate the recovery of the overhang of excess costs created by the prior regulatory regime, there remains the issue of the role of Ramsey pricing in a regime of unregulated competition after the transition. Proponents usually endorse such price discrimination as an efficient means of pricing in excess of incremental costs for firms enjoying increasing returns to scale (conditions of decreasing unit costs). Keeler's (1983) findings were that US railroads do not, implying that pricing at incremental cost would be sufficient to recover total costs, thereby removing the need for price discrimination to achieve revenue adequacy. Of course, other transportation industries undergoing a transition to deregulation might represent different circumstances and thus properly raise the issue.

5. Competition policy in the transition to deregulation

The retreat of regulatory institutions and state-owned monopoly has usually been accompanied by increased enforcement of traditional antitrust laws to ensure successful emergence of competition. In addition to the competitive concerns raised by the issue of competitive access, the transportation industries have posed a number of unusual problems of maintaining competition that arise because transportation firms operate in networks (Carlton and Klamer, 1983). Because they operate in networks, carriers often must cooperate as well as compete (Telser, 1987). Under the prior regulatory regime, that cooperation was often actually encouraged by regulatory policy and explicit antitrust immunity. Policy-makers often assumed implicitly that the efficiency benefits of cooperation under the prior regime would be preserved, while the benefits of head-to-head competition would nevertheless also be achieved. In the alternative, some antitrust enforcement agencies, accustomed to dealing with stand-alone firms requiring no coordination of service, simply did not understand the unique problems created by structuring a regime of both competition and cooperation and insisted on the complete elimination of even efficient cooperation.

A number of controversial joint activities by carriers will continue to raise similar issues, e.g. antitrust immunity for airline alliances and code-sharing agreements, joint operation of computerized airline reservation systems, frequent-flyer programs, ticket agency arrangements, and other joint ventures such as terminal construction and operation. The transition to deregulation and the loss or curbing of antitrust immunity precipitated numerous antitrust investigations and proceedings in each of the transportation modes, including major cases on alleged collusion and price fixing. Given the potential efficiency gains from cooperation in network industries, transportation carriers will continue to pose difficult problems to antitrust authorities of balancing competition with cooperation.

Merger policy has also been particularly difficult during the transition. Regulation was blamed for rail carriers inheriting uneconomic route structures and operating authority. These carriers also operated under merger approval procedures that allowed for a more permissive set of standards for gaining governmental approvals than for firms outside the transport sector. As a result, a wave of mergers was routinely approved by the US government authorities on the grounds that they improved uneconomic route structures and service inherited from the prior regulatory regime.

Chapter 27 of this volume address in more detail concerns that have been raised about the wisdom of these trends. To begin with, mergers in some modes were tending to eliminate the very competition that was the rationale for deregulation. Furthermore, setbacks among merged US rail carriers have raised serious questions about the validity of claims of substantial efficiency gains from mergers during the transition (Ingersoll, 1998). Even more serious, these setbacks pose a challenge to the entire process of deregulation and have called into question the permissive regulatory approach toward what has been labeled the "competitive access" problem, to which we now turn.

6. Pricing competitive access in the transition to deregulation

6.1 Significance of the access issue

The notion that the problems of the regulatory transition originate in the sunk costs from the prior regulatory regime can also help resolve "competitive access" issues. Problems of competitive access arise when either previous incumbents or entrants seek access to the "bottleneck facilities" of incumbent firms, interconnection is necessary for efficient supply in a network industry, and all links in the network do not achieve effective competition simultaneously during the transition (Armstrong et al., 1996; Laffont and Tirole, 1996; Larson and Lehman, 1997; Gabel and Weiman, 1998; Laffont et al., 1998). Solution of the competitive access problem may be critical to the success of the regulatory transitions because it is often a precondition to a regime of successful competition in network industries.

We begin our analysis of the competitive access problem with a puzzle (Grimm and Harris, 1998). Common prescriptions in the telecommunications, pipeline, and electric utility industries call for mandated "unbundling" of service elements to open up access of former monopolies to competition in the transition to deregulation and requirements to interconnect with other carriers, including competitors. In contrast, "voluntary negotiations" have been proposed and implemented as the solution for the US rail industry (McFarland, 1987; Reiffen, 1988; Kleit, 1993; Cunningham and Jenkins, 1997; but also see Grimm and Harris, 1983; Kaplow, 1985; Grimm et al., 1992) based in part on claims that the so-called "efficient component pricing rule," a.k.a. the "parity principle," uniquely achieves "static" (i.e. short-run productive) economic efficiency, and would be forthcoming in a regime of voluntary negotiations. The parity principle, as explained more fully below, indemnifies the incumbent monopolist from any risk of lost profit as a result of competition. Is there something unique about railroads that justifies exactly opposite treatment than for other network industries?

At first it might be supposed that railroads are different because their long-term decline created a severe revenue deficiency on the eve of regulatory reform. However, this cannot explain the difference in regulatory policy toward competitive access, because electric utilities, pipelines, and telecommunications carriers have faced or will face hundreds of billions of dollars in "stranded costs" (Holden, 1995).

The other possible explanation is that US rail regulators have accepted an economic paradigm that purports to show that regulatory policies to promote pro-competitive access policies will necessarily lead to static economic inefficiency (i.e. will fail to achieve weak competitive neutrality). However, the third economic insight identified above is that static economic efficiency is an extremely robust property of competitive access schemes. This insight is contrary to current US regulatory policy and to the views of prevailing economic theories and models, and permits a deeper understanding of the proposed solutions to the competitive access pricing problem.

6.2. The competitive access problem defined

Figure 1 illustrates the problem of rail competitive access (also known as the "rat tail" problem). Sponsors of this hypothetical problem (e.g. Baumol and Sidak, 1994, 1995) only look at the consequences of the competitive access regime for achieving static economic efficiency, in this case routing traffic to achieve minimal incremental costs over the competing routes (thereby ignoring sunk cost recovery), and thus we do so initially to conform to the spirit of the hypothetical problem. It involves a railroad company that owns the bottleneck portion of a route and that also participates in the competitive portion of a connecting route. The connecting carrier may be thought of as a new entrant or a firm that sank costs under the access rules and revenue divisions of the prior regulatory regime.

In the hypothetical problem, the entire route from X to Z is served by only one railroad (the "bottleneck"). Two railroads serve from Y to Z. In the prior regulatory regime, tariffs were "open routed" in the sense that an unaffiliated connecting carrier (serving only Y to Z) could propose a "through route" under a "joint rate" via the junction at Y and agreed-upon "divisions" of the revenue. The single-line route could then compete with the through route because the shipper could specify either routing. After regulatory reform, the connecting carrier is also allowed to rebate its share of the revenue division to encourage diversion to the joint-line route, and the single-line carrier can also compete on the basis of price.

Complaints of "vertical foreclosure" of this competition would arise if, during the transition to deregulation, the bottleneck carrier cancelled the joint rate applicable to the interchange at Y with a competing connecting carrier in order to favor its own single-line route from Y to Z (and they used their new rate-making freedom to do so on a widespread basis). New opportunities for such foreclosures also arise when a bottleneck carrier merges with one of the two carriers competing over the Y to Z service (and these were frequent under the new permissive merger standards). The non-merging carrier would lose a "friendly connection" at Y. It would then request "protective conditions" to the proposed merger to prevent post-merger cancellation of the joint rate and through service.

Proponents of the parity principle (or the efficient component pricing rule, ECPR) argued that the single-line rail carrier should be allowed freely to cancel the joint route and engage in "voluntary negotiations." It would then establish interconnection prices for its portion of the through rate that incorporated "opportunity cost," i.e. profit lost when the service moved over the joint rate (Baumol and Sidak, 1994). The logic was that if the price for the bottleneck portion were set below the parity principle price (even if that regulated price was the competitive price), a connecting carrier might be able to turn a profit even though its incremental costs were higher than those of the bottleneck carrier.

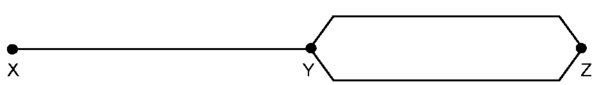


Figure 1. The rat tail problem in the US rail industry.

Box 1
Weak competitive neutrality

> Bottleneck and connecting carriers should compete for traffic on the basis of their competitive advantages or disadvantages arising from static efficiency differences (real resource costs going forward) in a winner-take-all competition. Violations of weak competitive neutrality result in either "uneconomic bypass" (successful competition by less efficient firms) or "foreclosure" (incumbents successfully impede effective competition by more efficient firms seeking access).

Ergo, it is argued, prices lower than the parity principle produce inefficiency (Baumol et al., 1997). The fact that the ECPR necessarily indemnified the monopoly carrier from any loss of profit resulting from competition, by including lost profit in the access price, was treated as an unfortunate consequence of the rule's alleged unique ability to achieve economic efficiency.

Examples such as these were used to infer that regulatory intervention, even to set a competitive price of access instead of the monopoly price, necessarily leads to a violation of weak competitive neutrality (Box 1).

However, concern over violation of weak competitive neutrality in access disputes is misplaced. In fact our third key insight identified above tells us that almost any arbitrary price of access will achieve static economic efficiency (i.e. efficient routing choices).

Tye and Lapuerta (1996) provide a mathematical proof of the third key insight (along with a possibly important exception). The mathematics are somewhat complex, but the logic of the proof relies on a simple proposition. Whereas ECPR starts with opportunity cost and derives a specific price for interconnection, an arbitrary alternative rule starts with an arbitrary access price and "imputes" it as an opportunity cost to the bottleneck owner. The same efficiency results prevail in either case. If the owner of the bottleneck treats the arbitrary price of access as an "imputed cost" (or opportunity cost) when it competes on the basis of price to provide the total service itself, its total incremental cost will go up by the arbitrary access price – just like the connecting carrier. The playing field will then be leveled and both carriers will be competing on the basis of the difference in real resource costs from Y to Z. Arbitrary access pricing works to achieve weak competitive neutrality just as good as ECPR does because price competition by the bottleneck carrier causes a range of arbitrary access prices to become opportunity costs of access, just like ECPR does for one particular value (the one that preserves the incumbent's profit) (Armstrong, 2001).

Obviously, the goal here is not to espouse arbitrary pricing of access. Rather, the point is that US public policy in the rail industry has been based on the idea that static economic efficiency was the chief economic goal of access pricing, and that this could be accomplished only by a system of voluntary negotiations leading to the parity principle. This misconception has led to endorsements of access

Box 2
Strong competitive neutrality

> Ownership of the bottleneck facility should be neither a competitive advantage nor a disadvantage as far as bottleneck owners and connecting carriers are concerned. Equally efficient bottleneck owners and connecting carriers going forward should have the same expectation of recovering their total costs. Strong competitive neutrality arises when an interconnection regime allows equally efficient competitors an equal opportunity to expect to earn their cost of capital, compensates the incumbent for any unavoidable and prudent legacies of regulation, and purges the interconnection price of monopoly rents (if any). Strong competitive neutrality includes weak neutrality, but not vice versa.

pricing schemes that, despite their severe drawbacks in ignoring the three key concepts identified above, were wrongly believed uniquely to satisfy static efficiency concerns (Interstate Commerce Commission, 1985b; Surface Transportation Board, 1996a,b, 1998). Secondly, the proof of the robustness of weak competitive neutrality suggests the need for a richer and more subtle test than mere static economic efficiency when choosing among alternative competitive regimes in the transition. This requirement leads to the notion of strong competitive neutrality (Box 2).

6.3. Economic logic underlying alternative policies on competitive access

Static efficiency properties are not unique to any particular access pricing rule, and are easily satisfied by alternative competitive access schemes. Whether or not a competitive access scheme satisfies weak competitive neutrality tells us almost nothing about its desirability (Tye and Lapuerta, 1996). The problem with Figure 1 as an alleged example of inefficient competition even at a competitive price of access lies in an assumption that is inconsistent with competitive markets. Proofs of the alleged efficiency of the ECPR implicitly assume that the bottleneck owner does not compete on price. In the more general case where price competition prevails, the parity principle does not uniquely achieve economic efficiency. It does, however, permanently protect the incumbent bottleneck monopolist from any adverse financial consequences arising from the introduction of competition via the access scheme.

There are additional problems with the parity principle as an access pricing mechanism in the presence of sunk costs, which are considerable in the rail industry. Because the only revenue available to the connecting carrier is the bottleneck owner's avoided incremental cost from Y to Z, it imposes a "price squeeze" on the carrier seeking access. Even an equally efficient connecting carrier (i.e. one with the same incremental costs) cannot recover total costs, including sunk costs. No connecting carrier would voluntarily sink investments if service depended on access dispensed under the parity principle, because the ability to recover total costs would depend on which carrier arbitrarily owned the bottleneck, not competition on the merits. The parity principle thus violates all of the above three key insights – first by ignoring sunk costs, second by imposing a regime to which no entrant would voluntarily agree prior to sinking costs (violating the contractual equilibrium), and finally by making it impossible for an equally or more efficient connecting carrier to recover total costs (violating strong competitive neutrality) (New Zealand Ministry of Commerce, 1995). In recognition of these difficulties, the New Zealand Telecommunications Act of 2001 banned application of the parity principle (there labeled the "Baumol–Willig rule") as a matter of law in that country.

6.4. Implications of weak versus strong competitive neutrality to the access pricing problem

There are good reasons for choosing among the alternative interconnection regimes, but achieving static economic efficiency (weak competitive neutrality) will not usually be one of them. The static efficiency issue, which has heretofore been the centerpiece of the debate, should simply be put aside. The choice among these pricing schemes therefore turns on assumptions about what other objectives should be accomplished in addition to static economic efficiency. Instead we must focus our attention on the hidden assumptions in the debate. Should US rail policy on competitive access continue to require that bottleneck carriers be indemnified from the financial consequences of competition (i.e. should the debate over competition start with a requirement that the bottleneck carrier not engage in price competition)? This revisits the oldest debate in the economics of the rail industry: is there an inherent conflict between intramodal competition and revenue adequacy, i.e. is rail competition inherently "cut-throat" and "destructive"?

Consideration of strong competitive neutrality offers additional objectives, such as the ability to recover sunk costs on equal terms, a consideration that its proponents believe will more likely lead to dynamic (i.e. long-run) efficiency gains. The parity principle supplements the objective of static efficiency implicitly with the assumption that the incumbent monopolist should be insulated from price competition and any profit erosion as a result of the introduction of competition. Critics of the parity principle view such a competitive asymmetry as a strange, permanent constraint to impose on a pricing regime designed to achieve a transition to deregulation and effective competition. However, some proponents of the rule often view it as vital to the recovery of stranded costs in the transition (Sidak and Spulber, 1996, 1997), an issue to which we now turn.

7. Revenue adequacy and stranded costs

Because of the legacy of regulation, incumbents often bear the burden of sunk costs that cannot be recovered because of the onset of competition during the transition ("stranded costs"). This has precipitated a debate over whether and how much of these costs should be recovered from customers and access charges to new entrants and by what mechanisms (Gordon, 1981). Arguments for and against the recovery of these costs are usually predicated on conflicting assertions regarding whether recovery contributes to or undermines static economic efficiency. Much of this discussion commits the same error of unduly focusing on static economic efficiency that arose above in the discussion of competitive access charges. Our previous analysis of competitive neutrality again resolves the issue: virtually any amount and method of recovering stranded cost will achieve static economic efficiency (Tye and Graves, 1996). These results hold because the stranded cost recovery mechanism can be restated as an access charge, and static economic efficiency is a robust property of access schemes. Again the issue really turns on whether dynamic efficiency objectives are hindered or hurt by the asymmetric ability of firms to recover total costs, the very same issue that ultimately must be answered in the access pricing debate. But here equity issues also play a role, because stranded costs by definition are permanently sunk, yet were prudently incurred in the prior regulatory regime.

Those arguing for competitive neutrality in the stranded cost recovery debate propose that the mechanism, combined with the obligation of the incumbent to amortize the sunk cost legacy from the past regulatory regime with the revenues provided by the recovery mechanism, should in combination be neither a competitive advantage nor disadvantage as far as incumbents and entrants are concerned. Equally efficient incumbents and entrants going forward should have the same (*ex ante*) expectation of earning their cost of capital going forward. Symmetry means only that both incumbents and entrants should be free to compete on the basis of their competitive advantages or disadvantages arising from true efficiency differences (real resource costs) going forward, a standard that is readily satisfied for almost any amount or method of recovery.

Intuitively, we may thus propose the following "smell test" for the competitive neutrality of a stranded cost recovery mechanism. Given the combination of the burden of amortizing the sunk costs of the past regulatory regime, the revenues from the offsetting stranded cost recovery mechanism for the incumbent, and no competitive advantages or disadvantages arising from other sources (i.e. assume "all else equal"), would an investor prefer to be the entrant or the incumbent? If the stranded cost recovery mechanism produces indifference to this choice, it has achieved symmetry as to revenue recovery, insofar as the recovery mechanism is incurred. All competitors would bear the full competitive advantages or disadvantages arising solely from differences in non-stranded costs, i.e. true efficiency differences (Tye and Graves, 1996).

According to this logic, allowing incumbents a fair opportunity for recovery of stranded costs during a transition to deregulation is not an impediment to competition on equal terms. By definition, a fair opportunity for recovery of stranded costs creates competition on equal terms (and conversely) because it imposes on all competitors the same degree of exposure to the legacy of sunk costs of the past regulatory regime. If done correctly, recovery of stranded costs achieves competition on equal terms because it removes an artificial financial and competitive advantage arising (in the absence of the stranded cost recovery mechanism) solely because entrants do not bear the sunk cost of the past regulatory regime and incumbents do.

But why should it matter that we preserve competitive neutrality with regard to the recovery of stranded costs in the transition? Critics of stranded costs recovery often claim that such costs are sunk and therefore need not be recovered to achieve efficiency on a going forward basis. The answer ultimately comes down to the equity and efficiency rationale for a transition mechanism. Should the transition preserve a fair opportunity for investors to expect to earn their cost of capital on prudent investments made in the prior regulatory regime? The economic rationale for maintaining that expectation in the transition is no different than the rationale for creating that expectation under regulation in the first place. Incumbent firms as a practical matter will continue to be the monopoly suppliers of certain services in the transition, so their financial stability can matter. As precedent, regulatory agencies such as the US Federal Energy Regulatory Commission have developed numerous rate methodologies to prevent arbitrary windfalls or losses during regulatory transitions (Kolbe et al.,1984; Myers et al., 1985).

Sometimes critics of stranded cost recovery also claim that investors need not be afforded such an expectation during the transition because they have automatically been previously compensated for the risks of stranded assets when the allowed rate of return is set equal to a market-based cost of capital in the prior regulatory regime (Stelzer, 1994a,b). As it turns out, the asymmetric risk of stranded costs implies that even if investors are fully cognizant of the risks, capital market prices fully reflected such risks, and regulators always set the allowed rate of return equal to the true cost of capital, it is mathematically impossible for investors to have been previously compensated automatically for these risks (Train, 1991; Kolbe and Tye, 1995, 1996).

Viewing the regulatory transition as arising from the legacy of sunk costs also contributes to understanding the "captive shipper" problem. In the rail industry, shippers were accorded regulatory protections in the transition only if they were deemed to be served by "market-dominant" railroads. Our first two insights suggest that the shipper's market dominance problem is simply the flip side of the carrier's stranded cost problem. Sunk costs from the prior regulatory regime are the key to identifying captive shippers and the need for residual regulation during the transition. The "contractual equilibrium" suggests that residual rate regulation for such captive shippers should look to long-term contracts signed by shippers and carriers prior to sinking costs for rate-making guidelines. Regulatory institutions should be designed to ensure that such designations expire in a timely manner, so that regulation does not yield to the incentives of some participants to preserve the competitive advantages conferred by these designations.

8. Conclusion

Successful regulatory transitions are critical to successful deregulation because they must address and solve the unique problems that arise because regulatory institutions previously replaced market mechanisms. The approach here emphasizes the application of three fundamental insights to the problem. Regulatory transitions must contend with the legacy of costs sunk by carriers and customers in the prior regulatory regime and not proceed as if there were a clean slate. Secondly, the "contractual equilibrium" is achieved when contracts supplant regulatory institutions. Regulatory intervention in the transition should encourage the signing of contracts similar to those that would have been in effect had the distorting effects of the legacy of sunk costs from the prior regime never occurred. As these contracts are signed, the regulatory institutions of the transition should wither away, leaving a marketplace undistorted by the legacies of regulation. Until then, a common regulatory problem during the transition is to define a set of residual, self-terminating, economic constraints that will satisfy political demands for equity – and other considerations created by the short- to medium-term continuation of sunk costs – without creating insurmountable obstacles to approaching a market-determined outcome in the long run.

The fact that transportation carriers operate in networks will continue to pose difficult antitrust issues of balancing competition and cooperation. The problem of competitors seeking access or interconnection to the facilities of incumbents is a rich opportunity to apply these principles. There, our third key concept demonstrates the robustness of static economic efficiency and reveals the significance of assumptions about who ought to get the net revenues from providing joint service. Recognizing the distinction between weak and strong competitive neutrality also provides insights into such issues as the recovery of stranded costs.

References

Armstrong, M. (2001) "The theory of access pricing and interconnection," in: M. Cave, S. Majundar and I. Vogelsang, eds, *Handbook of telecommunications economics*. Amsterdam: North Holland.

Armstrong, M., S. Cowan and J. Vickers (1994) *Regulatory reform: economic analysis and British experience*. Cambridge: MIT Press.
Armstrong, M., C. Doyle and J. Vickers (1996) "The access pricing problem: a synthesis," *Journal of Industrial Economics*, 44:131–150.
Bailey, E.E. and W.J. Baumol (1984) "Deregulation and the theory of contestable markets," *Yale Journal on Regulation*, 1:111–137.
Baumol, W.J. (1979) "Minimum and maximum pricing principles for residual regulation," *Eastern Economic Journal*, 5:235–248.
Baumol, W.J. and J.G. Sidak (1994) "The pricing of inputs sold to competitors," *Yale Journal on Regulation*, 11:171.
Baumol, W.J. and J.G. Sidak (1995) "The pricing of inputs sold to competitors: rejoinder and epilogue," *Yale Journal on Regulation*, 12:179–185.
Baumol, W.J. and R.D. Willig (1983) "Pricing issues in the deregulation of railroad rates," in: J. Finsinger, ed., *Economic analysis of regulated markets*. New York: St Martin's Press.
Baumol, W.J., J.A. Ordover and R.D. Willig (1997) "Parity pricing and its critics: a necessary condition for efficiency in the provision of bottleneck services to competitors," *Yale Journal on Regulation*, 14:145.
Carlton, D.W. and J.M. Klamer (1983) "The need for coordination among firms, with special reference to network industries," *University of Chicago Law Review*, 50:446–465.
Cunningham, P.A. and R.M. Jenkins, III (1997) "Railing at open access," *Regulation*, 20:42–47.
Friedlaender, A.F. and R.H. Spady (1981) *Freight transport regulation*. Cambridge: MIT Press.
Gabel, D. and D.F. Weiman (1998) *Opening networks to competition: the regulation and pricing of access*. Boston: Kluwer.
Goldberg, V.P. (1976) "Regulation and administered contracts," *Bell Journal of Economics*, 7:426–48.
Gómez-Ibañez, J.A. (1999) "Pricing," in: J.A. Gómez-Ibañez, W.B. Tye and C. Winston, eds, *Essays in transportation economics and policy: a handbook in honor of John R. Meyer.* Washington, DC: Brookings Institution Press.
Gómez-Ibañez, J.A., W.B. Tye and C. Winston, eds (1999) *Essays in transportation economics and policy: a handbook in honor of John R. Meyer.* Washington, DC: Brookings Institution Press.
Gordon, K. (1981) "Deregulation, rights and the compensation of losers," in: K.D. Boyer and W.G. Shepherd, eds, *Economic regulation: essays in honor of James R. Nelson*. East Lansing: Michigan State University.
Grimm, C.M. and R.G. Harris (1983) "Vertical foreclosure in the rail freight industry: economic analysis and policy prescriptions," *ICC Practitioners' Journal*, 50:508–531.
Grimm, C.M. and R.G. Harris (1998) "Competition access policies in the rail freight industry, with comparisons to telecommunications," in: D. Gabel and D.F. Weiman, eds, *Opening networks to competition: the regulation and pricing of access*. Boston: Kluwer.
Grimm, C.M., C.M. Winston and C.A Evans (1992) "Foreclosure of railroad markets: a test of Chicago leverage theory," *Journal of Law and Economics*, 35:295–310.
Holden, B.A. (1995) "Shift to deregulation may cost electricity industry $135 billion," *Wall Street Journal*, Aug. 7:B4.
Ingersoll, B. (1998) "Board appears ready to seek rail remedies," *Wall Street Journal*, April 20:A20.
Interstate Commerce Commission (1985a) *Coal Rate Guidelines, Nationwide* 1 I.C.C. 2d 520, aff'd sub nom. *Consolidated Rail Corp.* v. *United States*, 812 F.2d 1444 (3rd Circular 1987).
Interstate Commerce Commission (1985b) *Intramodal Rail Competition* 1 I.C.C. 2d 822, aff'd sub nom. *Baltimore Gas & Electric* v. *United States*, 817 F.2d 108 (DC Circular 1987).
Joskow, P.L. and N.L. Rose (1989) "The effects of economic regulation," in: R. Schmalensee and R.D. Willig, eds, *Handbook of industrial organization*, Vol. II. Amsterdam: North-Holland.
Kaplow, L. (1985) "Extension of monopoly power through leverage," *Columbia Law Review*, 85:515–556.
Keeler, T.E. (1983) *Railroads, freight and public policy*. Washington, DC: Brookings Institution.
Kessides, I.N. and R.D. Willig (1998) "Restructuring regulation of the rail industry for the public interest," in: *Railways: structure, regulation and competition policy.* Paris: OECD.
Klein, B., R.G. Crawford and A. Alchian (1978) "Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics*, 21:297–326.
Kleit, A.N. (1993) "Problems come before solutions: comment on pricing market access for regulated firms," *Logistics and Transportation Review*, 29:69–74.

Kolbe, A.L. and W.B. Tye (1995) "It ain't in there: the cost of capital does not compensate for stranded-cost risk," *Public Utilities Fortnightly*, 133:26–28.
Kolbe, A.L. and W.B. Tye (1996) "Compensation for the risk of stranded costs," *Energy Policy*, 24:1025–1050.
Kolbe, A.L., W.B. Tye and M.A. Baker (1984) "Conditions for investor and customer indifference among regulatory treatments of deferred income taxes," *RAND Journal of Economics*, 15:434–446.
Laffont, J.-J. and J. Tirole (1996) "Creating competition through interconnection: theory and practice," *Journal of Regulatory Economics*, 10:227–256.
Laffont, J.-J., P. Rey and J. Tirole (1998) "Network competition: I. Overview and nondiscriminatory pricing and II. Price discrimination," *RAND Journal Economics*, 29:1–56.
Larson, A.C. and D.E. Lehman (1997) "Essentiality, efficiency, and the efficient component pricing rule," *Journal of Regulatory Economics*, 12:71–80.
McFarland, H. (1987) "The economics of vertical restraints and relationships between connecting railroads," *Logistics and Transportation Review*, 23:207–222.
Meyer, J.R. and W.B. Tye (1985) "The regulatory transition," *American Economics Review*, 75:46–51.
Meyer, J.R. and W.B. Tye (1988) "Toward achieving workable competition in industries undergoing a transition to deregulation: a contractual equilibrium approach," *Yale Journal on Regulation*, 5:275–276.
Myers, S.C., A.L. Kolbe and W.B. Tye (1985) "Inflation and rate of return regulation," *Research in Transportation Economics*, 2:83–119.
New Zealand Ministry of Commerce and Treasury (1995) *Regulation of access to vertically-integrated natural monopolies*. Wellington: New Zealand Ministry of Commerce and Treasury.
Reiffen, D. (1988) "Partial ownership and foreclosure: an empirical analysis," *Journal of Regulatory Economics*, 13:227–244.
Stelzer, I.M. (1994a) "Stranded investment: who pays the bill?" In: remarks at the Southwestern Electric Exchange, American Enterprise Institute, March 30.
Stelzer, I.M. (1994b) "Restructuring the electric utility industry: further tentative thoughts," *Electricity Journal*, 7:36–41.
Sidak, J.G. and D.F. Spulber (1996) "Deregulatory takings and breach of the regulatory contract," *New York University Law Review*, 71:978–981.
Sidak, J.G. and D.F. Spulber (1997) *Deregulatory takings and the regulatory contract: the competitive transformation of network industries in the United States*. Cambridge: Cambridge University Press.
Surface Transportation Board (1996a) Statement of economists before the Surface Transportation Board, Docket No. 41242, et al., *Central Power & Light Company* v. *Southern Pacific Transportation Company*, October 21.
Surface Transportation Board (1996b) Decision in *Central Power and Light Co.* v. *Southern Pacific Transportation Company,* Docket No. 41242, et al., served December 31.
Surface Transportation Board (1998) Ex parte *No. 575, review of rail access and competition issues*, February 20.
Telser, L.G. (1987) *A theory of efficient cooperation and competition*. Cambridge: Cambridge University Press.
Train, K.E. (1991) *Optimal regulation: the economic theory of natural monopoly*. Cambridge: MIT Press.
Tye, W.B. (1991) *The transition to deregulation: developing economic standards for public policies*. New York: Quorum Books.
Tye, W.B. and F.C. Graves (1996) "Stranded cost recovery and competition on equal terms," *Electricity Journal*, 9:61–70.
Tye, W.B. and C. Lapuerta (1996) "The economics of pricing network interconnection: theory and application to the market for telecommunications in New Zealand," *Yale Journal on Regulation* 13:419–500.
Tye, W.B. and H. Leonard (1983) "On the problems of applying Ramsey pricing to the railroad industry with uncertain demand elasticities," *Transportation Research A*, 17:439–450.
Vickers, J. and G. Yarrow (1988) *Privatization: an economic analysis*. Cambridge: MIT Press.
Viscusi, W.K., J. Vernon and J. Harrington (1995) *Economics of regulation and antitrust*, 2nd edn. Cambridge: MIT Press.

Wilson, G.W. (1980) *Economic analysis of intercity freight transportation*. Bloomington: Indiana University Press.
Winston, C., T.M. Corsi, C.M. Grimm and C.A. Evans (1990) *The economic effects of surface freight deregulation*. Washington, DC: Brookings Institution.

This Page Intentionally Left Blank

Chapter 4

ECONOMIC DEREGULATION IN THE USA[a]

ROBERT J. WINDLE
University of Maryland, Baltimore, MA

1. Introduction

The USA, as well as many other countries, has undertaken extensive deregulatory efforts over the past 25 years. These efforts have included a dismantling of many of the regulations faced in the transportation sector, but also included deregulation of many other industries. These deregulated industries often have the characteristic of being what we currently refer to as network industries. Network industries generally provide a good or service over a defined network. The network can be a series of railroad tracks, wires that go into houses, pipes that transport oil or gas, or routes that deliver people or goods to various locations. The common theme to these networks is that they often involve the outlay of fixed costs to build the network and the existence of externalities created by the network. These externalities generally result from the benefits of multiple users employing the same network. Deregulation has focused on these industries largely due to the fact that regulation was often extensive in these network industries.

The move to deregulate these industries began in earnest in the USA in the mid-1970s. This has allowed researchers approximately 25 years to assess the results of deregulation in some US industries (e.g. railroads, trucking, airlines, and long-distance telephone service). The track record for other deregulation efforts is somewhat shorter (e.g. ocean shipping, local telephone service, and electricity). The results of this deregulation wave have generally been reported to be positive by researchers who have looked at the impact of deregulation. Deregulation has not been a panacea for all the problems in these industries, and calls for re-regulation occasionally surface, often during times of economic duress.

This chapter will focus primarily on the deregulation of transportation industries, but will refer to deregulation in other industries with similar characteristics. Section 2 will provide a brief overview of why these industries were regulated in

[a]The author would like to thank Martin Dresner for helpful comments.

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

the first place. Section 3 will then review the history of deregulation in airlines, railroads, trucking, ocean shipping, and related network industries. Section 4 will review the results of deregulation in each of these industries, and discuss continuing problems. Section 5 provides some conclusions. Although much of the discussion pertains to economic regulation and deregulation in general, most of this discussion focuses on the US regulatory system.

2. Rationale for economic regulation

The history of regulation in US network industries goes back to the nineteenth century. Railroads were the first of the network industries to fall under regulatory rule, with the passage of the Interstate Commerce Act in 1887, which created the Interstate Commerce Commission (ICC). The arguments for regulation of the US railroads would be repeated as additional network industries fell under the regulatory umbrella over the next 50 years.[a]

The basic arguments revolved around the natural monopoly characteristics of railroads. Railroads involved an enormous investment in land and tracks to provide service. These investments resulted in a large percentage of railroad costs being fixed. Spreading these costs over increasing units of output (either freight or passenger service) results in declining unit costs. The result, from a cost standpoint, was that it often made sense to have only a single railroad serve individual markets. The resulting monopoly power of the railroad generated complaints from shippers who were at the mercy of the monopoly pricing practices of railroads. If a second railroad entered the market, it might mitigate the pricing power of the first railroad, but might also result in "destructive competition." The high fixed cost nature of the industry implied that marginal costs were extremely low. The competitive impulses of the two railroads would drive prices down to marginal cost where neither railroad is able to cover the extensive fixed costs of the network, resulting in bankruptcy of one or both railroads.[b]

In response to these issues, a regulatory bargain was created. The ICC would limit the level of competition in the industry by restricting entry, setting minimum rates, allowing some collusion among the railroads, and allowing a regulated return on railroad investment. In return, the railroads were required to act as common carriers with an obligation to serve all customers in a non-discriminatory fashion at reasonable rates. Rates would have to be approved by the ICC, and many firm decisions were subject to ICC review. Railroads were also prohibited

[a]See Kahn (1989) for a complete description of the theories of natural monopoly and destructive competition.

[b]See Viscusi et al. (1997) for a history of regulation and deregulation in various network industries.

from abandoning routes without ICC approval. This regulatory paradigm became known as public utility or rate of return regulation.

Public utility regulation was later adopted in many other network industries. The Motor Carrier Act of 1935 subjected the US trucking industry to ICC regulation, including decisions regarding entry, exit, and pricing. Economists argued from the onset of motor carrier regulation that trucking did not have any natural monopoly characteristics. In a recurring theme, this legislation was prompted in large part by the competitive threat that motor carriers posed to the regulated railroad industry. The Airmail Act of 1934 subjected airmail rates to ICC jurisdiction. A competitive bidding process was used by the ICC to award route contracts. Excessive entry and airline bids below costs throughout the 1930s resulted in severe financial problems for US airlines. As a result, the airline industry was brought more extensively under the regulatory umbrella in 1938 with the passage of the Civil Aeronautics Act that formed what was later to become the Civil Aeronautics Board (CAB). The CAB was the airline counterpart to the ICC, and regulated prices, entry, and exit in the airline industry.

The ocean shipping industry has a much longer history than the above transportation industries. Shipping cartels arose in the USA in the nineteenth century to counteract the tendency of the industry to overbuild. These cartels set rates and divided output and revenue among its members. As such, the cartels operated in a similar fashion to the regulatory agencies. The Shipping Act of 1916 granted these shipping conferences limited antitrust immunity in the USA. As with airlines, the US telecommunications sector was originally regulated by the ICC, but, starting in 1934, was regulated by the newly formed Federal Communications Commission (FCC). US electric utilities were regulated primarily at the state level by public utility commissions. These were mostly in place by 1930. The consolidation of the electric utility industry across state boundaries resulted in some issues concerning the power of the state agencies to regulate electric rates. The Public Utilities Holding Company Act and Federal Power Act were both passed in 1935 to clarify the power of the states to regulate electric utility rates. Cable TV fell under FCC regulation by 1966, based on the language of the Communications Act of 1934 and the request of commercial television stations, who were intent on limiting the burgeoning competitive threat posed by cable TV in the USA. Telecommunications, electric utilities, and cable TV all involved substantial investments in constructing a network of wires to distribute their product. These investments created substantial fixed costs, and as a result declining unit costs in these industries. All three could be classified as natural monopolies.

Public utility regulation dominated these industries for decades in the USA. Over the years economists began to point out various failures of this regulatory regime. In particular they pointed out that rate regulation focused on the price of the service provided, but failed to control for the quality of service. This could result in the production of products with poor quality or the transfer of destructive competition from price to service aspects of the business. In addition to quality concerns, public utility regulation may provide only a limited incentive for firms to engage in innovative behavior. This is particularly true when there is a monopoly provider being regulated. Cost savings were generally required to be passed on to consumers in the form of lower prices, and product improvements would not allow for an increase in the rate of return on investment. Finally, public utility regulation has to set reasonable rates. Economic efficiency requires that prices be set equal to marginal cost. If firms really were natural monopolies, marginal cost pricing would involve a loss for the firm and thus require some subsidy or the adoption of a "second-best" pricing scheme. Even if the industry was not a natural monopoly, determining marginal cost is a complicated exercise. Refinements to public utility regulation over the years have tried to deal with these issues.[a]

3. History of network deregulation

In the 1960s and 1970s various economic studies were published that cast doubt on the efficacy of public utility regulation. Many of these articles highlighted the issues noted above. In addition to these basic problems with the regulatory regime, there were empirical studies that questioned the assumptions and results of regulatory policy in particular industries. Advances in computing power and the introduction of flexible functional forms for estimating cost functions allowed for a more rigorous investigation of the structure of costs in transportation industries. A number of empirical articles were written that indicated that airlines, railroads, and trucking were not characterized by increasing returns to scale, and therefore were not natural monopolies.

Perhaps even more influential were a series of articles (Levine, 1965; Jordan, 1970; Keeler, 1972) comparing the regulated airline industry with the largely unregulated intrastate airline industry in the USA. There were two US states (Texas and California) that were large enough geographically and with a sufficiently large enough population that they could support their own airlines. Federal regulation of firms occurs when a firm is involved in interstate commerce. Since these airlines operated totally within a single state's boundaries, they were not subject to federal airline regulation. Comparisons of these intrastate markets to the regulated interstate markets revealed much lower prices in the unregulated markets. Additional research indicated that regulation did not result in financially healthier airlines. Airlines simply shifted their competitive endeavors to the

[a]See Train (1991) for a discussion of the problems with regulation and attempts to correct these problems through improvements in the regulatory process.

service aspects of flying, such as flight frequency, meals, and the purchase of new aircraft. Service competition raised the costs of doing business, thus limiting profitability.

In the US railroad industry it was the financial performance of the railroads that had the biggest impact on the move toward deregulation. Due to extensive overbuilding and the competition from the expanding motor carrier industry, railroads had come under increasing financial strain throughout the 1920s and 1930s. Bringing motor carriers under the ICC's regulatory umbrella was designed to limit the intermodal competition that was thought to be one of the railroads' main problems. The expansion of the automobile in the post Second World War period further devastated the railroads by attracting much of the rail passenger traffic. Abandonment of passenger traffic and regulatory oversight did not solve the railroads' financial woes. The bankruptcy of the Penn Central Railroad in the 1970s raised the specter of a possible wholesale bail-out of the railroads by the government.

While the US railroads were suffering financially under the regulatory regime and wanted relief, the trucking industry seemed to be prospering financially and had no desire to see significant regulatory change. However, evidence was accumulating that regulation of the trucking industry resulted in higher prices and inefficient operations. As with the US airlines, academic studies comparing regulated interstate carriers with unregulated intrastate carriers indicated higher prices and lower levels of service in regulated markets. In addition, the high value of motor carrier operating certificates indicated that economic rents were accruing to industry participants. There were indications that some of these rents were being passed on to labor in the form of above-market wages.[a]

In the US telecommunications industry it was primarily potential competitors that initiated the move toward deregulation. The US telecommunications service has historically been divided into three different markets: equipment, long-distance service, and local service. The introduction of microwave technology in the transmission of voice traffic significantly reduced the economies of scale associated with providing telecommunications services, particularly the provision of long-distance telephone service. In addition, the dominant US telephone operator, AT&T, undercharged for local telephone service by charging long-distance rates that exceeded the cost of providing service. This type of cross-subsidization was possible with a single monopoly provider. The new technology combined with the subsidization of local rates through higher long-distance charges prompted the rise of competitors, notably Microwave Communications, Inc. (MCI), who challenged the regulated monopoly (AT&T) in court. Equipment manufacturers also took AT&T to court in the 1950s and 1960s in an attempt to

[a]See Gallamore (1999) and Teske et al. (1995) for a review of these regulatory issues.

produce and sell telecommunications equipment to the retail customer. This is a story similar to the US trucking/railroad situation. New competitors were eager to provide competitive services, often to the disadvantage of the existing regulated company.[a]

In the US electric utilities there was less direct evidence that regulation was a problem. State regulation continued over both the production and distribution of electric power. Deregulation in this sector was stimulated by the theoretical shortcomings of public utility regulation, and in the 1980s the perceived success of deregulation in other network industries. Shortcomings identified by economists included the preference of regulated utilities to adopt capital-intensive solutions (the so-called Averach–Johnson effect) and the lack of incentives for regulated utilities to improve their productivity.[b]

With strong evidence of existing regulatory shortcomings, a movement to deregulate US transportation industries commenced in the mid-1970s. Airline deregulation began with administrative steps within the CAB to allow a greater degree of pricing freedom in 1976 and 1977. The Airline Deregulation Act of 1978 set a timetable for the elimination of most restrictions on pricing, entry, and exit in the industry. These changes were supposed to be phased in over a 7 year period, but in practice the domestic industry was mostly deregulated within the first year. The international routes of airlines have also benefited from a liberalization of the international rules. International routes are subject to regulation established by bilateral negotiations between countries. Since the late 1970s there has been an increase in the number of bilateral agreements that allow for increased pricing and route freedom. Deregulation of these international routes cannot be accomplished unilaterally, requiring agreement between the signatory countries.

The path to US railroad deregulation began with the Railroad Revitalization and Regulatory Reform Act of 1976, which provided some rate freedom to the railroads for selected commodities and confidential contracts. The impact of this act was lessened due to the resistance of the ICC commissioners to regulatory change. Appointments of new and more pro-deregulation commissioners in the late 1970s, combined with the passage of the Staggers Act in 1980, provided the impetus for a more fundamental change in the regulatory environment. The result has been the elimination of most railroad rate regulation through explicit exemption of many commodities or the allowance of private contracts between shippers and the railroads. Even where rate regulation remains, the onus has shifted from the railroads to show that rates are "just and reasonable" to the shippers to show that the rate exceeds 180% of variable costs, that there is no

[a]See Brock (1986) and Crandall (1991) for a thorough discussion of these issues.

[b]See Train (1991), Joskow and Schmalensee (1985), and Joskow (2000) for a discussion of these issues.

effective shipping alternative, and that the rate is unreasonable. Deregulation has also made it much easier for railroads to abandon unprofitable routes and encouraged mergers that promote efficiency (Grimm and Windle, 1998; Gallamore, 1999).

As with the US airlines, US motor carrier deregulation began even before the passage of formal legislation. Administrative deregulation provided increased pricing flexibility to motor carriers in the late 1970s. The passage of the Motor Carrier Act of 1980 resulted in complete deregulation of pricing, entry, and exit in the industry.

The US ocean shipping industry has also moved toward deregulation over the past 20 years. The Shipping Act of 1984 permitted individual service contracts between shippers and ocean carriers. This allowed ocean carriers to deviate from the cartel set rates, and compete on price. The Ocean Shipping Reform Act of 1998 has further enhanced the ability of ocean carriers to compete on price by increasing the confidentiality of individual service contracts. Previously, agreements required public disclosure of many terms. The 1998 Act made most of the service contract terms confidential, including the price (Federal Maritime Commission, 2000).

Other US network industries have also been deregulated to some extent over the last 20 years. The Cable Television Deregulation Act of 1984 deregulated cable TV rates, although entry is still granted by local governments through a franchise agreement. The franchise is supposed to be awarded to the low bidder. It is this bidding process that is supposed to keep rates and costs under control.[a]

A good deal of US telecommunications deregulation came about as a result of competitors challenging the monopoly provider (AT&T) in court. By 1977 the equipment market had been opened up by court-mandated requirements to use a standard plug and jack. The FCC required a separate unregulated subsidiary to manufacture and sell telecommunications equipment in 1980. MCI's challenge to AT&T eventually resulted in the FCC allowing MCI to enter the long-distance telecommunications market in 1975. In 1974, partially as a result of the evidence brought by potential competitors over the years, the US Justice Department brought an antitrust case against AT&T, accusing it of illegally monopolizing the industry. This case resulted in the break up of AT&T into seven local operating companies and AT&T. The seven local operating companies provided a local phone service in different geographical areas where each had a monopoly. AT&T continued to operate in the now competitive equipment and long-distance markets.

Access to many of the facilities owned and operated by the local Bell operating companies in the USA was necessary in order to provide long-distance service, as

[a]See Viscusi et al. (1997) for a discussion of the use of franchise bidding to generate a competitive market.

well as the expanding wireless services. Attempts to provide access to these local network elements was the primary motivation behind the 1996 Telecommunications Act. This legislation required local operating companies to unbundle these network elements and make them available for resale. For instance, a wireless carrier needs to connect with the local exchange and use the wire line into a customer's home to provide service. The local exchange and the wire into the home are owned by the local operating company, who now must sell these services to the wireless operator. These facilities are generally considered too costly to replicate and thus represent a significant barrier to entry for potential competitors. This has resulted in extensive regulatory issues regarding the pricing of these unbundled elements (Crandall and Hausman, 2000).

In the US electric power industry, deregulation has proceeded on two fronts. One is the requirement that utilities must purchase power from alternative power producers. The 1978 Public Utility Regulatory Policy Act required utilities to purchase power from companies that were able to generate electricity as a by-product of their main line of business (co-generation) and from those using renewable energy sources to produce electricity. The Energy Policy Act of 1992 allowed for the creation of independent wholesale power producers. The second development in the US electric power industry that has encouraged the formation of independent power producers is the requirement in the Energy Policy Act of 1992 that utilities must transmit the power produced by independent power producers over their transmission lines. Thus, a power producer in Kansas can sell power to a utility in Ohio by transmitting its power over lines owned by the utilities in Missouri, Illinois, and Indiana (Joskow, 2000).

Large parts of the US telecommunications and electric power industry remain regulated. This is primarily due to the existence of natural monopoly characteristics in the distribution of power and local telephone service. Most transportation network industries have been largely deregulated.

4. Results of network deregulation and continuing issues

This section will examine the results of the US deregulation efforts of the past 25 years. There are several summaries of the overall impact of deregulation; for example, Winston (1993, 1998), Grimm and Windle (1998), and Grimm and Winston (2000). In general the results have been viewed by economists as being beneficial, although the most controversy remains in those industries that are characterized by large economies of scale (telecommunications and electric power).

The results of US airline deregulation have been analyzed by many authors. Some of the more important studies include Bailey et al. (1985), Morrison and Winston (1986, 1989, 1995, 2000), and the Transportation Research Board (1991). In general these studies have found that prices are lower and consumer welfare has increased after the industry was deregulated.[a] Given the freedom to reconfigure their route structures, many airlines adopted a hub-and-spoke route system. This increased flight frequencies for many routes by combining passengers with many different destinations on a single flight into a hub airport. Since passengers place a high value on being able to leave at their desired time, the increase in frequencies led to improvements in consumer welfare. Deregulation resulted in a flood of new entrants immediately following the relaxation of entry restrictions. Southwest Airlines has been the most successful of these new entrants.[b] Although many of these initial entrants have since merged or gone bankrupt, new carriers, such as JetBlue Airways, continue to enter the industry.

Continuing problems in the deregulated environment in the USA include charges of predatory pricing by established airlines, airport dominance, limited access to airport infrastructure, and the poor financial performance of the industry. Large existing airlines have been accused of cutting fares and expanding service when a new entrant begins service on a route. The intent of such actions is to drive the new entrant out of business. Attempts by the US Department of Transportation to define and limit this practice have been unsuccessful. The US Justice Department lost a case in court accusing American Airlines of engaging in predatory pricing. Researchers and policy-makers have found it difficult to separate competitive behavior from predatory behavior. Two factors that would seem to argue against systematic predatory behavior are that new carriers continue to enter the industry and the poor financial performance of the industry. If predatory behavior is being practiced, it has not been very successful in raising carrier profitability.

Airport dominance has been an issue since deregulation was first adopted. The adoption of hub-and spoke systems in the USA has resulted in a large percentage (60–90%) of an airport's departures being accounted for by a single airline. The concern is that the dominant carrier can raise fares on routes that begin or end at the hub. This ability to raise fares is generally attributed to market power exerted by the dominant carrier. This market power could be the result of the increased frequencies enabled by the hub operations. Hub premiums, as they are known, have been estimated at anywhere between 2% and 27%, with the preponderance of studies estimating the effect at about 20%. Although this has been a continuing concern, Morrison and Winston (2000) provide some evidence that hub premiums may be less of a concern than previously thought.[c]

[a]Morrison and Winston (2000) estimate that US fares are 27% lower under deregulation and that consumer benefits exceed US $20 billion.

[b]Southwest was originally an intrastate carrier in Texas and therefore is not, strictly speaking, a new airline, but was a new entrant to the interstate market.

[c]Morrison and Winston (2000) attempt to control for the impact of Southwest Airlines on fares at hub airports. They also point out that hubs provide some benefits that may justify higher fares.

As US airports have become more congested, the lack of airport infrastructure (gates and landing slots) has seriously limited the ability of carriers to compete on certain routes. This is a problem largely caused by the explosion of travel resulting from the lower fares generated under deregulation. One of the causes of this problem is failure to appropriately price infrastructure. While traffic in 2001 to 2003 is significantly reduced due to the economic recession and the events of September 11, 2001, it is likely that traffic growth will resume over the next several years and airport and airspace infrastructure shortages will once again present difficulties for the industry.

The financial health of the US airline industry has remained a concern throughout the past 25 years. The industry is impacted greatly by changes in the demand for airline services. Since many of its expenses are fixed in the short run, any downturn in demand will result in lower revenues, without a corresponding reduction in costs. Demand uncertainty is generally related to economic conditions, but is also impacted by events such as the 1988 Pan Am bombing and the September 11, 2001 terrorist attacks. In addition, the old regulated carriers have still not completely adjusted their labor rates and working conditions to the deregulated environment. As a result, new entrants often have a significant cost advantage over these older carriers even 25 years after deregulation.

US rail and trucking deregulation have often been evaluated jointly in an analysis of surface freight deregulation. Winston et al. (1990) is an important study in this area. Winston (1993), Teske et al. (1995), and Grimm and Windle (1998) provide a summary of other studies. Gallamore (1999) provides a rail-only evaluation. These studies agree that the financial condition of the US railroads has improved dramatically under deregulation. This turnaround seems to be primarily due to productivity improvements and cost savings as the railroad industry has restructured. The restructuring included significant abandonment of rail lines and increased merger activity. The kilometers of track operated by Class I railroads in the USA fell by 31% between 1975 and 1987. The level of service provided by railroads has also risen significantly as railroads are free to adjust rates, enter into individual contracts with shippers, abandon unprofitable routes, and merge with other railroads. The results for trucking in the USA are equally impressive. In addition to the surface freight studies listed above there have been trucking-only studies done by Chow (1991) and Corsi (1994). The results of these studies indicate that the US trucking service has improved, rates have fallen, and productivity has increased as a result of deregulation.

One of the advantages of deregulation in the USA is that it has sent cost-based price signals to rail and truck customers. Under regulation, rates were set by the ICC. There is ample evidence that rates were often set too high for some commodities and too low for others. These rates would determine the modal shares of trucks and railroads. Rates that did not accurately reflect the true costs of each mode would result in distortions in the modal shares and a loss in consumer welfare. Deregulation has largely eliminated this distortion, although regulation of certain rail rates remains.

The railroad business is still difficult and costly to enter. The result is that some US shippers are captive to a single railroad. This is particularly true for commodities where trucks do not provide a viable alternative (coal being the primary example). The Surface Transportation Board still regulates the rates of these captive shippers. Although trucking has been deregulated at the interstate level, there are still states that regulate intrastate rates in much the same manner as the ICC used to do for interstate rates.

The results with regard to the US ocean shipping industry are less extensive, due to the more recent and less extensive status of deregulation in this industry. Some results are presented by the Federal Maritime Commission (2000). This report concludes that an increasing number of rates are being set by individual contracts between shippers and ocean carriers. These replace the practice of setting rates through conferences. Individual contracts result in a wider variety of rates and service levels that appear to be driven by market forces.

While the results for transportation deregulation in the USA have been overwhelmingly positive, the results for other network industries have provided mixed results. Crandall and Waverman (1995) provide an assessment of the AT&T divestiture, while Crandall and Hausman (2000) look at the impact of the 1996 Telecommunications Act. Assessment of these regulatory changes is difficult, since much of the regulatory apparatus remains in effect in the telecommunications business. Deregulation was much more extensive in the US transportation sector, with only minor economic regulation remaining in effect. Deregulation has eliminated the cross-subsidies that existed in the US telecommunications sector. In long-distance telecommunications, rates have fallen dramatically, and AT&T's market share fell from 90% in 1984 to 35% as of December 31, 2002 (Rosenbluth, 2003). The 1996 legislation was intended to promote competition in the local telephone market. It required the local telephone companies to provide unbundled elements of the network to competitors. The major issue that remains subject to negotiation and regulation is the price to be charged for these unbundled elements. In order to provide an incentive for local companies to open up their networks to potential competitors, the 1996 act also allowed these local companies to enter the long-distance market if they can show that they face competition in their local markets. Progress on this front has been slow, although by 2003 many consumers in the USA now have a choice of local providers. It is still too early to determine how effective this local competition will be, but many parts of the USA are now seeing local and long-distance services being sold as a package by multiple providers.

Cable TV deregulation in the USA resulted in higher cable rates. As franchise agreements came up for renewal, it proved difficult to remove the current franchise holder. Without the threat of losing the franchise, and lacking any competition, cable TV providers proceeded to raise rates and were also noted for less than spectacular service. This resulted in the passage of the Cable Television Consumer Protection Act of 1992. This Act reintroduced some rate regulation to the industry. In recent years, satellite TV systems have provided actual competition to the cable TV franchises in the USA. The competitive environment promises to keep rates under control and raise the level of service. This reduces the need for public-utility-like regulation of the industry.

The experience with electric utility deregulation in the USA is still rather limited, and the results have often been less than encouraging. Since US electric utility regulation is primarily controlled at the state level, different states have proceeded down the deregulation path at different speeds. Those states furthest along the deregulatory path are attempting to create markets where none existed previously. In a competitive world, purchasers of electricity could enter into an agreement with any producer to provide them with electric power. Joskow (2000) provides details on the problems and progress on this front. Some of the difficulties include the ability to transmit power from the producer to the consumer. Often this may require negotiations with multiple owners of transmission facilities. Prices need to be specified in an environment with a great deal of demand uncertainty and no practical method of storing the product. This leads to situations where supply and demand may both be extremely inelastic in the short run. As a result a large increase in demand due to hot weather, or a reduction in supply due to a power plant outage, may result in a huge increase in price. The creation of markets to handle these issues is in its early stages. The experience in the California energy market in 2000–2001 illustrates the potential problems with creating efficient energy markets. Wholesale rates for electric power purchased by California utilities were essentially unregulated, while retail rates remained regulated. In this case, a supply and demand imbalance resulted in large increases in wholesale rates that could not be passed on to retail customers. California utilities found themselves selling power at a price below the purchase price, resulting in the bankruptcy of two of the nation's largest utilities. This experience provides a cautionary tale about the potential pitfalls of partial deregulation.

5. Conclusion

The evidence suggests that deregulation in the USA has worked best where regulation was perhaps unwarranted or where technology has introduced new forms of competition into a formerly regulated industry. The economic rationale for regulation often involves claims that the industry has large economies of scale, making it a natural monopoly. In some cases this simply was not the case. Trucking was brought under regulatory control in the USA because it made regulating the railroads easier. Airline regulation began in the USA largely in response to industry financial difficulties. Portions of the US telecommunications industry (equipment) and electricity industries (power generation) were regulated due to their ownership by a vertically integrated regulated monopoly. Over the years, evidence accumulated that these industries could be competitive if the firms were allowed to compete. In these cases deregulation has worked reasonably well.

In other cases in the USA, new technology provided competition in an industry that previously had the characteristics of a natural monopoly. Trucking provided a competitor for railroads in many markets. Microwave technology provided competition for the provision of long-distance telephone service. In these instances the old regulated industry often lobbies to prevent the entry of the new competitor or argues that the new competition should also be subject to regulation. As a result, trucking was brought under the ICC regulatory umbrella in the 1930s. It took MCI over a decade to get approval to provide competitive long-distance service, primarily due to objections brought by AT&T. These sectors have generally had a successful deregulatory experience. Competition has flourished and generally rates have come down and service and productivity have gone up.

Two industries in the USA that may also be going down this path are cable TV and local telephone service. Satellite TV provides a competitor for cable TV. Wireless telephones may provide a competitive alternative to the local telephone monopoly. To the extent that there proves to be effective competitors, these industries may also be able to replace regulation with a competitive marketplace. The advent of VOIP (voice over Internet protocol) as a possible substitute for telephone calls (both local and long distance) provides another potential competitor for the local telephone monopoly. Since Internet connections are increasingly provided by cable companies in the USA, regulatory authorities will have to deal with the further blurring of industry boundaries in the telecommunications industry.

The benefits realized through the deregulation of the industries noted above has established a clear preference for marketplace solutions over regulatory control in the USA. As a result, the federal and state governments in the USA have proceeded with deregulatory efforts even in industries that still possess huge economies of scale and may well be natural monopolies. The primary example of this would be the distribution of electric power. The local phone network may also exhibit natural monopoly characteristics, but the growth of wireless technology may provide a technological advance that allows entry without replicating the local phone distribution network. These natural monopoly markets pose a more difficult challenge for those in favor of marketplace solutions. It may be difficult or impossible to create competitors for certain segments of the industry. Some level of regulation may be necessary to guarantee access at a fair price to facilities that are too costly to duplicate. Residual regulation in the US railroad industry is designed to deal with just such monopoly conditions in certain markets. As a result the US telecommunications and electricity sectors may require regulatory involvement for the foreseeable future. However, if we have learned anything in the past century regarding regulation, it is that circumstances are constantly changing. New technologies create new competitors. Too often in the past, regulators have stymied competition by protecting regulated industries. It is a lesson we should not forget.

References

Bailey, E.E., D.R. Graham and D.P. Kaplan (1985) *Deregulating the airlines*. Cambridge: MIT Press.

Brock, G.W. (1986) "The regulatory change in telecommunications: the dissolution of AT&T," in: L.W. Weiss and M.W. Klass, eds, *Regulatory reform: what actually happened*. Boston: Little, Brown.

Chow, G. (1991) "US and Canadian trucking policy," in: K. Button and D. Pitfield, eds *Transport deregulation: an international movement*. London: MacMillan.

Corsi, T.M. (1994) "Motor carrier industry, structure and operations," in: *Conference proceedings 3: International Symposium on Motor Carrier Transportation*. Washington, DC: National Academy Press.

Crandall, R.W. (1991) *After the breakup: US telecommunications in a more competitive era*. Washington, DC: Brookings Institution.

Crandall, R.W. and J.A. Hausman (2000) "Competition in US telecommunications services: effects of the 1996 legislation," in: S. Peltzman and C. Winston, eds, *Deregulation of network industries: what's next?* Washington, DC: AEI-Brookings Joint Center for Regulatory Studies.

Crandall, R.W. and L. Waverman (1995) *Talk is cheap*. Washington, DC: Brookings Institution.

Gallamore, R.E. (1999) "Regulation and innovation: lessons from the american railroad industry," in: J. Gómez-Ibáñez, W.B. Tye and C. Winston, eds, *Essays in transportation economics and policy: a handbook in honor of John R. Meyer.* Washington, DC: Brookings Institution Press.

Grimm, C. and R. Windle (1998) "Regulation and deregulation in surface freight, airlines and telecommunications," in: J. Peoples, ed., *Regulatory reform and labor markets*. Boston: Kluwer.

Grimm, C. and C. Winston (2000) "Competition in the deregulated railroad industry: sources, effects and policy issues," in: S. Peltzman and C. Winston, eds, *Deregulation of network industries: what's next?* Washington, DC: AEI-Brookings Joint Center for Regulatory Studies.

Jordan, W.A. (1970) *Airline regulation in America: effects and imperfections*. Baltimore: Johns Hopkins University Press.

Joskow, P.L. (2000) "Deregulation and regulatory reform in the US electric power sector," in: S. Peltzman and C. Winston, eds, *Deregulation of network industries: what's next?*. Washington, DC: AEI-Brookings Joint Center for Regulatory Studies.

Joskow, P.L. and R. Schmalensee (1985) *Markets for power: an analysis of electrical utility deregulation*. Cambridge: MIT Press.

Kahn, A.E. (1989) *The economics of regulation*. Cambridge: MIT Press.

Keeler, T.E. (1972) "Airline regulation and market performance," *Bell Journal of Economics*, 3:399–424.

Levine, M. (1965) "Is regulation necessary? California air transportation and national regulatory policy," *Yale Law Journal*, 74:1416–1447.

Morrison, S.A. and C. Winston (1986) *The economic effects of airline deregulation*. Washington, DC: Brooking Institution.

Morrison, S.A. and C. Winston (1989) "Enhancing the performance of the deregulated air transportation system," in: M. Neil Bailey and C. Winston, eds, *Brookings papers on economic activity: microeconomics, 1989*. Washington, DC, Brookings Institution.

Morrison, S.A. and C. Winston (1995) *The evolution of the airline industry.* Washington, DC: Brooking Institution.

Morrison, S.A. and C. Winston (2000) "The remaining role for government policy in the deregulated airline industry," in: S. Peltzman and C. Winston, eds, *Deregulation of network industries: what's next?* Washington, DC: AEI-Brookings Joint Center for Regulatory Studies.

Rosenbluth, T. (2003) *Standard & Poor's industry surveys, telecommunications: wireline.* New York, Standard and Poor.

Teske, P., S. Best and M. Mintrom (1995) *Deregulating freight transportation*. Washington, DC: AEI Press.

Train, K.E. (1991) *Optimal regulation*. Cambridge: MIT Press.

US Federal Maritime Commission (2000) *The Ocean Shipping Reform Act: an interim status report*. Washington, DC: US Federal Maritime Commission.

US Transportation Research Board (1991) *Winds of change: domestic air transport since deregulation*. Washington, DC: US Transportation Research Board.

Viscusi, WK., J.M. Vernon and J.E. Harrington, Jr (1997) *Economics of regulation and antitrust*. Canbridge: MIT Press.

Winston, C. (1993) "Economic deregulation: days of reckoning for microeconomists," *Journal of Economic Literature*, 31:1263–1289.

Winston, C. (1998) "US industry adjustment to economic deregulation," *Journal of Economic Perspectives*, 12:89–110.

Winston, C., T.M. Corsi, C.M. Grimm and C.A. Evans (1990) *The economic effects of surface freight deregulation*. Washington, DC: Brookings Institution.

This Page Intentionally Left Blank

Chapter 5

TENDERING OF SERVICES

JOHN PRESTON
University of Oxford

1. Introduction

This chapter will make international comparisons of the tendering of transport services. In doing so, it will draw on research carried out for the European Commission, principally the ISOTOPE (Improved Structure and Organisation for urban Transport Operations of Passengers in Europe) and MARETOPE (Managing and Assessing Regulatory Evolution in Local Public Transport Operations in Europe) projects. For comparisons beyond Europe, this chapter draws heavily on the biannual International Conference on Competition and Ownership in Land Passenger Transport, most recently held in Rio de Janeiro, Brazil (see http://www.its.usyd.edu.au/conferences/thredbo).

2. Forms of tendering

Tendering is a generic term that for the purposes of this chapter involves firms bidding for the right to operate services. It has a long pedigree, dating back at least to the work of Chadwick (1859), and is seen as a means of introducing competition for the market as an alternative to competition in the market. Its origins were as a means of regulating natural monopolies but it has become a way of introducing competition for unprofitable services. As a result, tendering schemes are sometimes referred to as limited competition, controlled competition, or competition for the market models. Such models have had a long history in the public utilities sector of certain countries (particularly France). In the transport sector they were given a boost by experiments with the tendering of bus services in San Diego (1979), London (1984), and Copenhagen (1989). Such models are now relatively common for the provision of both bus and rail services, on which this chapter will concentrate. They have also been applied to subsidized air and ferry services, while competitive contracting-out of activities has become a widely applied business practice (e.g. Domberger, 1998).

The key feature of these models is some form of competitive bidding procedure, akin to an auction, but within public transport markets three main variants have

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

emerged (ISOTOPE, 1997). First, the so-called Scandinavian model (although pioneered in Europe by London) is the classic form of tendering. It is based on minimum production cost (a contract form also known as "gross cost") tenders at a route level and is the dominant organizational form for public transport services in Denmark and Sweden. Secondly, the French model (which is dominant in France outside Paris) is based on network management contracts. In such cases, the authority typically provides the vehicles and related infrastructure, and firms bid for the right to manage these resources. This is thus an operating contract, in contrast to the complete contracts typical in Scandinavia, in which operators usually supply the vehicles. The third category is a hybrid contract (referred to by ISOTOPE as the Adelaide model), which involves contracting out of some tactical (e.g. fares and service planning) as well as operational functions.[a] Such contracts will often also attempt to expose bidders to revenue risk as well as production risk. Typically, operators will keep the revenue, and bids will be in terms of minimum subsidy (a contract form known as "net cost" – although note that such contracts can also feature in variants of the Scandinavian model). For profitable routes, bidders might pay a premium in order to have the right to operate. Hybrid contracts have become particularly prevalent for rail systems, where they are sometimes called "franchises" or "concessions," as there is greater scope for price discrimination while the long life and specificity of assets encourages long contracts. Concessions are also common for new light-rail and road schemes, and may involve financial arrangements such as BOT (build, operate, and transfer) and DBOM (design, build, operate, and maintain), but consideration of these forms of contracting-out is beyond the scope of this chapter.

Given the above, the structure of this chapter will be as follows. First, some of the theoretical evidence on the impact of tendering will be reviewed. Next, the empirical evidence on the impact of tendering, with respect to both the bus and rail sectors, will be examined. Lastly, an overview will be provided and some conclusions will be drawn.

3. Theoretical evidence on tendering

Muren (2000) analyses three contractual forms: gross cost contracts, net cost contracts, and contracts with measured service quality included as an explicit variable. The objective is to minimize costs subject to a minimum level of service. Using an analytical framework associated with Lewis and Sappington (1991) and Laffont and Tirole (1993), it is shown that under the gross cost contract the operator must make some profit even after a competitive bidding process, otherwise

[a]See Van de Velde (1999) for more details on the "strategic, tactical, operational" classification applied to organizational forms in public transport.

it would pay the operator to cut service quality. By reducing the length of the contract period, this profit, which is a transfer from the authority to the operator, can be reduced. The net cost contract gives, in principle, a possibility of achieving the desired level of service quality with a lower profit accruing to the operator. This is because the operating firm can lose in two ways if it cuts quality of service: it reduces its chances of getting its contract renewed, and it will lose fare revenues. The net cost contract thus gives the operator stronger incentives to produce quality of service. However, if there is high variation in the number of passengers, the net cost contract may need to compensate the operator for taking the revenue risk. Such compensation reduces the relative advantage of the net cost contract over the gross cost contract. The incentive contract with quality made explicit in the contract resembles the net cost contract in that it is risky for the operator. These contracts may be useful to the extent that it is possible to find quality variables that are easy to measure and for which the operator can predict the revenue effect of investment in service quality with relative certainty.

Another important area of theoretical study is the application of auction theory to explain bidding strategies (for a useful summary see Kennedy, 1995). At least two types of auctions are possible. Independent-value auctions occur where bidders have different valuations for the good being auctioned. For example, a bus company will require lower amounts of subsidy if the contract fits in well with existing work or can be easily served from an existing depot compared with a bidder for whom these characteristics do not apply. In such a case, the bid will increase (i.e. the amount of subsidy will decrease) as the number of bidders increases. Common-value auctions occur where all bidders have the same valuation of the good being auctioned but are uncertain about the value of that good. In such auctions, they may increase their bids as the number of bidders increases in order to win the auction, but if they do this they may run the risk of the "winner's curse." In other words, they may overestimate the value of the good being auctioned and pay more for it than it is worth. To avoid the winner's curse, bidders may be particularly cautious when there are a large number of bidders. The two effects may cancel out, and, as a result, in a common-value auction the bid price is not expected to vary with the number of bidders.

Given that operating costs for a specific service will vary between operators depending on the availability of staff, vehicles, and depots, particularly for relatively small amounts of service, gross cost contracts may have the characteristics of an independent-value auction, and hence bids may increase with the number of bidders. By contrast, it may be expected that revenue from a given bus service under a net cost contract with fixed fares would have a common value for bidders. Net cost contracts may therefore have more features of a common-value auction, and hence the number of bidders may be less important.

A number of other aspects of auction theory may be worth considering. Vickrey auctions award contracts to the highest bidder based on the price of the second highest bidder (Vickrey, 1961). This is believed to reduce strategic bidding, and could have a role in public transport, and has been suggested for the auctioning of train paths and take-off and landing slots at airports. A two-stage bidding process with separate bids based on price and quality could also be considered (referred to in the literature as the Brook's law procedure). Menu auctions (combined bids) are also possible, whereby, for example, bidders make bids for routes separately and in various combinations. This enables bidders to exploit economies of scope but may also encourage strategic behavior, for example by permitting cross-subsidy between routes. Van de Velde and Sleuwaegen (1997) raise further important issues concerning pre-selection procedures, the use of qualitative criteria, additional negotiations after selection, etc.

There have been a number of attempts to simulate the impacts of tendering. For example, Preston et al. (2000) have undertaken detailed modeling of competition for the track in the UK. Overall, their analysis suggested that a move to longer (12 year), exclusive and loosely regulated franchises could lead to an annual subsidy reduction of 25% compared with the proposed regime. The analysis also suggested that if there are only one or two bids per franchise, subsidy requirements are likely to increase. For example, a reduction in the number of serious bidders from four to one was forecast to increase subsidy requirements from between 50% (for a 5 year franchise) to almost 80% (for a 10 year franchise). Winning bid forecasts based on up-to-date financial information were estimated and validated with data on actual bids. It was found that franchises let at the outset of the franchising program were generally awarded for less than their forecast "market" value, whereas franchises awarded toward the end of the process were awarded for substantially higher than their forecast market value.

4. Empirical evidence on bus tendering

There have been a large number of studies of bus tendering, also referred to as off-the-road competition. Some of these are summarized in Table 1, with particular emphasis on changes in unit operating costs (per bus-kilometer or service hours). The evidence is dominated by variants of the Scandinavian model, with some evidence beginning to emerge on hybrid models. There is little published evidence, at least in English, on the French model.

4.1. Evidence on the Scandinavian model

Unsurprisingly, evidence comes mainly from Scandinavia, but also from the UK and other European countries, as well as Australasia and the USA. Jansson (1994) has reviewed the early stages of tendering in Sweden, where tendering has been

Table 1
Empirical evidence on the impact of bus tendering. See text for full details of each study

Source	Country	Impact on unit costs (%)
Jansson (1994)	Sweden	–12
Alexandersson et al. (1998)	Sweden	–13
Alexandersson and Pyddoke (2003)	Sweden	–5
Jansson (1994)	Denmark	–10
Johansen & Senstadvold (1996)	Norway	–20
YTV (1996)	Finland	–29
Glaister and Beesley (1991)	UK (London)	–16
Kennedy (1995)	UK (London)	–16
dctlparCox et al. (1997)	New Zealand (Auckland)	–34
Carlquist and Johansen (1999)	Sweden (Helsingborg)	–27
Walters and Cloete (2003)	South Africa	+25
Marcucci (2003)	Italy (Rome)	–14 (a)
Cox and Duthion (2003)	USA	–40 (b)
Wallis (2003)	Australia (Adelaide)	–38
Wallis (2003	Australia (Perth)	–22
Unweighted average		–18

Notes: (a) unweighted average of three tenders; (b) unweighted average of seven cities (Dallas–Fort Worth, Denver, Los Angeles, Minneapolis–St Paul's, San Diego, San Francisco, and Seattle).

gradually introduced since July 1989. Cost savings from the initial rounds of tendering varied from 0 to 45%, with an average of around 12%. Later work in Sweden has been undertaken by Pyddoke (1996), who found that moving from one to two bids reduces costs by 12%, moving from one to three bids reduces costs by 17%, and moving from one bid to four bids reduces costs by 20%. Some diseconomies of scale were found in that contracts with large amounts of bus-kilometers had higher unit costs. Costs per kilometer were found to be 16% lower in sparsely populated areas where speeds will be high and peakiness of demand low. A further review of Swedish experience, provided by Alexandersson et al. (1998), suggests that competitive tendering has reduced unit costs by 13%. There was no trace of indirect effects. Also referred to as ripple effects, they involve cost-savings spreading from areas that have been exposed to competitive tendering to those that have not. The latest work by Alexandersson and Pyddoke (2003) suggests that the reductions in real unit operating costs over the period 1987 to 2001 may only be around 5%. However, the problem of the counterfactual is compounded by the long time period studied.

In particular, the assumption that labor costs would have only increased in line with inflation in the absence of tendering is probably unrealistic. Jansson (1994) also reviews the early impacts of tendering in Copenhagen where costs were reduced by around 10% between 1989 and 1992. A feature of the initial

Copenhagen model was that the municipal operator was not allowed to bid, although a higher number of bids were still received per contract than in Sweden. Copenhagen has subsequently standardized on contract lengths of 6 years, with combined bids limited to a maximum of three packages. A value analysis model has been developed to adjust for variations in, for example, the age of vehicles, staff experience, and past quality of service. Furthermore, bids have to be based on average salary costs to prevent operators competing on working conditions. Another important feature of the Copenhagen model is the quality measuring system that has been developed in order to provide operators with financial incentives. This system has had a number of iterations but currently is based on ten attributes, nine of which are based on passenger perceptions derived from a survey of 30 000 customers per year. Since tendering round 9 (2000) a malus/bonus scheme has been applied, in which operators can be fined up to 1.5% of the contract sum if they fail to meet quality standards, while conversely there is a reward of 1.5% of the contract sum if quality standards are exceeded. This bonus is varied depending on service reliability, being only 0.45% if less than 99.86% of service is operated, rising to 1.8% if 99.96% or more of service is operated. An inspection system is also in operation that permits on-the-spot fines of between €30 and €400 for various minor breaches of contract, with provisions for contract withdrawal for major breaches. The results appear to be a high level of customer satisfaction, some initial increases in patronage (up 8% between 1993 and 1997) but subsequently decreases and some further reductions in costs (down by 24% between 1990 and 1997) but with some increases since (up 14% between 1997 and 2003) (HUR, 2003). The Copenhagen scheme seems to have had some benefits but has evolved into a highly bureaucratic regime, the overall efficacy of which has yet to be proven (e.g. see the contradictory results of the welfare analysis in MARETOPE (2002)).

In Norway, tendering in Lillehammer initially reduced unit costs by 20%, while frequencies increased by 5%, patronage by 30%, and revenue by 20% (Johansen and Senstadvold, 1996). Unit costs have subsequently increased in Lillehammer but against a background of a continuing rise in patronage (now up almost 50% since tendering) as a result of improved service quality (Søberg, 2001).

A study of the tendered bus market in the UK outside of London (which comprises only between 15 and 20% of the total bus market) indicated that gross cost contracts may reduce subsidy by 13% compared with net cost contracts provided there is strong competition (White and Tough, 1995). UK data also indicate that vehicle size and age specifications may increase subsidy by between 5 and 10% (Pickup et al., 1991).

Glaister and Beesley (1991) found that in London, where tendering was gradually introduced between 1984 and 1994, the initial rounds reduced costs by 16% when administration and monitoring costs were taken into account. Combination bids (one firm bidding for a package of routes at a lower subsidy requirement than their bids for each individual route) were believed to distort the market. Glaister and Beesley also found that publicly owned firms had higher unit costs than privately owned firms, suggesting that the market was not yet in equilibrium. However, follow-up work by Kennedy (1995) found such equilibrium, with the reduction in unit costs similar to those of Glaister and Beesley (when administration costs are taken into account). Later work on tendering in London (Toner, 2001) indicated that the switch from gross cost to net cost contracts led to an increase in subsidy requirements, albeit in a system where operators only handle around 30% of revenue. It was believed that, due to concerns over revenue risks, small operators were reluctant to bid for net cost contracts. This is also likely to have been a factor in White and Tough's analysis discussed above. There was also some evidence in London that small tenders achieved higher bid prices per bus-kilometer than large tenders.

There has been much concern in the UK about the rising costs of tenders, but this is only just beginning to show through in the aggregate statistics, with public transport support administered largely through tenders reaching a minimum level in 1997–1998 but by 2000–2001 had increased by 64%. Initial work on this issue, based on case studies of the administrative areas of Cheshire, Kent, Staffordshire, Suffolk, and Tyne and Wear, found a number of reasons for such cost increases, including rising staff and vehicle costs (which in some cases was exacerbated by rural bus grants), falling revenue, and higher required rates of return. Contracts with open "break" clauses might encourage early termination of contracts and subsequent higher prices. However, there was no clear correlation between the number of bids and unit costs, suggesting that the tendered market is contestable at the time of bidding in that the threat of potential entry is effective (UK Department of Environment, the Regions and Transport, 1999).

In Finland, an initial round of tendering in Helsinki cut subsidy costs by 29%, allowing ticket prices to be reduced by 8% and vehicle-kilometers to increase by 3% (YTV, 1996). Recent work on the impact of bus tendering in Italy, in Rome, indicates that a batch of three tenders led to unit operating cost reductions of between 8 and 26% (Marcucci, 2003).

Cox et al. (1997), in a worldwide review, emphasize that immediate conversion to competitive tendering produces the greatest cost reduction rates. Examples quoted include Auckland (unit costs down 33.5%, service levels up 16.5%, 1990–1996) and Las Vegas (unit costs down 33.3%, service levels up 243%, 1993–1994). Phased conversion produces more moderate cost reductions. Short-term conversion (5 years or less) typically reduces unit costs by 15% (Adelaide, Copenhagen, Perth, Stockholm). Ripple effects were observed with long-term conversion programs, as in San Diego (tendering commenced in 1979) and London (tendering commenced 1984). *Ad hoc* competitive tendering programs, which are typical in the USA (Cox (1999) examines 30 such programs), tend to produce savings on the competitively tendered services only, and hence no ripple effects (this was initially true also of Sweden). The most recent work by Cox and Duthion

(2003) includes evidence on bus tendering from several US cities. The main impact appears to be reductions of unit operating costs of around 40%.

4.2. Evidence on hybrid models

Adelaide is often cited as a pioneer of a tendering approach that gave operators some responsibility for network planning, although initial implementation was cautious, with only modest cost reductions, some service improvements, and no strong evidence of changes in patronage (Radbone, 1997). Subsequently, a more ambitious revenue incentive-based regime has been introduced, which has seen service levels increase by 15%, unit costs reduce by 38%, and patronage increase by 8%. A similar form of tendering in Perth has led to an increase in service levels of 32%, unit cost reductions of 22%, and a patronage increase of 26% (Wallis, 2003).

In the Netherlands, tendering was initially limited to two experiments in rural areas. Bidders were asked to suggest a better network for the same amount of subsidy as the present operator. In Limburg, this resulted in a US company (Vancom) winning, with 30% more bus-kilometers for the same subsidy. In Sealand, the incumbent won the contract, with 15% more bus-kilometers than the year before. This approach was criticized for being based on a coarse supply oriented set of selection criteria with unclear incentives, but it did illustrate that competition was feasible in the Netherlands (Van de Velde, 1995). Some recent regional tendering experiments have led to a 60% increase in service hours (Amersfoort), a 10.5% increase in service hours, a 6% reduction in subsidy (South Holland), and an 18% increase in service hours (North West Utrecht) (Van de Velde and Pruijmboom, 2003).

A further example of a hybrid system is Helsingborg in Sweden (population approximately 110 000). The bus system was tendered in 1992, and an agreement between the operator and the authority drawn up in 1994. The operator was given responsibility for service and fare planning, subject to authority controls concerning minimum levels of frequency and network density. An immediate cost reduction of 27% was achieved. Service increased by 4%, and out-turn fares by 22% between 1994 and 1996. Despite this, patronage remained stable, and the cost recovery ratio increased from 33% in 1991, to 60% in 1993 and to 73% in 1996. The operator can lose the contract if patronage falls by more than 3% in a year or if customer ratings of quality fall below a certain standard. In 1997 a non-tendered net cost contract was introduced above the existing contract (Carlquist and Johansen, 1999). Quality targets have been introduced, with the operator getting an extra payment corresponding to 5% of revenues if the target is met. If quality is below a certain threshold, a penalty must be paid. The Helsingborg experiment is being repeated in Sundsvall, where, since June 1999, the operator is required to increase patronage by 4% per year. If this fails, the operator has to spend 4% of the net reimbursement per year on marketing. Further experiments in which the operator is given responsibility for planning are being undertaken in the city of Södertälje in the Stockholm region (where the authority defines minimum frequencies and maximum walk distances) and for two new express bus services in Skåne (where reimbursement will be related to demand) as well as in a few other places in Sweden.

4.3. Overview of bus tendering

Overall, empirical analysis of tendering suggests that cost reductions of between 10% and 20% can be achieved if there is no restructuring of the bus industry, while reductions of 35% or more can be achieved if there is also restructuring (fragmentation and privatization), particularly if there is labor market flexibility. Further aggregate evidence on the impact of bus tendering at the European level came from ISOTOPE (1997), which in a cross-sectional study for over 100 cities in the mid-1990s suggested that limited competition models could reduce costs per vehicle-kilometer by 36% and increase vehicle-kilometers per member of staff by 18% compared with traditional regulated/municipal operations models. This issue was further examined by in MARETOPE (2002), which examined a pooled cross-sectional and time series database of 31 cities over the period 1991–2000. For bus systems, it was found that tendering increased vehicle-kilometers per member of staff by 36%. If operators bore the production risk, unit costs were reduced by 10%, compared with a 9% reduction if they bore both production and revenue risk. By contrast, for combined bus, tram, and rail, it was found that if operators bore production risk unit costs were reduced by 27%, compared with a 30% reduction if they bore both production and revenue risk. The analysis undertaken in MARETOPE attempted to analyze the dynamics of the impact of tendering. The figures quoted above are short-run estimates of the impact of the first year. Typically, the long-run impact over a 10 year period would be around double these impacts, albeit with the vast majority of change (90%) occurring in the first 4 years. A welfare analysis suggests that total social cost (i.e. operator costs plus user costs) per passenger was reduced by 17% when operators bore both production and revenue risk, and by 13% when operators bore production risk only.

ISOTOPE concluded that tendering had clear advantages in terms of efficiency in production – that is, the ability to produce a given output at minimum cost. However, it also concluded that tendering did not have clear-cut advantages in terms of efficiency in consumption – that is, the ability to produce the quantity and quality of service demanded by consumers at the most appropriate price. Therefore, it was unable to forecast the impact of tendering on overall economic efficiency. MARETOPE reconfirmed the advantages of tendering in terms of productive efficiency, and has led to a better understanding of the dynamics of impacts and of attribution to different measures. There was some weak evidence to suggest that for the bus industry gross cost tenders (in which the operator bears only the production risk) may be more effective than net cost tenders (in which the operator bears both the revenue and production risk). This analysis also suggests that net cost tenders may be more appropriate in the rail industry (possibly because the greater proportion of off-vehicle sales permits a greater degree of price discrimination). With respect to efficiency in terms of consumption, MARETOPE was unable to detect any strong effects other than tendering was associated with increased service levels per unit area, although this does not necessarily imply causation. Demand analysis failed to pick up any effect associated with tendering over and above the impact of such reforms on price and service levels. With respect to overall economic efficiency, MARETOPE produced some evidence that welfare per passenger is positively correlated with the introduction of competition and with making operators bear the production risk. These enhancements in overall economic efficiency are due almost exclusively to increases in productive efficiency.

The main evidence on the impact of bus tendering comes from cities with developed economies. However, there have also been experiments with tendering in cities with developing or transitional economies (Gwilliam, 2003). Net cost area schemes have been implemented in Bahrain, Kuala Lumpar (Malaysia), and Kingston (Jamaica). Net cost route-based tenders have been implemented in Bishkek (Kyrgyzstan), Kazakhstan, Rostov (Russia), Santiago (Chile), South Africa, and Uzbekistan. Gross cost route-based tenders have been adopted in Bogata (Colombia) and Curitiba (Brazil). These experiments have generally been less successful than those in the developed world, due to various institutional frailties, problems with immature markets, and inauspicious macroeconomic conditions. Unexpected results may occur. For example, the introduction of tendering in South Africa led to a 25% increase in subsidy. However, it also led to a reduction in the average age of the bus fleet from 12 to 6 years, and was introduced in the context of a conventional urban bus sector, which, in the face of competition from informal transport, was on the verge of bankruptcy (Walters and Cloete, 2003).

5. Empirical studies of rail tendering/franchising

Empirical studies of rail franchising are less common than for bus. Table 2 summarizes some findings concerning off-track competition. Two extremes can be detected. The first is the Latin American concession model based on long contracts, of up to 50 years plus extensions. The franchisee owns and maintains the rolling stock and maintains the infrastructure, although infrastructure ownership

Table 2
Summary of passenger rail tendering/franchising

Case study	Length (years)	Extension (years)	No. of bidders	Major investment	Maintenance of rolling stock	Maintenance of infrastructure	Contract specification	Award criteria
UK	5–15	15–20	5.4 per franchise (range 3–8) (a)	Varies	Rolling stock companies (ROSCOs)	Track authority (Railtrack now Network Rail)	Passenger service requirements Fares Operating performance	Net cost
Argentina	30 (b)	10 (b)	4 (b)	Yes	Franchisee	Franchisee	Maximum fares	Three stages: experience, investment, and net cost
Brazil – Metro	20	20	4	No	Franchisee	Franchisee (state undertakes upgrades)	Flat tariff Frequency, reliability, safety, and comfort targets	NPV of concession fee
Brazil – Flumitrens	25	25	5	No	Franchisee	Franchisee (state undertakes upgrades)	Flat tariff Frequency, reliability, safety, and comfort targets	NPV of concession fee
Germany – VRR (two routes)	5	0	3/1	No	Franchisee	Track authority (DB AG)	Fares Timetable performance	Gross cost
Germany – VRS (three routes)	15	0	3/2/2	No	Franchisee	Track authority (DB AG)	Fares Timetable performance	Net cost
Sweden Jönköping Länstågen	4	1	2	No	Franchisee (heavy maintenance by JLT)	Track authority (Banverket)	Fares Timetable performance	Gross cost

Source: Preston et al. (2001). *Key*: NPV, net present value. *Notes*: (a) 4.2 bids per re-franchise (range 2–9); (b) inter-city services.

rests with the state. In the Brazilian model, major infrastructure upgrades are undertaken by the state. In the Argentinian model, upgrades are undertaken by the franchisee, but with arrangements for compensation if the contract is prematurely terminated, while regulation is limited to a relatively loose control of fares. Regulation is more comprehensive in the Brazilian model, in addition covering frequency, reliability, comfort, and safety. Operators bear the revenue risk, and hence contracts are based on a net cost basis.

The second extreme is what might be termed the European tendering model. Tenders are relatively short (5 years or less), the winning firm has no responsibility for infrastructure, and fare and service levels and other aspects of performance are heavily prescribed by the state (often acting at a local level). In Sweden, local authorities usually bear the revenue risk, and hence local tenders are usually awarded on a gross cost basis, although regional tenders are awarded on a net cost basis. Germany also has a mixture of gross and net cost contracts. Elsewhere (the Netherlands, Switzerland, and, more recently, Denmark), net cost contracts are the norm.

It should be noted that the so-called franchising regime in the UK is somewhere between these two extremes, initially resembling the European tendering model, except based on net costs. Subsequently, the Strategic Rail Authority's proposals for re-franchising intended a move toward a variant of the Latin American model (longer contracts with some operator involvement in investment). These proposals were not implemented, and instead the emphasis on short franchise extensions has seen a shift back in the UK toward the European tendering model.

The concession model has not just been restricted to passenger services but has been successfully applied to freight railways, particularly in Latin America and Africa. Thompson (2003) reports on concessions in Argentina (five concessions), Bolivia (two), Brazil (seven), Chile (three), Guatemala, Mexico (three), Peru, and Panama, as well as Cote d'Ivoire/Burkina Faso and Malawi. Thompson reports that in all cases traffic has increased and that on average freight rates have reduced by more than 30%.

Evidence on the impact of competition for the rail market is limited. In the UK, franchising reduced direct subsidy to passenger rail operations from central and local government from UK £2319 million in 1995–1996 to UK £1146 million in 2002–2003 (in 1999–2000 prices) – a decrease of over 50%. However, this ignores direct support now paid to Railtrack/Network Rail, which raises the total subsidy in 2002–2003 to UK £2420 million – an increase of 4% on the 1995–1996 level. Moreover, work by Knowles (2003) indicates that the subsidy received by train-operating companies in 2001–2002 was almost 20% greater than that forecast at the time the original franchises were let.

There is also some important evidence on rail tendering from Sweden. Alexandersson and Hulten (2003) find that gross cost tenders have reduced subsidy by between 10 and 40%, while net cost tenders have reduced subsidy by between 10 and 100%. Detailed analysis of two routes in Sweden by Preston et al. (2001) found that, for one route, based on a net cost contract, subsidy had reduced by 50%. However, for the other route, based on a short gross cost contract, a reduction of subsidy of only 10% was achieved. Moreover, on re-tendering, subsidy went up by 50% from its original level, largely due to improved rolling stock specifications. Overall, tendering/franchising appears to have been more problematic in the rail sector than the bus sector. There appears to be greater scope of strategic game playing, including predatory pricing and attempted regulatory capture, and a greater risk of being afflicted by the winner's curse. This in turn may be related to the greater technical complexity and capital intensity of rail compared with bus operations. In any event, there have been some spectacular business failures, for example in Melbourne (Australia) and Stockholm (Sweden), leading some to question the efficacy of tendering approaches (Stanley and Hensher, 2003).

6. Overview

Preston and Shaw (2000) have highlighted a number of issues concerning contract competition, contract specification, contract award, contract enforcement, and contract renewal. These will be discussed in turn.

With respect to contract competition, recent evidence from London, Scandinavia, and elsewhere indicates that proactive measures are needed to ensure that tendering is competitive in initial and subsequent rounds. As a minimum, an active private-sector competitive fringe is required. It seems sensible to unbundle large state-owned transport enterprises and to commercialize their management, possibly in preparation for privatization. However, in extreme cases this might involve maintaining some public sector presence in the bus market. For example, Jansson (1994), highlights the important residual role of publicly owned operators in Sweden. Anti-trust legislation needs to be alert to the potential for re-agglomeration to lead to local monopolies or cartels, although the evidence that large firms have advantages in competitive tendering is relatively weak, at least for route-based contracts. Policy should be proactive in encouraging the participation of small and medium enterprises and, for developing countries, engaging the so-called informal sector.

Contract specification is clearly of critical importance. In practice, contract lengths vary between 6 months and 99 years. Contract lengths of 4–5 years seem to be appropriate for bus contracts, but contracts should be longer for rail, particularly if operators are responsible for providing rolling stock. Operating contracts in which authorities and/or third parties provide the rolling stock to train operators may be a useful device to inject more competition into the rail market. There is little consensus on the relative merits of gross cost and net cost contracts despite some empirical support for the former for bus and the latter for rail. However, the lack of consensus may become increasingly irrelevant given the move toward hybrid contracts based on gross cost plus revenue incentives (so-called output-based funding). Another important issue with respect to contracts that is unresolved is their areal extent. Route-based and area-wide contracts both have their supporters, although it is probably fair to say that too small contracts are less of a problem than too large contracts because they can be re-packaged by bidders into larger contracts. Contract specification should also ensure integration within and between public transport modes. Furthermore, it may be possible to consider contracting-out aspects of the planning function as well as the operating function. Evidence from cities that have made some progress in this respect, in Australia, the Netherlands, and Sweden, appears to be promising. To determine optimal specifications it may be necessary to experiment directly in the market place. Alternatively, simulation models might be helpful in determining appropriate contracts, as, due to strategic game playing by operators, demonstration projects and other experimental trials may be misleading.

The contract award process should be transparent, with as much information as possible in the public domain. In certain countries specific legislation has been required to ensure this, e.g. the Sapin law in France.

Contract enforcement costs can be substantial, with detection of contract breaches often difficult. Contract withdrawal seemed to be the most effective deterrent. Although financial penalties have a role to play, they need to be at relatively high levels (and accompanied by adverse publicity whenever these penalties are incurred) in order to provide sufficient incentives. Contract renewals need to be dealt with carefully. Contracting authorities should be prepared for contractors to go bust. They should have contingency plans for such eventualities to ensure continuous service delivery that do not involve bailing out the original contractor. Tendering bodies need to be particularly cautious about mid-term contract re-negotiations and automatic extensions.

7. Conclusions

The empirical evidence presented tends to back some form of competitive tendering for public transport largely because of its advantages in terms of achieving productive efficiency, with decreases in unit costs of around 20% typical, starting from a situation of an unchallenged public monopoly operation. It has also been shown that there are many tendering options to choose from. Important issues include length of contract, ownership of vehicles, ownership of terminals, initiation of fare policy, initiation of service levels, the basis of the award, penalties for physical underperformance, bonuses for customer satisfaction, and vehicle specification (Preston, 1997). However, international evidence does suggest some possible ways forward. It appears preferable to introduce tendering quickly and comprehensively, and it seems sensible to avoid cost-plus contracts. Service delivery penalties and vehicle size and age specifications can improve performance but can also increase costs. At least for gross cost contracts, theory suggests that the number of bidders might increase the bid price (reduce the subsidy requirement), although empirical evidence on this is mixed. A number of important issues remain unresolved by empirical evidence, including the relative merits of route or network, gross cost or net cost, and long or short contracts. However, there appear to be some preferences for short, route-based, gross-cost-based contracts for bus operations, and longer, area-based, net cost contracts for rail.

More recently, competitive tendering has come under attack. The good performance of public transport in cities such as Dublin, Sydney, and Toronto that have not had competitive tendering (although have had the threat of tendering) has been highlighted, as has the high-profile failures of tendering in cities such as Melbourne (Mees, 2002). Evidence of rising tender prices is invoked, although as yet rarely quantified, with the result that attribution of the causes of any such cost increases is not clear. Critics will also point out that competitive tendering has not had conspicuous success in increasing the usage of public transport and has led to service ossification. This is perhaps a little unfair. Experiments with various forms of incentives have led to service innovations and patronage increases of between 8% and 26% in bus-tendering experiments in Australia, and such schemes are continuing to evolve. Contracting out of the strategic and tactical functions associated with network planning is also being considered in addition to the more traditional contracting out of operations. One possibility is to give operators greater responsibility for tactical functions, as in the hybrid approaches typified by Adelaide and Perth in Australia, Helsingborg in Sweden, and a number of other Dutch and Swedish towns and cities. There might be further scope for the contracting out of tactical functions to third parties (the so-called architect model), although there are obvious issues about contract specification and evaluation of bids.

There is some evidence that tendering (and its variants) has proved more problematic for rail. There may be greater problems of ensuring that bidding is competitive, of incumbency advantages, and of the inefficiency of the winning bid (because of the greater role of economies of scale, scope, and density). There may also be added complications for contract specification, evaluation, and enforcement, given the greater scope for product differentiation in the quality dimension. Second best, Vickrey auctions might reduce strategic game playing. Similarly, variable-term franchises might restrict the scope for contract re-negotiations, at least in cases where up-front capital investment is required (Engel et al., 2001). What is clear is that there is likely to be continued innovations in tendering processes in the future, with a move away from simple variants to more sophisticated approaches. Whether in practice these have the desired effects in promoting productive, allocative, and dynamic efficiency will remain a largely empirical issue.

References

Alexandersson, G. and S. Hulten (2003) "European regulation and the problem of predatory bidding in competitive tenders – a Swedish case study," in: *8th International Conference on Competition and Ownership in Land Passenger Transport*. Rio de Janeiro.

Alexandersson, G. and R. Pyddoke (2003) "Bus deregulation in Sweden revisited: experiences from 15 years of competitive tendering," in: *8th International Conference on Competition and Ownership in Land Passenger Transport*. Rio de Janeiro.

Alexandersson, G., S. Hulten and S. Fölster (1998) "The effects of competition in Swedish local bus services," *Journal of Transport Economics and Policy*, 32:203–219.

Carlquist, E. and K.W. Johansen (1999) *Local public transport systems. Financial organisational frameworks in Norway and abroad*, TØI Report 451. Oslo: TØI.

Chadwick, E. (1859) "Results of different principles of legislation in Europe: of competition for the field, as compared with competition within the field of service," *Journal of the Royal Statistical Society*, 22:381–420.

Cox, W. (1999) "US competitive tendering: comprehensive cost analysis," in: *6th International Conference on Competition and Ownership in Land Passenger Transport*. Cape Town.

Cox, W. and B. Duthion (2003) "Private participation in US public transport: issues and perspectives," in: *8th International Conference on Competition and Ownership in Land Passenger Transport*. Rio de Janeiro.

Cox, W., J. Love and N. Newton (1997) "Competition in public transport: the international state of the art," in: *5th International Conference on Competition and Ownership in Land Passenger Transport*. Leeds.

Domberger, S. (1998) *The contracting organization: a strategic guide to outsourcing*. Oxford: Oxford University Press.

Engel, E., R. Fisher and A. Galetovic (2001) "Least-present-value-of-revenue auctions and highway franchising," *Journal of Political Economy*, 105:993–1020.

Glaister, S. and M. Beesley (1991) "Bidding for tendered bus routes in London," *Transportation Planning and Technology*, 15:349–366.

Gwilliam, K. (2003) "Land passenger transport in the developing world and World Bank policy," in: *8th International Conference on Competition and Ownership in Land Passenger Transport*. Rio de Janeiro.

HUR (2003) *From Copenhagen Transport to Greater Copenhagen Authority, Transport Division*. Copenhagen: HUR.

ISOTOPE (1997) *Final report*. Luxembourg: Office for Official Publications of the European Communities.

Jansson, K. (1994) "Swedish competitive tendering in local and regional public transport. Overview and comparative case studies," in: *PTRC Summer Annual Meeting*. Warwick.

Johansen, K.W. and M. Senstadvold (1996) *Optimal Pricing is Important. Competitive Tendering in Public Transport in Oppland County*. TØI Report, 1037, Oslo.

Kennedy, D. (1995) "London bus tendering: the impact on costs," *International Review of Applied Economics*, 9:305–318.

Knowles, R. (2003) "Impacts of privatising Britain's rail passenger services – franchising, refranchising and Ten Year Transport Plan targets," in: *RGS-IBG Annual Conference*. London.

Laffont, J.-S. and S. Tirole (1993) *A theory of incentives in procurement and regulation*. Cambridge: MIT Press.

Lewis, T. and D. Sappington (1991) "Incentives for monitoring quality," *Rand Journal of Economics*, 22:370–384.

Marcucci, E. (2003) "Local public transport reform in Italy: the case of the city of Rome," in: *8th International Conference on Competition and Ownership in Land Passenger Transport*. Rio de Janeiro.

MARETOPE (2002) *Assessment of impacts of change*, Deliverable D5. Brussels: European Commission.
Mees, P. (2002) *Competitive tendering of bus services: international experience and lessons for Toronto*, Report. Toronto: Toronto Transit Commission.
Muren, A. (2000) "Quality assurance in competitively tendered contracts," *Journal of Transport Economics and Policy*, 34:99–112.
Pickup, L., G. Stokes, S. Meadowcroft, P. Goodwin, B. Tyson and F. Kenny (1991) *Bus deregulation in the metropolitan areas*. Avebury: Oxford Studies in Transport.
Preston, J.M. (1997) *The restaurant at the end of the regulatory cycle. A review of the options for regulatory reform of the British local bus market*, Working Paper 852. Oxford: Transport Studies Unit, University of Oxford.
Preston, J. and A. Shaw (2000) "Tendering and competition in public transport," *Transport Reviews*, 20:23–329.
Preston, J., G. Whelan, C. Nash and M. Wardman (2000) "The franchising of passenger rail services in Britain," *International Review of Applied Economics,* 14:99–112.
Preston J., T. Holvad, N. Sykes and J. O'Reilly (2001) *Review of empirical and theoretical evidence*, Deliverable D2. *Development of market models for increased competition in railroad passenger traffic*, Project for SJ AB, Working Paper 906. Oxford: Transport Studies Unit, University of Oxford.
Pyddoke, R. (1996) *The competition effects on costs in tendering of bus contracts in Sweden.* Stockholm: Department of Economics, Stockholm University.
Radbone, I. (1997) "The competitive tendering of public transport in Adelaide," in: *5th International Conference on Competition and Ownership in Land Passenger Transport*. Leeds.
Søberg, O. (2001) "Experiences with tendering in land passenger transport in Norway," in: *7th International Conference on Competition and Ownership in Land Passenger Transport*. Molde.
Stanley, J. and D. Hensher (2003) "Performance-based contracts in public transportation," in: *8th International Conference on Competition and Ownership in Land Passenger Transport*. Rio de Janeiro.
Thompson, L. (2003) "Changing railway structure and ownership: is anything working?" *Transport Reviews*, 23:311–356.
Toner, J. (2001) "The London bus tendering regime: principles and practice," in: *7th International Conference on Competition and Ownership in Land Passenger Transport*. Molde.
UK Department of Environment, Transport and the Regions (1999) *Tendered bus services: government response to the select committee report.* London: DETR.
Van de Velde, D. (1995) "The experience of the Netherlands: towards competition," in: *4th International Conference on Competition and Ownership in Land Passenger Transport*. Rotorua.
Van de Velde, D. (1999) "Organisational forms and entrepreneurship in public transport," *Transport Policy*, 6:147–158.
Van de Velde, D. and E. Pruijmboom (2003) "First experiences with tendering at the tactical level (service design) in Dutch public transport," in: *8th International Conference on Competition and Ownership in Land Passenger Transport*. Rio de Janeiro.
Van de Velde, D. and L. Sleuwaegen (1997) "Public transport service contracts: searching for the optimum," *International Journal of Transport Economics*, 24:53–74.
Vickrey, W. (1961) "Counter-speculation, auctions and competitive sealed tenders," *Journal of Finance*, 16:8–37.
Wallis, I. (2003) "Regulation and competition in the land transport industry in Australia and New Zealand," in: *8th International Conference on Competition and Ownership in Land Passenger Transport*. Rio de Janeiro.
Walters, J. and D. Cloete (2003) "Regulation and competition in the land transport industry in Southern Africa," in: *8th International Conference on Competition and Ownership in Land Passenger Transport*. Rio de Janeiro.
White, P. and S. Tough (1995) "Alternative tendering systems and deregulation in Britain," *Journal of Transport Economics and Policy*, 29:275–89.
YTV (1996) *YTV regional transport tenders 1994–1996*. Helsinki: Pääkaupunkiseudun yhteistyövaltuuskunnan.

This Page Intentionally Left Blank

Chapter 6

PERFORMANCE EVALUATION FRAMEWORKS[a]

DAVID A. HENSHER
University of Sydney

1. Strategic thinking, competitive advantage, and effective performance

Transport businesses in the private and public sector are increasingly committing themselves to performance evaluation. There are many reasons why such an interest is growing, including the desire to deliver services more efficiently and effectively as well as an opportunity to understand better the contribution of an organization's range of activities to its overall performance. Performance indicators are designed, through reflection, to encourage continuous improvement in the sense of "doing things right." The very process, however, can assist in clarifying objectives and responsibilities by moving beyond "doing things right" to signaling an opportunity to "do the right things." It can make performance more transparent regardless of the nature of markets in which a transport business is operating and regardless of whether the intent of business is profit maximization of some set of social objectives.

Benchmarking against best practice has provided one basis for setting standards of performance. In addition, in markets that are not subject to direct competition, benchmarking against best practice has become a surrogate for the ideals of competitive performance, referred to as yardstick or benchmark competition. Regulatory agencies often rely on such evidence to determine if an organization not subject to competition is acting in the interests of customers. The transport sector has come under particular scrutiny in recent years as governments, in their role as regulator and service supplier, seek to engender increased performance and productivity from all transport sectors – road, rail, shipping, ports, airports, airlines, trucking, bus, and coach.

These ideas might be best depicted as elements of a performance framework, interpreted herein to encompass a strategic focus on the business and not simply the establishment of operational effectiveness, the latter defined as best practice

[a]The comments of Bill Waters have materially improved this chapter.

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

in respect of cost or productive efficiency associated with the supply of a well-defined and essentially equivalent service. The most discerning feature of strategic thinking is differentiation – seeking out a set of activities (known as an activity system) that make a business "different" in a way that denies potential competitors the opportunity to imitate an entire activity system and hence erode profits and market share. Unlike the more narrow idea of performance measurement through benchmarking, designed to identify and implement effective performance, which works to eliminate differences, incorporation of strategic advantage in a performance framework promotes the creation and preservation of differences. The striving for international competitiveness through achieving best practice status on a set of (partial) performance indicators must be seen as a useful but very incomplete framework within which to offer advice on how a business might achieve a competitive advantage (Porter, 1985). Succeeding on individual measures of best practice such as labor productivity (e.g. net tonne kilometers per employee) is not sufficient to ensure strategic advantage – it is at best a necessary condition.

The success of Southwest Airlines' "no-frills" service illustrates how integrated niche differentiation (IND) is a recipe for strategic advantage (Freiberg and Freiberg, 1996). The Southwest Airlines activity system has many individual elements that are relatively easy to imitate. These include no baggage transfers, flexible union contracts, low fares, high compensation of employees, no seat assignments, limited use of travel agents, automatic ticketing machines, a standardized fleet of 737 aircraft, no connections with other airlines, no meals, 15 minute gate turnaround time, and a high level of employee share ownership. When all of these "process inputs" are combined into an integrated business we have an airline that has for over 15 years established and protected its competitive advantage through a niche market emphasis of limited passenger services, frequent reliable departures, high aircraft utilization, very low air fares, short-haul point-to-point routes between mid-size cities and secondary airports, and lean highly productive ground and gate crews. Airlines trying to compete, such as United Express, have failed to succeed because they have been unable to date to replicate and improve on all elements of the activity system. Porter (1985) attributes this type of strategic advantage to the fact that the fit locks out imitators by creating a chain that is as strong as its strongest link. United Airlines integrated its express service into a more complex non-niche activity system servicing the more costly main carrier, which denied any opportunity for IND in respect of the total activity system that delivers Southwest Airlines strategic advantage.

Competitive advantage thus derives from the set of activities and their mutual reinforcement. Performance measures, to be useful in a strategic sense, must encapsulate the integrated nature of business to truly replicate their competitive strength. This includes extending the idea of strategic gain to include not only the narrower commercial ideals of business but also the externalities that are the product of social commitment, especially applicable to public agencies.

This chapter focuses on frameworks within which to promote performance evaluation. It does not detail the quantitative tools for measuring performance, much of which has been presented in Chapters 20 and 35–39 in Volume 1 of this series (Hensher and Button, 2000).

2. Promoting a holistic framework: STO

As we develop ideas of performance of a transport business, we must not reduce the framework to one of inputs, outputs, and singular critical success factors (CSFs) but rather recognize the way in which these inputs "fit" together in the development of an overall strategic framework. This mindset should apply to all transport organizations, regardless of whether they are a public authority or a private (privatized) business operating in a monopoly, competitively regulated, or deregulated market.

A setting that has proven to be especially useful within which to position the obligations of organizations and stakeholders is the STO framework. It recognizes that policy, planning, and operations exist within a hierarchy of objectives functionally split into three interdependent layers – strategic, tactical, and operational. This organizational framework offers an attractive setting within which to evaluate mechanisms consistent with a holistic (or system-wide) perspective on service delivery. The main features of the framework are represented by three STO levels (van de Velde, 1999, 2001; Viegas and Macário, 2001):

- the strategic level – where the focus is on the establishment of broad goals and objectives and guidance on ways of achieving outcomes consistent with such goals ("what do you want to achieve" – van de Velde, 1999);
- the tactical level – which highlights the supporting mechanisms (e.g. the regulatory process) to achieve the strategic goals;
- the operational level – which focuses on delivering the desired services to the market consistent with the strategic intent and aided by tactical mechanisms.

The STO framework provides the context in which we can put to the test policies and practices that to various degrees support mixtures of what I have termed the Napoleon and the Anglo-Saxon codes on delivery of services to the market and the community at large (Hensher, 2001a). The Napoleon code focuses on transport as an input into a wider socio-political–economic framework. It is a device for achieving a range of policy objectives, and as such it is argued that the sector should be heavily regulated and controlled by government. The Anglo-

Saxon philosophy is that transport is just another sector in the economy that should be provided as efficiently as possible in its own right. Markets are preferred to government intervention, and private participation and commercial criteria are dominant. Economists describe this mixture as the ability to deliver social-welfare-maximizing outcomes under conditions of cost efficiency that can still support stakeholders in the supply chain whose pricing objective is profit maximization (i.e. strictly commercial). Regardless of policy objective, a necessary but not sufficient condition is the provision of a given level of service at the lowest cost. Establishing the commitment to particular codes influences the performance evaluation framework.

To give some practical detail to the STO framework, let us translate it into the specific sectoral setting of public transport provision. At the strategic level we have three goals – efficiency, equity, and environmental sustainability (see below for more detail). These goals set reference points against which more practical objectives can be assessed, such as the three As associated with the achievement of networks of transport systems: accessibility to vehicles and infrastructure; affordability in terms of tariff levels; and availability in terms of coverage of services. The three As suggest a greater emphasis on efficiency and equity; however, there is tacit recognition of the importance of environmental quality. Although the three goals are admirable, to be operational they need translation into a set of performance measures. An example of a strategic objective is an environmental target defined very specifically as an increase in public transport share, such that public transport operators and regulators can have an unambiguous interpretation of their obligations.

At the tactical level, where there is an emphasis on planning, actions should be driven by a set of policies emanating from a strategic focus that have definition, interpretation, and specification. An example of a planning task linked to the environmental target of an increase in public transport share is the establishment of a partnership agreement between the regulator, the infrastructure supplier, and the service operator on measures to attain the public transport target. Clear definitions of the allocation of such measures to all involved parties is essential. This involves not only specifications of service enhancements but also all support measures such as infrastructure and information systems.

At the operational level, the independent regulator, the support services suppliers, and the service suppliers should work together, with incentives and non-compliance penalty clauses imposed on all parties to achieve an agreed quality of service (Hensher, 2001a). Importantly, the tendering or regulatory authority should be subject to the same set of incentives and penalties as the end-suppliers in recognition that all parties are responsible for the specified outcomes. For example, if there is an agreed package between local authorities and operators to enhance services involving the provision of specific transport priority measures (e.g. bus priority, pedestrian access, and increased rail service frequencies), not only should the operator incur a penalty if they fail to deliver the service frequency but the public authority would also have to pay the operator if it fails to deliver the priority measures.

This holistic perspective is designed to reveal the necessary conditions for performance evaluation, while recognizing that the creation of differentiation in the market through judicious definition of "doing the right things" is aided by the building blocks associated with "doing these things right."

3. A useful checklist of broad principles for selecting performance measures

There are (at least) five key principles for selecting performance indicators and rules for operationalizing these principles that are consistent with the responsibilities of management throughout all organizations. These are:

- The indicators must relate to the objectives of an enterprise (which include internal and external considerations).
- They must be clearly definable and unambiguous in their interpretation such that a particular numerical value or change in value is unambiguously good or bad.
- Indicators must adequately distinguish between factors outside the control of an organization and those within it over well-defined time periods.
- They must be simple to comprehend by those who are in a position to influence the numerical magnitude, including those who directly contribute to the outcome.
- The results must be related to the overall analysis of performance. This requires an unambiguous definition of an improvement in performance (with links to the STO framework).

These principles recognize that we need overall measures of enterprise performance (e.g. total factor productivity, TFP) and a number of partial indicators that have a well-defined relationship with the overall measures to ensure that actions of owners of indicators are compatible with the achievement of improvement in overall performance. To operationalize this set of principles it is desirable to establish some interface with the responsibilities of management throughout each organization in the STO framework. This is best represented by choice-determining basic principles:

- a roll-up principle in which data collected at one level of the organization should be capable of being "rolled up" to the next level above it;
- a responsibility principle whereby managers should only be called to account for performance in their areas of control, responsibility, and authority;
- the hands-on principle under which managers are assessed on the basis of agreed performance indicators focusing on outcomes;

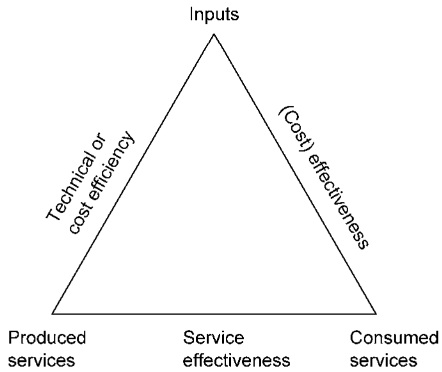


Figure 1. The essential dimensions of performance measurement.

- an ownership principle requiring that a performance indicator be "owned" by one or more managers having responsibility and authority for all outcomes measured by that indicator.

In exercising these principles, the choice of performance indicators has the character of negotiation of a contract. At each level of disaggregation, it is important to identify the types and magnitudes of inputs used to produce the various outputs. This is complicated at some levels of disaggregation by the sharing of inputs. A transport operator, for example, has a number of levels of operation to consider, although many of them can be easily accommodated provided that the most disaggregate unit of activity or service provision is well-defined in respect of data. This encourages all operators to start thinking in a bottom-up mode rather than a "traditional" top-down mode.

4. The traditional dimensions of performance evaluation

Performance has many dimensions. In broad terms we distinguish between efficiency and effectiveness (Figure 1). The efficiency of an enterprise represents the manner in which the physical inputs are used to produce the physical (intermediate) services. This is only one of the generally accepted (economic) definitions of efficiency – namely technical efficiency when no particular assumption on behavior of an organization (e.g. profit maximization) is imposed. The economic dimension – allocative efficiency on the input side – is combined with technical efficiency when an explicit assumption about cost minimization behavior is made. Some studies refer to this as overall or productive efficiency – i.e. combining of inputs to produce a given amount of output so as to minimize the costs of production. Effectiveness has two essential components: cost effectiveness – the relationship between inputs and consumed services – and service effectiveness – the relationship between produced services and consumed services. An organization can be effective without being efficient. All of these global measures are relative measures.

The capture of efficiency has been represented by TFP in applications using mainstream econometric models and index numbers, and by relative efficiency (RE) in applications using optimization models of the linear programming form. The latter is referred to as data envelope analysis (DEA). Where DEA concentrates on physical inputs and outputs, it is measuring technical efficiency; where it considers input prices and cost minimization, it is measuring productive efficiency. Waters (2000) provides an overview of TFP methods with applications by Nash (2000) to rail, Forsyth (2000) to airports, and Oum et al. (2000) to airlines. Rouse and Putterill (2000) review the DEA approach with an application to roads.

In many applications, notably rail and bus, economists measuring technical and/or productive efficiency have used consumed services (i.e. final outputs) to measure output, which is a cost-effectiveness relationship. The role of intermediate or produced services is ignored. This is equivalent to imposing no well-defined restrictions on the underlying production technology, and treating production as a "black box." To appreciate the importance of this point, it is useful to think of performance as a two-step strategy:

- In step one, a study of purely technical efficiency of an organization's production would use intermediate outputs to demonstrate how well the inputs are combined to produce outputs in the main under the control of the business (e.g. the application of labor, energy, materials, and capital in the railways to produce train capacity kilometers). This would translate into productive efficiency under conditions of cost minimization.
- In the second step, where the emphasis is on the effectiveness of the organization's service offerings and marketing policies, a model is constructed to represent the optimal choice of intermediate outputs, given the operating environment and final output demands. The price of final outputs, together with the price of competing outputs, and regulatory and institutional constraints, together with the produced services, all influence service effectiveness. For example, total annual passenger kilometers (i.e. final output) are influenced by train capacity kilometers (i.e. intermediate output), rail fares, competing mode use costs, and levels of service of competing modes.

The studies that ignore this two-step strategy have "collapsed" the measure of performance into a single dimension, namely cost effectiveness. As useful as this dimension is, it fails to make explicit the optimization of intermediate outputs, which are usually under the control of the organization, and hence limits the usefulness of the performance measure in identifying the sources of inefficiency that have arisen due to actions under the control of management of an enterprise. The ability to distinguish sources of performance that are controlled and are not controlled by the supplier is crucial in the development of suitable performance-based incentive and disincentive schemes. Sanctions for non-compliance are not just where the "owner" of a performance indicator has no control on the outcome.

5. Broadening the performance evaluation framework to capture the spirit of STO

The framework in the previous section while general may give the misleading impression that it focuses on the internal performance of an organization. To highlight the importance of both internal and external criteria we present a re-interpretation.

External criteria are of particular concern to public sector agencies but also for private enterprise, who are increasingly encouraged to be more environmentally and socially responsible. For example, a policy to introduce higher public transport fares requires evidence not only that this is in line with increased costs, but that the market responsiveness is not against the interests of broad external criteria such as environmental and social sustainability. Where there is a case that the mapping of prices with costs produces a loss of net consumer benefit, governments often prefer to maintain loss-making (i.e. non-commercial) prices and to provide direct operating subsidies to cover the gap, in the name of community service obligations (CSOs) and/or competitive neutrality. The establishment of gains in net social benefit per dollar of subsidy, for example, is one very explicit performance measure that incorporates internal (i.e. cost efficiency) and external (i.e. social justice) criterion. The challenge for a business subscribing to such external indicators of performance is to work on both best practice cost efficiency (i.e. delivering a given level of service for the lowest input cost) and identification of the externality impacts of specific practices. The latter provides the basis for determining the effectiveness of the organization in achieving its stated objectives within the STO framework.

Externalities should be defined broadly to cover all dimensions of costs and the set of objectives that are not accounted for by the narrower commercial definition of cost efficiency. The latter is limited to identification of the set of internal inputs (labor, capital, materials, energy, etc.) used to produce a given level of output at the lowest cost under an essentially commercial objective. Externalities thus become as much a part of the internal processes of operating a business as they are about the focus on how a business responds to external influences. We strongly support the position that a necessary but not sufficient condition for an effective enterprise is that it is cost efficient. One has to recognize, however, that the final determination of effectiveness must be guided by more than cost efficiency. The essential inputs are (marginal) costs (assumed to be delivered efficiently, which

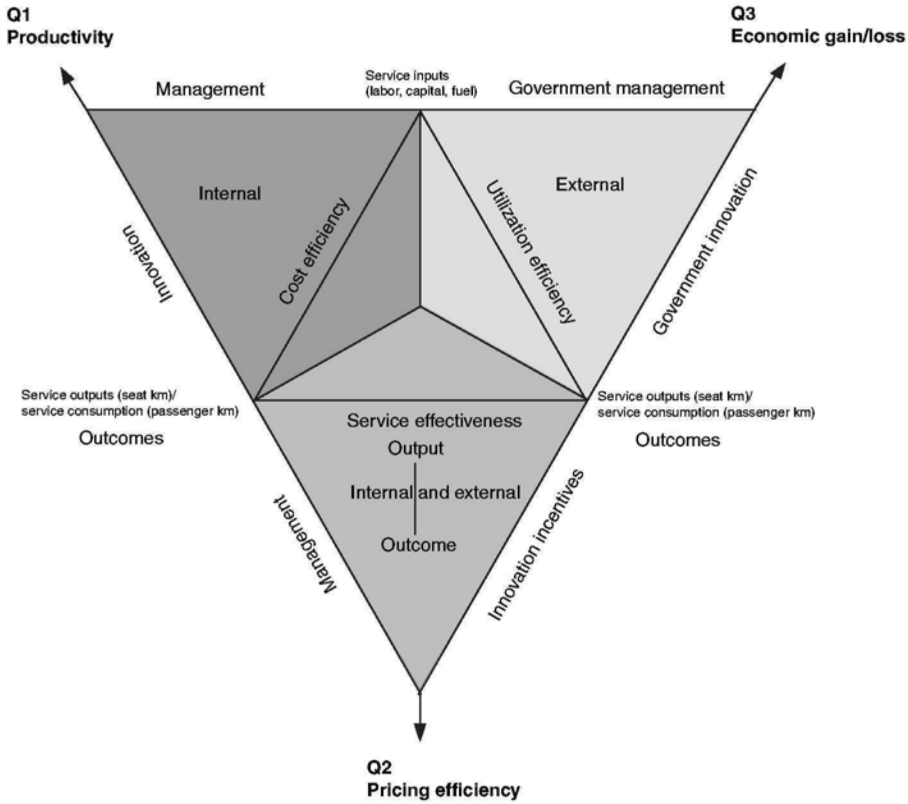


Figure 2. An overview of the components of internal and external performance determination.

can include outsourcing), direct and cross-price elasticities of demand for services (which identify how the market responds to the goods and services on offer), and the pricing strategy of the business (i.e. profit maximization, social welfare maximization, and constrained social welfare maximization, where the last focuses on covering average total costs). The combination of these three "inputs" overlaid by the rules of regulation and/or market forces determines the price to charge in the market. It is essential that we consider performance indicators that embrace more than simply best practice in terms of cost efficiency.

Traditionally, external objectives have been weakly linked to these dimensions of performance, yet it is essential to strengthen the links if we are to preserve the value of performance measurement in organizations where the objectives have a strong sense of external conformation. We might usefully summarize the internal and external dimensions of performance through an extended performance measurement framework as presented in Figure 2. Such a framework recognizes the role of the customer, the internal business, financial commitments, innovation, and learning. Each perspective is defined by a hierarchy of critical success factors such as productivity gain, organizational learning, financial improvement, and quality enhancement. The interlocking triangles in Figure 2 are a “folded-out” performance pyramid in which measures of customer satisfaction, flexibility and productivity link the strategic vision of the transport business to a set of underlying operational measures such as quality, delivery, process time, and cost (Rouse and Putterill, 2000).

Triangle Q1 measures the overall productivity of a business as defined by concepts such as TFP and DEA. TFP and DEA measure the productivity of all heterogeneous inputs to the production process. Its strength lies in its ability to recognize that many heterogeneous factor inputs are used to produce a number of heterogeneous outputs. It links three dimensions of corporate performance: cost efficiency, management quality, and innovation. Management effectiveness depicts the relationship between service inputs and overall productivity. Innovative efficiency depicts the relationship between service outputs and overall productivity.

Q2 measures the market effectiveness of a business through the idea of pricing efficiency. Market effectiveness depicts the relationship between output and the consumption of that output in the market place. As in Q1, two dimensions link pricing efficiency with output and consumption. First, the management dimension links output with pricing efficiency. Here, all the skills and quality of management come into play. As an entrepreneur, the manager must ensure that the pricing of goods/services brought to market are such as to ensure sufficient penetration of markets to enable a sufficient return on investment in the long run. Thus, entrepreneurial versatility, fund-raising ingenuity, ambition, and judgment are all brought into play in the decisions leading to the pricing of output. This activity complements the equally important issue of input pricing by management, covered in Q1. Second, the innovation dimension links consumption and pricing efficiency. It represents the improvement in quality in the goods/services offered at the price that customers are willing to pay.

The exigencies of relatively integrated policy-making require that major activities such as provision of transport services be measured for the efficiency with which they utilize economic resources, to determine whether the provision of service results in an economic gain or an economic loss to society. While it is important to measure service consumption, it is imperative to measure service output. This is particularly so in complex markets such as public transport, where over-servicing is prevalent and caused by many factors.

The final triangle, Q3, is labeled “economic gain/loss.” This is deliberately broad so as to avoid terms such as “consumer surplus” or “social benefit–cost ratio,” and avoids argument of what should be included in weighing the benefits of service provision against its costs. Two further dimensions link service input and service output. First there is the management dimension, labeled government management, which covers the whole gamut of intervention including regulation, taxation, subsidization, etc. It is useful for a business to know if it was winning or losing as the result of the impingement of government management upon its activities (an important interface between T and O in the STO framework). This applies as forcefully to the public sector for government business enterprises as it does to the private sector, and possibly more so, given the added complexity of multiple pricing strategies. Second, there is innovation that is government-sponsored innovation aimed at expanding the efficiency of business. This innovation may take the form of major improvements in infrastructure, the implementation of total quality management principles, etc.

Each triangle can be given a hierarchical feedback structure in a framework in which each triangle is represented by a surface that links the vision, goals, and objectives of the business with managerial measures, as defined by the critical success factors, to a set of underlying process drivers. The feedback between goals and measures is critical in the monitoring of the effectiveness of particular process drivers (often referred to as policy instruments in the public policy literature). The pyramid paradigm recognizes that process drivers or policy instruments associated with one triangle or plane of the performance pyramid can influence the other triangles, and hence are cross-linked. Decomposition of each overall critical success factor associated with each triangle (e.g. TFP, net social benefit per dollar of subsidy, market price of services) to establish the influence of a range of process drivers (e.g. input mix, incentive plan, CSO obligations) and broad contextual effects (e.g. location, scale) can assist managers in understanding the impact of process drivers on one or more result areas. Analytical tools such as TFP, DEA, and social cost–benefit analysis (SCBA) all provide ways of quantifying the global status of performance within a framework designed to capture the commitment of a business to multiple objectives, goals, and a vision.

To illustrate the broader basis of performance evaluation, we use an example of a proposed carbon tax of Aus. $0.20/kg for automobiles in Perth (Western Australia) and a range of performance measures. A number of key performance indicators are shown in Table 1, derived from the application of TRESIS, a transport and environmental strategy impact simulator (see Hensher 2001b; Marshall, 2003). Average vehicle operating cost after equilibration would increase by 18.18%. Total government revenue would increase by 14.82%. The total end-user money cost would increase by 7.34% while total end-user generalized cost would be reduced by 0.344%. Modal commuter shares for automobile trips would decrease, while those for public transport would increase. Total annual vehicle kilometers would reduce by 2.309%, and total greenhouse gas emissions would reduce by 2.392%. One can quantify these indicators for each application year in the evaluation period, and calculate the accumulated impact of the policy over a given period.

Table 1
Summary results for carbon tax: 2004, Perth, Western Australia

Indicators	Base case	Carbon tax	Difference (%)
Automobile operating cost			
AvOpCost (c/km)	6.553	7.744	18.18
VehOpCost (Aus. $)	6.351×10^8	7.333×10^8	15.451
Government revenue			
TGovtCarbT (Aus. $)	0	1.128×10^8	NA
TGovtExcise (Aus. $)	3.799×10^8	3.708×10^8	–2.391
TGovtPark (Aus. $)	9.867×10^7	9.847×10^7	–0.207
TGovtPT (Aus. $)	7.137×10^7	7.417×10^7	3.95
TGovtSales (Aus. $)	5.588×10^7	5.631×10^7	0.772
TGovtVehReg (Aus. $)	1.142×10^8	1.142×10^8	–0.132
Subtotal	7.200×10^8	8.268×10^8	14.823
Total end user cost			
TEUC.MC	1.371×10^9	1.472×10^9	7.341
TEUC.TC	7.75×10^8	7.724×10^8	–0.344
Commuter mode share			
TDA (share/No.)	69.90%/1.489×10^8	69.63%/1.483×10^8	–0.387
TRS (share/No.)	19.98%/2.027×10^7	19.90%/2.019×10^7	–0.387
TTain (share/No.)	1.99%/4.24 × 106	2.06%/4.392×10^6	3.603
TBus (share/No.)	8.13%/1.733×10^7	8.41%/1.791×10^7	3.399
Greenhouse gas emissions			
TCO2 (kg)	2.524×10^9	2.464×10^9	–2.392
Passenger vehicle kilometers			
TVKM (km)	9.693×10^9	9.469×10^9	–2.309

6. Conclusion

An essential requirement of successful business practice, be it in the private or public sector, is "to know your business." In practical terms, this translates into a need to both understand the market in which a business supplies goods and services (i.e. the nature of demand and competition), and the way in which the business uses its factor inputs (i.e. capital, labor, materials, energy, management expertise, technology) to produce goods and services that are consumed by the market. Although this process of good practice is undeniable, its execution is often inadequate, resulting in the inefficient and ineffective use of inputs in the provision of produced services as well as a mismatch between produced services and those demanded in the market (i.e. excess demand or excess supply).

Organizations in both the private and public sectors compile quantitative measures of the relationship between inputs and intermediate (i.e. produced) and final (i.e. consumed) outputs to "indicate" the performance of an enterprise. These performance indicators are typically ratios of one input and one output (e.g. amount of labor per vehicle kilometer of produced service) and ratios of two inputs (e.g. full-time mechanics per bus). Such indicators are partial measures of performance. Although useful information in its own right, especially for managers responsible for particular sections of a business, partial indicators are inadequate in describing the overall performance of an enterprise, and links to the broader STO setting within which much transport activity is increasingly required to be evaluated.

The purpose of this chapter has been to introduce the notion of a performance evaluation framework and to show how one can use this as a framework within which to quantify meaningful measures of performance that satisfy both holistic and more specialized objectives.

References

Forsyth, P. (2000) "Models of airport performance," in: D.A. Hensher and K.J. Button, eds, *Handbook of transport modelling*. Oxford: Pergamon Press.

Freiberg, K. and J. Freiberg (1996) *Nuts! Southwest Airlines' crazy recipe for business and personnal success*. Austin: Bard Press.

Hensher, D.A. (2001a) *Establishing value for money in incentive-driven quality contracts: the bus reform agenda in New South Wales*. Sydney: Institute of Transport Studies, University of Sydney.

Hensher, D.A. (2001b) "A systematic assessment of the environmental impacts of transport policy: an end use perspective," *Environmental and Resource Economics*, Special Issue.

Hensher, D.A. and K.J. Button, eds (2000) *Handbook of transport modelling*. Oxford: Pergamon Press.

Marshall, S. (2003) "The street: integrating transport and urban environment," in: D.A. Hensher and K.J. Button, eds, *Handbook of transport and the environment*. Oxford: Elsevier.

Nash, C.A. (2000) "Modelling performance – rail," in: D.A. Hensher and K.J. Button, eds, *Handbook of transport modelling*. Oxford: Pergamon Press.

Oum, T.H., C. Yu and M.Z.F. Li (2000) "Modelling performance: measuring and comparing unit cost competitiveness of airlines," in: D.A. Hensher and K.J. Button, eds, *Handbook of transport modelling*. Oxford: Pergamon Press.

Porter, M.E. (1985) *Competitive Advantage*. New York: Free Press.

Rouse, P. and M. Putterill (2000) Highway performance," in: D.A. Hensher and K.J. Button, eds, *Handbook of transport modelling*. Oxford: Pergamon Press.

van de Velde, D.M. (1999) "Organisational forms and entrepreneurship in public transport," *Transport Policy*, 6:147–157.

van de Velde, D.M. (2001) "The evolution of organisational forms in European public transport," in: *International Conference on Competition and Ownership of Land Passenger Transport*. Molde.

Viegas, J. and R. Macário (2001) "Pitfalls of competitive tendering in urban public transport," in: *UITP Conference*. London.

Waters, W.G. (2000) "Productivity measurement," in: D.A. Hensher and K.J. Button, eds, *Handbook of transport modelling*. Oxford: Pergamon Press.

This Page Intentionally Left Blank

Chapter 7

PRIVATIZATION IN TRANSPORT

CHRIS NASH
University of Leeds

1. Introduction

By the 1970s, in much of the world, transport had become a largely public sector activity. Roads, railways, airports, and many ports were publicly owned. Rail, most bus and coach, and many air services were provided by public sector operators. The one big exception to the rule was road haulage, which was often provided by a mix of large and small private sector companies, and where in any case much activity was undertaken by private companies on their own account.

In the 1980s, transport policy moved progressively in the direction of the market approach, and widespread privatization of transport operations and sometimes even infrastructure took place. One of the pioneers in this development was the UK, under the government of Margaret Thatcher. Significant events were:

- deregulation of express coach services in 1980;
- deregulation and privatization of most local bus services following the 1985 Transport Act;
- privatization of the major airports, ports, and of British Airways in the mid-1980s;
- privatization of rail infrastructure and operations under the 1993 Transport Act;
- removal of direct responsibility for building and maintaining roads from the UK Department of Transport to the Highways Agency;
- examination of a number of options for achieving private investment in roads, culminating in the "shadow toll" system of paying for DBFO (design, build, finance, and operate) road schemes, whereby road users are not charged a direct toll, but private road builders are paid by central government according to the level of use of the road.

The UK was not alone in these developments. Railways were privatized in Japan, although the reform of 1987 was, strictly speaking, simply restructuring

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

into a number of separate companies; sales of shares in three of these companies took place over 1992–1997. In North America those parts of the main line rail network that were in public ownership (namely Conrail and Canadian National) were privatized, as were the railways in New Zealand and parts of Australia. In much of Central and South America and Africa, rail operations have been franchised to the private sector (Campos, 2000). In a number of countries, particularly in South America, bus operations have been privatized; or where this did not happen explicitly, privately operated buses or para-transit gained a large market share. Light rail systems have been built on DBFO or similar approaches in many countries. Privately owned toll roads, ports and airports have been developed throughout much of the world. Only Western Europe (excluding the UK) seems to have been immune from this trend, though even here bus operations in many countries (such as France and Spain) include extensive private sector operations, and the rail market has been opened up to a degree of competition from private sector operators, while some local rail services have been franchised out (Nash and Toner, 1999).

This chapter will consider, in turn, the arguments for and against privatization of transport operations and infrastructure, alternative ways of carrying out such privatization, and assess the results it has achieved, taking as a case study rail and bus transport in the UK, before drawing conclusions.

2. Arguments for and against privatization

Before considering the case for privatization, it is useful to ask first why it is that transport infrastructure and operations so often were provided by the public sector even in predominantly capitalist economies. Historically there seem to be a range of reasons. Strong among them seems to have been the natural monopoly argument, that economies of scale meant that it only made sense to have a single rail or road network or network of public transport services. Secondly was an argument that transport was so fundamental to all economic and social activity that it required a degree of central planning and control (in the UK this was part of the policy of the Labour government of the late 1940s to bring the "commanding heights" of the economy into public control). Perhaps this may be interpreted as a belief that transport systems have large external benefits throughout the economy; more recently, increasing attention has been paid to their external costs in the forms of pollution, congestion, and accidents.

In principle, government control may be thought to allow all decisions to be taken to give an optimal outcome in terms of the objectives of government transport policy, balancing their financial, social, and environmental effects. However, by the 1970s, discontent with this approach was growing. Firstly, there was increasing recognition that government decision-taking may not always be competent and may give undue weight to short-term political advantage rather than long-term objectives. The public choice school of thought formulated explanations as to why public ownership might systematically fail to maximize social welfare in terms of incompatibility with the incentives faced by politicians and managers. Secondly, the construction and operation of transport services may be inefficient, because the organizations involved lack strong incentives to achieve high-quality services at minimum cost. Principle-agent theory offered insight into these problems. Finally, the transport sector has heavy requirements for investment, which may place strain on public sector borrowing, particularly in circumstances where concern about inflation places heavy macroeconomic constraints on public borrowing and spending. By contrast, privatizing assets could provide a once and for all contribution to meeting the financial needs of the public sector.

It should be noted that these objectives may not necessarily be fully compatible with each other, and the form privatization takes may depend on the relative weight given to them. For instance, technical efficiency is likely to be maximized by a competitive approach, whereas revenue from asset sale is highest when the company concerned retains a monopoly.

Experience of nationalized industries in practice suggested that the lack of clear objectives, and constant political interference with decision-taking, made it difficult to measure performance and to hold management accountable for it. Partly in order to overcome these problems, transport operators were generally constituted as public corporations with a fair degree of independence from central government, and progressively with clearer objectives and a distinction between commercial and social operations. Nevertheless, many thought the problems remained (e.g. Foster, 1992).

Privatization would lead to clear and explicit objectives, in which the management of operators and infrastructure providers were motivated by profits, and politicians would need to make explicit arrangements, through regulation, taxes, or subsidy, to achieve their political and social objectives. This development alone, together with the "hard" budget constraint and takeover threats faced by a private company, would be enough to increase efficiency (Vickers and Yarrow, 1988). However, privatization has most often been accompanied by action to open up the market to competition, by removing regulatory and other barriers to entry. Competition would lead to provision of services and infrastructure at minimum cost and with maximum innovation. However, we have already commented that much of the transport sector was seen as a natural monopoly, where competition could not generally be expected, and to the extent that it occurred it could actually be harmful by losing economies of scale. An important part of the revolution was a growing belief that the natural monopoly in transport was confined to the infrastructure, and that it was perfectly possible to have competing operators providing services over the same infrastructure. That has always been the case for roads, airports, and ports, and now the same philosophy was applied in many countries to railways. Partly this philosophy was built on the new theory of contestability, which suggested that economies of scale themselves were not a barrier to effective competition. What mattered was the absence of sunk costs and barriers to entry and exit. If these were low, then even the monopoly provider of rail or public transport on a route would be forced to behave as a competitive company by the threat of "hit and run" entry. Only the infrastructure, with its huge sunk costs and difficult authorization procedures for new construction, was inevitably protected from such a threat of competition.

Infrastructure thus remained a greater challenge for would be privatizers (Estache and de Rus, 2000). If infrastructure owners were to be prevented from abusing their monopoly power, it would be necessary to regulate them, and old-style cost-plus regulation had been shown to have many of the problems of public ownership in terms of lack of incentives for efficient operation. Moreover, since in the case of a natural monopoly marginal cost is below average, efficient pricing would require subsidies. As a result, much thought was devoted to alternative ways of privatizing the provision of transport infrastructure, and indeed also those transport services that were seen as still possessing natural monopoly elements and/or requiring subsidy. One effect was the re-emergence of interest in various forms of franchising; a solution that had been put forward by Chadwick in the nineteenth century and repopularized in the literature by Demsetz (1968). Under this approach, the competition was for the market rather than in it; private competitors would place bids in terms of what they would provide and at what price for a stipulated length of time, and the public authorities would select the winning bid. While this approach avoids the problem of lack of competition in natural monopoly markets, it does however introduce a host of new problems of its own. If the franchise places the minimum of constraints upon the actions of the franchisee, they will still be free to adopt the pricing, service level, and investment decisions of a monopolist. If on the other hand the franchise takes the form of a complete contract, covering all these issues, then a public body will still be responsible for taking all major decision about what to produce at what price, with scope for poor decisions and political interference, and the freedom even to adjust to changing market conditions is very limited. Drawing up a complete contract covering all relevant dimensions of quality may be very difficult, and the franchisee has an incentive to provide the minimum level of service unless a higher level is directly recompensed by increased revenue. Moreover, there is a problem toward the end of the franchise where the franchisee has little incentive to invest in anything that will only be justified beyond the end of franchise, particularly if the stage comes where they know that they will not be awarded a new franchise period.

3. Alternative approaches to privatization

We have seen in the last section that there are a number of different approaches to privatization to be found in the transport sector. In this section we will examine the options more closely.

3.1. Simple privatization without other measures

Simple privatization comprises selling the organization without changes to its structure or environment. This may be accomplished by outright sale to another company, by a management and/or employee buy-out, or by sale of shares to the public. Simple privatization was essentially what was done in the UK with British Airways, and in New Zealand with the railways. In both cases it was thought that there was sufficient competition (from other airlines in the former case and from road transport in the latter) that monopoly power was not a problem. Provided that is the case, then there is every reason to suppose that the privatization will be successful. In the case of British Airways, competition has been strong, particularly in the face of further worldwide deregulation of airlines. In the case of New Zealand railways, the story has been more problematic. Financial pressures have led to the company seeking to cut services, at the same time as there has been strong environmental and social concern to increase both freight and suburban traffic on rail rather than road. The result is that the budget constraints have not proved as hard as they appeared, and the issue of the public sector funding, or even taking back into ownership at least the track, has re-emerged.

3.2. Privatization plus regulation

Where a natural monopoly is sold outright and no or limited competition is expected, there may be a combination of privatization plus regulation. The history of this approach to privatization is far from untroubled. For instance, in the UK, it is the approach essentially applied to the airports formerly owned by the British Airports Authority, which included all three major London airports, and to rail infrastructure in the form of Railtrack. Both cases illustrate the difficulty faced by a regulator who needs to balance the need to give users appropriate price incentives with the need to avoid excessive profits or inefficiency on the part of the regulated company, and to provide the incentive and financial means for an appropriate level of investment. In the case of British airports, it is argued that the ability to make money from property and other commercial activities at airports has required the regulator to set landing fees inefficiently low to control monopoly profits. In the case of Railtrack, the problem has been a failure of incentives for the company to invest adequately in renewals and capacity enhancement; when the full implications of the level of renewals needed became clear, the company became bankrupt (Nash, 2002).

3.3. *Privatization plus deregulation*

This approach relies on deregulation providing new competitors rather than regulation of the privatized undertaking to provide the element of competition. Of course, it is inapplicable to a true natural monopoly. Where it is most applicable is to services such as bus or air services where the monopoly is thought to be largely the result of regulatory control of entry rather than cost characteristics. Even here it is often thought necessary to promote competition by restructuring the company at privatization. For instance, in the UK, the publicly owned National Bus Company was restructured into many separate companies that were sold individually in order to promote competition. That such restructuring may be of limited benefit is illustrated by the subsequent history of bus operators in the UK; a spate of mergers and takeovers has led to the dominance of the industry by three large groups (Table 1). This was not expected given the widespread evidence of limited economies of scale in bus operations, but it appear that the benefits of common marketing, purchase of vehicles and fuel, and perhaps diminished competition have led to widespread consolidation, which the limited UK legislation on mergers has been powerless to control (Dodgson and Topham, 1988). To a degree the same has happened in the UK in the rail passenger business. The history of restructuring the rail freight business in the UK to promote competition was even less successful; although the former freight sector of British Rail was split into six different companies at privatization, five of the six were then sold to the same consortium! The problem here was that the businesses being disposed of were not financially very attractive. There is also a conflict between the desire of the government to get the best price for the assets it is selling and the desire to promote competition.

Restructuring may be seen as worthwhile even where it does not lead to direct competition. For instance, in Japan, the national rail system was divided into seven regional companies at privatization (Mizutani, 1999). Where the natural monopoly element is seen as relating to specific routes or parts of the network, rather than to the company as a whole, this approach has advantages. It permits what is known as yardstick competition, whereby companies compete for public reputation even when their services do not directly compete. It permits a diversity of approaches, which may help to identify what works and what does not, thus promoting innovation and technical change. It is also of considerable help to the regulator of a regulated industry, since they can observe and learn from the results

Table 1
Bus and rail concentration in the UK in 2000

Bus		Rail	
Company	Turnover (%)	Company	Train-km (%)
First Group	22.1	National Express	33.0
Stagecoach	16.6	Firstgroup	13.1
Arriva	15.8	Connex	13.0
Go-Ahead	7.2	Arriva	9.8
National Express	5.9	Virgin	8.9
Other	32.4	Stagecoach	8.5
		Other	7.8

Source: *TAS Bus Monitor* (2000) and *TAS Rail Monitor* (2000).

of several different companies and apply best practice in reaching decisions on prices and services.

3.4. Franchising

The franchising approach to privatization is very different from those described above. It is most relevant where it is seen as impossible or undesirable to introduce competition directly in the market, but where competition between firms for the franchise is possible. It also permits the attachment of conditions regarding prices, services, or investment to the franchise, so it is appropriate where the authorities are not happy to leave these entirely to commercial decisions, and the provision of subsidy in a way that reduces the degree to which the subsidizing authority is at the mercy of a monopoly provider. It has become the dominant method of privatizing road and rail infrastructure and rail services, including new light rail systems, as well as being used for other public transport such as buses and ferries in a number of cases.

There are a number of key decisions to be made when setting out to franchise transport infrastructure or services. These are franchise length, degree of control over prices and services, degree of transfer of risk, and nature of the bidding process (Preston et al., 2000). Each will be discussed in turn, but of course these factors are interdependent.

Franchise length

There is an important trade-off in determining franchise length. Short franchises maximize competitive pressure, but they may also lead to a limited incentive to invest, not just in physical assets but also in product development such as improved marketing, timetables and fare structures, staff training, or even simply the maintenance of the assets in good shape. In fact, the incentive is reduced to undertake any activity that involves immediate costs with benefits spread over a longer period of time than the franchise. By contrast, longer franchises improve the incentives to invest, at least in the earlier years of the franchise, but they grant the incumbent a monopoly for a substantial period of time, and by reducing the possibility for any other company to have recent experience of the franchises in question may reduce competition even when the franchises is re-let.

But the choice is not quite as clear-cut as it sounds. Firstly, the incentive to invest may be improved if the incumbent wishes to secure a further franchise period, and performance in the existing franchise is one of the criteria on which franchises are let. Secondly, this effect is stronger if there is a steady stream of franchises coming up for re-letting so that the chances of winning at least one franchise is increased and failure to win a particular one does not remove the incentive. The incentive is also improved if the asset in question may be transferred to another franchise that the company wins. Finally, if necessary, the franchising authority can secure investment by making it a condition of the franchise, or by underwriting it to ensure that the investor receives remuneration for it after the end of the current franchise (e.g. by requiring the successor franchisee to lease it). Ultimately, the franchising authority could actually retain ownership of the assets itself, although this will fail to achieve the financial benefits that the public sector often looks for from franchising.

Examples of all these approaches may be found in the transport sector. For instance, operation of regional rail passenger services in Sweden is franchised out for very short periods of 2–3 years, but the regional authority typically supplies the rolling stock and depots, while a separate infrastructure authority supplies the track. In the UK, rail services were initially franchised out for periods of 8–15 years, but investment was often part of the franchise agreement, rolling stock and infrastructure were both owned by other companies (rolling stock leasing companies and Railtrack, respectively), and the Office of Passenger Rail Franchising (later absorbed by the Strategic Rail Authority) was given powers to underwrite investment where necessary. In South America, regional rail companies have been franchised out on very long franchises of 30–40 years, with the franchisee responsible for the infrastructure and the operations. Even with long franchises, safeguards against poor performance may be built in by performance requirements and incentives, conditions relating to the condition of the assets at the end of the franchise (and backed up by a performance bond), and by break points in the franchise if performance is not adequate. Perhaps the biggest problem with long franchises is the need to foresee likely developments over the period and to make provision for them either directly in the franchise agreement or by permitting renegotiation on stipulated terms. Obviously the more the franchise agreement stipulates matters such as service levels, charges, and investment, the more likely it is that it will need renegotiation in the light of changing market conditions. Longer franchises are generally more appropriate where the public authority does not wish to exert close control over such matters.

Degree of public control

We have already explained that the degree of public control is an important factor in designing a franchise agreement. The absence of public control may maximize the price the public authority receives for the franchises but will lead to monopoly exploitation of the consumers of the product. Stipulation of maximum prices and minimum levels of service may restrict such exploitation, but will reduce the price raised by the franchise. Ultimately, if the public authority requires the franchisee to supply a set of services that in total is unprofitable, then the franchise agreement will be for a given level of subsidy rather than a given payment. The widespread presence of natural monopoly and of external costs and benefits in the transport sector means that usually public authorities will want to retain a considerable amount of control.

Transfer of risk

If the public authority simply invites bids for a franchise while imposing no conditions on fares or services and requiring the franchisee to take complete responsibility for provision of the necessary assets, then in principle the transfer of risk is complete. For a new DBFO project, this will include all risks associated with designing, building, financing and operating the project, including both cost and revenue risk. Numerous other options, e.g. BOOT (build, own, operate, and transfer) and DBOT (design, build, operate, and transfer), exist where the contractor is not responsible for all these functions. Complete transfer of risk, even when intended, may not prove to be the case however if the franchisee gets into financial difficulties; usually, public authorities will be reluctant to see the complete cessation of transport services or availability of infrastructure in the event of bankruptcy of the franchisee, but it may be difficult to arrange alternative provision in the short run, and also costly to refranchise after a financial failure, as this will certainly not be as attractive to new franchisees as if the incumbent had been a success. The result is that it may be tempting, or even essential, that the franchising authority renegotiates the franchise to keep it running in the short term. Such renegotiation has been a feature of rail franchises in many countries, including the UK and Australia, and also of a number of privately built toll roads (e.g. in Hungary). However, the more it becomes accepted that franchises will be open to renegotiation once awarded, the more encouragement is given to opportunistic behavior in which companies base their bids on an assumption that they will be able to renegotiate the terms after the franchise has been awarded.

In the early days of privatization in the UK, the UK Treasury sought to achieve the maximum transfer of risk in all cases. This led to some extraordinary arrangements to transfer revenue risk. For instance, the approach to private maintenance and construction of roads that were free of tolls was to pay a shadow toll based on the volume of traffic, in order to transfer revenue risk even in a situation where there was no direct revenue! In fact, growth of traffic on the road was likely to be much more due to government policy and external events than to the actions of the franchisee, and to the extent that the latter could take action to encourage traffic, for instance by encouraging development at motorway interchanges, this was directly counter to other aspects of public policy. Later road franchises have shifted the emphasis to payment according to quality of service rather than traffic volume.

A more sensible approach is to allocate the risk according to who is best able to control it. Thus, to the extent that revenue risk is associated with the actions of the franchisee, it is sensible to transfer that risk, but to the extent that it is associated with government policy or actions (e.g. land use planning, securing of planning permission for developments, fuel taxes, or development of competing roads or modes of transport) the government will have to pay dearly to transfer the risk to the private sector. Where the franchising authority controls fares and services, the ability of the operator to control revenue risk is reduced, and the likelihood that it will require considerable compensation to bear those risks increased. Cost side risks, however, are more due to the operator, although the government may choose to bear the risks of general movements in fuel prices or wage rates over which the operator has no control. For instance, in the UK, passenger train franchisees are protected against changes in the charges levied by the infrastructure provider; these are passed straight through into the franchise payments.

The bidding process

Interrelated with many of the above factors is the actual process the bidding takes. If the franchising authority wishes to exert complete control over fares and levels of service, then the bid will of necessity take the form of a level of cash payment, either that the franchisee is willing to pay or that it will require to operate the franchise. If on the other hand the franchising body is most concerned to achieve a given financial outcome, then the bids may be in the form of the level of charges and/or service that the franchisee will provide. For instance, in the UK, rail passenger franchising initially was based almost entirely on the financial requirement to operate a given level of services with a given degree of regulation of fares, whereas road infrastructure franchises were awarded on the shadow toll they would charge (although in this case, because the shadow toll is paid by the public sector, the difference is more apparent than real).

In practice, the position is rarely so clear-cut. Franchising authorities will always want to satisfy themselves that the body they are awarding the franchise to has a reasonable prospect of fulfilling it, given the costs and other difficulties that will arise if it fails. Thus, there is usually a pre-qualification procedure so that only serious contenders are allowed to bid. Even then the reputation of the bidder may be taken into account, as may what it offers in terms of price, quality of service, and investment, even where the primary criterion for award of the franchise is financial. It will be obvious, however, that the more criteria there are to be taken into account, the more difficult the decision will be. It will also be more difficult to ensure that the industry sees the award as fair and transparent, and if the industry loses faith in this it may make bidding less competitive. Ultimately, corruption may be a fear.

Summary

Thus, it will be seen that a decision as to whether franchising is an appropriate way forward is not an easy one. Many different approaches to franchising are possible, and which is the most appropriate depends very much on individual circumstances. One factor is clear, however: franchising is only worthwhile if a reasonable level of competition for the franchises can be sustained. This is an important point both for the decision whether to franchise and for the design of the franchising scheme. For instance, a scheme that works in terms of a single large long-term franchise will not maintain a level of activity in the market that will maintain a number of credible experienced bidders for future franchises. Other things being equal, a series of smaller, shorter franchises will maintain competition more effectively, although of course there may be other costs in terms of loss of economies of scale and disincentives to invest, as discussed above. In the UK, in London, where the number of bidders for the operation of bus services has been a concern, one factor that influences the size and geographical coverage of invitations to tender is the wish to encourage new operators, who are often small and confined to a particular locality, to bid.

4. The success of privatization – the UK experience

As explained above, the UK has undertaken a more complete privatization of its transport sector than almost any other country. In particular, it has almost completely privatized its rail and bus industries. Moreover, these sectors include examples of privatization with deregulation, privatization with regulation, and franchising. As a case study we will consider in turn the experience of the bus and national rail industries in the UK, and finally comment briefly on the different approach taken in the case of the London Underground metro system.

Privatization of buses outside of London went hand-in-hand with deregulation, and it is impossible to separate out the effects of the two. As from October 1986, the former entry controls on the industry were abolished, and anyone who satisfied the conditions for an operator's license was free to offer whatever bus services they chose at whatever fares they chose, subject to giving 42 days notice to the Traffic Commissioners. If local authorities wished to supplement these services they had to do so by means of competitive tendering. The former state-owned bus company, the National Bus Company, was split up and privatized; many local authority owned bus companies followed suit, and those that did not were legally required to be set up as arms length commercial organizations with no subsidies other than those provided by means of concessionary fare arrangements (which applied to all operators) and competitive tendering.

In London a different approach was adopted. Over a period of time, London Transport buses were split up into separate companies and privatized. Instead of deregulation, London Transport moved toward a system of universal competitive tendering, in which it controlled routes, fares, and timetables, but the actual operations were privatized. Initially it used gross cost contracts, in which it took the revenue risk. There were a number of reasons for this. Firstly, since it controlled fares and service levels, operators had limited scope to influence revenue, and some of the things they could do, such as racing to operate directly in front of other companies' buses, were undesirable. Moreover, a large amount of revenue was from travelcard-type tickets (allowing travel on any bus, regardless of company), which had to be apportioned between operators anyway. Secondly, they found that small operators were very reluctant to bid for contracts where they bore revenue risk, so they achieved a greater level of competition by bearing it themselves. Incentives to perform were built in via measurement of key parameters such as punctuality and lost kilometerage. Nevertheless, after pressure from the government, they gradually moved toward net cost contracts whereby the operator bore the revenue risk.

Mackie et al. (1995) examined the experience of the first 10 years of the new system, and this is summarized in Table 2. It is seen that in both London and the rest of the country a major reduction in costs and subsidies was achieved, and fares and service levels both rose. But in London these changes were associated with an increase in bus patronage; elsewhere there was a continuing sharp fall. While there are many differences between London and the rest of the UK, such as congestion, parking difficulties, and levels of car ownership, it is clear that London has managed to retain its bus patronage much better than other cities. The suspicion remains that this is partly due to the regulatory environment.

As mentioned above, one concern with bus privatization in the UK is the maintenance of adequate levels of competition either on the street or for competitive tenders. It appears that some of the most acute barriers to entry come from the difficulty of obtaining suitable depots from which to operate in large

Table 2
Trends in the bus industry in the UK: 1985/1986 to1993/1994

Area	Cost/ bus-km (a)	Real subsidies (b)	Bus-km	Passenger fares	Passenger journeys	Cost/ passenger journey
London	–35.1	–47.0	+25.6	+29.2	–3.0	–20.0
English metropolitan	–46.5	–42.5	+20.5	+48.9	–35.5	0.0
English shires	–36.8	–20.7	+24.5	+8.8	–20.2	0.0
Scotland	–40.0	–30.5	+26.6	+2.3	–21.6	–2.0
Wales	–46.8	–33.3	+33.7	NA	–20.3	–9.5
Total	–39.8	–38.3	+24.2	+19.2	–22.5	–3.8
Total excluding London	–41.9	–34.9	+24.0	+17.4	–27.4	0.0

Notes: (a) excluding depreciation; (b) loss-making services plus concessionary travel reimbursement.

cities. At privatization, depots, and many bus stations, were sold with the companies. Often these were subsequently sold for property development. While this might have been an efficient outcome, it has hampered the level of competition. It might have been better to have separated out this fundamental bus infrastructure and leased it, in order to maintain a sufficient infrastructure to support the necessary number of companies.

When it came to main-line rail services, the approach taken to privatization was totally different from that taken to buses (Nash, 2002). Separation of infrastructure from operations was immediately followed by outright privatization of the infrastructure manager (Railtrack) and the freight operators, and by franchising of passenger services for periods of 7–15 years. Competition was introduced along the supply chain by means of completely contracted-out infrastructure maintenance and renewal, rolling stock supply and maintenance, and many other activities to separately privatized parts of British Rail and other private companies (Nash, 2000). The natural monopoly part of the business was considered to be Railtrack, and its charges and services were controlled by an independent regulator. Thus, rail privatization comprised a mixture of restructuring and outright privatization, privatization with regulation, and franchising.

Until the year 2000 there was a widespread view that, although problems had been found, the UK experience was overall positive. Passenger traffic had risen to its highest levels since before the major cuts to the rail network under Dr Beeching in the 1960s (Table 3). Freight traffic had also started to grow, although not to the same extent. While the form of privatization had led inevitably to a big increase in government grants in the short term, as assets were sold off and the train-

Table 3
Rail traffic on the UK national rail network

Period	Total passenger-km	Total freight tonne-km	Revenue support grants to domestic passenger services (UK £ × million)	
			From central government	From PTE (a)
1986–1987	30.8	16.6	755	70
1987–1988	32.4	17.5	796	68
1988–1989	34.3	18.1	551	70
1989–1990	33.3	16.7	479	84
1990–1991	33.2	16.0	637	115
1991–1992	32.5	15.3	902	120
1992–1993	31.7	15.5	1194	107
1993–1994	30.4	13.8	926	166
1994–1995	28.7	13.0	1815	346
1995–1996	30.0	13.3	1712	362
1996–1997	32.1	15.1	1809	291
1997–1998	34.7	16.9	1429	375
1998–1999	36.3	17.4	1196	337
1999–2000	38.3	18.4	1031	312
2000–2001	38.2	18.1	847	283
2001–2002	39.1	19.4	731 (b)	306

Source: Strategic Rail Authority, *National Rail Trends 2002–03*, quarter 3.
Notes: (a) PTE grants are grants paid by Passenger Transport Executives for local rail passenger services in the main conurbations excluding London; (b) in 2001–2002, a Direct Rail Support grant to Railtrack to contribute to infrastructure costs was introduced; in that year it totaled £684 million.

operating companies had to pay commercial rates for using them, subsidies (excluding receipts from the sale of assets) were rapidly declining.

The problems that were emerging with the system were essentially threefold. Firstly, some aspects of quality of service were a cause for concern; punctuality and reliability were declining, and despite a high degree of public control over train service levels through the franchising process, through ticketing and information designed to maintain network benefits, there were problems in achieving well-integrated timetables and fares in a system with many different train operators.

The second big problem was in the field of investment. It was argued that the level of investment, particularly by Railtrack, although increasing, remained inadequate and was not providing sufficient capacity for a growing industry. Moreover, where train operators had ordered new rolling stock there were long delays in getting it into service, partly due to problems with the manufacturers, but more especially due to difficulties in getting it through the Railtrack safety regime.

The third problem was the financial difficulties that some of the train operators, particularly those in the less profitable parts of the industry that had built their franchises bids around big reductions in operating costs, were having. The result was doubt as to whether the reduction in subsidy was fully sustainable or whether in fact some franchisees would go out of business, meaning that some franchises would need to be re-let, probably with higher levels of subsidy.

These problems might in part be seen as inevitable problems of a system that combines separation of infrastructure from operations with franchising. However, solutions to them were pursued through the creation of a new public sector body, the Strategic Rail Authority, to take the lead in planning and investment, restructuring of rail access charges and conditions to improve the incentives to Railtrack, restructuring of the franchises to provide a smaller number of much longer-term franchises, and injection of more public money into both investment and subsidy in the short term with the continued aim of improving long-term financial performance as well as quality and capacity.

The latter part of the year 2000 saw the rail industry in the UK in crisis. Following the Hatfield rail accident, gauge corner cracking was found extensively in rails all over the system. Railtrack imposed severe speed restrictions, which greatly increased journey times and led to reduced frequencies and poor reliability. The problems were compounded by other factors, including severe flooding and shortages of rolling stock. There were serious falls in traffic and profitability. Under pressure from the resulting increase in spending on maintenance and renewals, as well as claims for compensation from train-operating companies, Railtrack was declared bankrupt. Its successor, Network Rail, is a company that – although nominally still private sector – does not have shareholders or a profit motive; rather, it is controlled by a membership comprising the Strategic Rail Authority, the train-operating companies, and rail users.

Some (including the then chief executive of Railtrack) assert that these difficulties were a likely result of a fragmented structure to the industry, with separation of infrastructure from operations, a privatized monopoly infrastructure provider, and extensive subcontracting of maintenance and track renewals. The evidence on this is still not clear, although certainly the evidence produced so far does point to severe weaknesses in the management of some of those interfaces, and particularly that between Railtrack and its maintenance contractors. Moreover, it has become clear that the philosophy that government only has a role in securing provision of services, and that the infrastructure manager can be left to respond to market forces, has not worked; the Strategic Rail Authority has taken an increasingly stronger role in planning and financing infrastructure investment and use.

There is evidence that franchising did indeed produce a significant improvement in efficiency within the train-operating companies over the period in question

(Nash, 2002), although British Rail had achieved a similar improvement in the years prior to privatization as well. Given the increases in both passenger and freight traffic, it seems reasonable to conclude that overall franchising and privatization of the freight operators was a success (Fowkes and Nash, 2004). The big problems predictably surrounded the natural monopoly element: Railtrack. No other European railway has privatized its rail infrastructure, and the system in Sweden, which has many similarities to that in the UK except that the infrastructure has remained in public hands, has had far fewer problems than the UK.

The London Underground metro system was not included in this privatization process, and indeed an element of privatization has only just been introduced. This takes the form of yet another variant on the list. Train operations on the London Underground remain in public hands, but the infrastructure has been privatized on long-term franchises, with the new companies committed to maintaining it and investing in new signaling, stations, and other measures to improve capacity and performance. The motivation has been to improve efficiency and to obtain private funding for investment; the main controversy is over whether the approach is cost-effective. Private companies will generally require a higher rate of return than the rate of interest paid on public borrowing, and, given the experience of Railtrack, perhaps a sizable risk premium as well. It has been argued that the result is to make this approach much more expensive than simply funding the investment through public borrowing (Glaister et al., 1999).

5. Conclusions

We have seen above that there are arguments for and against privatization in transport. There are also many different ways of carrying it out. Both the case for privatization and the best way of carrying it out depend on circumstances. It is thus difficult to reach any general conclusions.

In sectors that are potentially competitive, and where no strong case exists for public control of services and charges, it is hard to see any particular reason for public ownership to exist. Arguably these conditions hold in general in air, water, and road freight transport operations. In all cases there are concerns about the external effects of congestion, pollution, and accidents, but these may be better dealt with by regulation and taxation than by public ownership.

For rail and bus services, franchising of passenger services and outright privatization of freight appear to have achieved both efficiency improvements and improved services. Outright privatization accompanied by total deregulation is a less common way forward, and in the UK bus industry it has achieved mixed results; costs have been greatly reduced but demand has continued to decline.

Privatization of transport infrastructure is more problematic. While sometimes there may be adequate competition from other facilities, especially in the case of seaports and airports, this is frequently not the case, and some type of regulation or franchising is likely to be necessary. While there have been success stories concerning privatization, there have also been a number of failures, of which Railtrack in the UK is perhaps the most spectacular.

References

Campos, J. (2000) "Rail reform in Latin America: what could Europe learn/avoid," in: E. Niskanen and C. Nash, eds, *Proceedings of the Seminar on Rail Infrastructure Charging*. Helsinki.

Demsetz, H. (1968) "Why regulate utilities?" *Journal of Law and Economics*, 11:55–65.

Dodgson, J.S. and N. Topham, eds (1988) *Bus deregulation and privatisation: an international perspective*. Aldershot: Gower.

Estache, A. and G. de Rus (2000) *Privatisation and regulation of transport infrastructure. Guidelines for policy makers and regulators.* Washington, DC: World Bank.

Foster, C.D. (1992) *Privatisation, public ownership and the regulation of natural monopoly.* Oxford: Blackwell.

Fowkes, A.S. and C.A. Nash (2004) *Rail privatisation in Britain – lessons for the rail freight industry.* Paris: ECMT.

Glaister, S., R. Scanlon and T. Travers (1999) *The way out. An alternative approach to the future of the Underground. LSE London Discussion Paper*, No. 1. London: London School of Economics.

Mackie, P., J. Preston and C. Nash (1995) "Bus deregulation ten years on," *Transport Reviews*, 15:229–51.

Mizutani, F. (1999) "An assessment of the Japan railway companies since privatisation: performance, local rail service and debts," *Transport Reviews*, 19:117–139.

Nash, C.A. (2000) "Privatisation and deregulation in railways – an assessment of the British approach," in: B. Bradshaw and H. Lawton Smith, eds, *Privatisation and deregulation of transport*. Basingstoke: MacMillan.

Nash, C.A. (2002) "Regulatory reform in rail transport – the UK experience," *Swedish Economic Policy Review*, 9:287–288.

Nash, C.A. and J.P. Toner (1999) "Competition in the railway industry," *Journal of Competition Law and Policy*, 1:197–227.

Preston, J.M., G. Whelan, C. Nash and M. Wardman (2000) "The franchising of passenger rail services in Britain," *International Review of Applied Economics*, 14:99–112.

Vickers, J. and G. Yarrow (1988) *Privatisation – an economic analysis*. Cambridge: MIT Press.

This Page Intentionally Left Blank

Chapter 8

COORDINATION, INTEGRATION, AND TRANSPORT REGULATION

DIDIER M. VAN DE VELDE
Erasmus University, Rotterdam

1. Introduction

A number of characteristics of the functioning of transport markets have led through time to a call for "coordination." The topic has been a much-debated issue for several decades. The literature from the 1930s until the 1970s paid considerable attention to the topic and generated numerous definitions of coordination. Lack of clarity remained a problem throughout the whole period during which coordination policy was fashionable, as exemplified by Peterson (1930), Tissot van Patot (1938), and Ponsonby (1969). The main elements of the concept are presented here, starting with the definitions suggested by Peterson:

- Coordination is the assignment, by whatever means, of a given facility to those transport tasks that it can perform better than other facilities, under conditions that will ensure its fullest development after assignment to the ideal tasks.
- Coordination is the creation, by any means, of effective joint services by agencies that are directly complementary.
- Coordination is the attainment of a compromise between monopoly and competition that will ensure the continuance of essential agencies, maintain the maximum variety of service, eliminate undue waste, and preserve effective incentives to improvement.
- Coordination is the avoidance of duplication through subordination of rival agencies.

These definitions are not necessarily inconsistent, but they involve radical differences in emphasis, and suggest divergent programs of execution.

These definitions relate to topics of continuing relevance in the transport world, even if their treatment has substantially evolved over the course of years. The primary interest is how to secure the public interest. This was originally defined as an avoidance of inefficiencies, such as overlapping or duplication of infrastructure

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

and services that would ultimately cause a financial burden for society. The challenge was to ensure a division of tasks between transport modes according to their inherent advantages, and under prices reflecting the social costs of the provision of the services.

Essential issues of the debate are:

- How can it be ensured that modes of transport will carry the traffic for which they are best suited?
- What is the optimal balance between free competition and regulation in the transport sectors?
- What is the optimal balance between stimulation of improvements and variety and avoidance of waste?
- What are the implications of network aspects in transport provision and in transport infrastructure for the regulatory needs of the transport sectors?
- Does regulation have to vary from sector to sector?
- Does intermodal competition have to be regulated differently from intramodal competition?

Additional issues are those linked to the integration of the concept of coordination in a wider social context. This can be seen in the public transport sector, where social issues were also taken into consideration, such as the provision of a basic level of mobility to the transport disadvantaged. Other examples are the coordination with land use planning and issues pertaining to the stimulation of regional economic development by the development of transport infrastructure and services.

Section 2 presents a theoretical perspective on the coordination issue. Section 3 presents a historical perspective on this issue. Before drawing a few conclusions, the fourth section focuses on coordination in the public transport sector, as this sector continues to witness – much more than the private transport sector – a strong coordinative intervention.

2. Theoretical perspective on coordination

Micro-economists are essentially interested in the functioning of markets. In an ideal free market economy the perfect competition benchmark is that welfare is maximized. However, welfare maximization cannot be reached if certain conditions are not fulfilled; in such cases the market is said to fail. The main reason to intervene in transport markets is, from this theoretical perspective, to correct market failure. It should be mentioned, though, that numerous interventions in the transport sector are in fact unrelated to this goal. Transport regulation is often based on goals located outside the transport sector in the strict sense. Transport is often used to serve social redistributive aims by providing, for example, fare rebates or a minimum level of mobility. It is also used for specific regional economic redistributive aims through investments in specific transport infrastructure in an attempt to stimulate economic development in a disadvantaged region. These forms of intervention have little to do with dysfunctions of the transport markets *per se*.

One of the most important forms of market failure in most transport sectors is the existence of externalities in the form of pollution and congestion. Theory advises their internalization into the decision-making of the market actors. A second important form of market failure in some transport sectors is the existence of natural monopolies, as in the railway sector, for which various forms of economic regulation have been developed through time. These issues are discussed at greater length elsewhere in this book. Two other potential sources of market failure that have relevance for the topic of coordination will be discussed here. The first is the contentious issue of excessive, "cut-throat," competition. The second is the issue of the network effects in transport provision.

2.1. Market failures and cut-throat competition

The call for a "coordination" policy arises from various perceived dysfunctions of the transport markets, the relevance of which depends on the transport sector considered. These dysfunctions have received many names. "Wasteful competition" (Foster, 1963), "excessive competition," "cut-throat competition," or "ruinous competition" (Willeke, 1977) are all terms that can be found in the literature on transport markets.

Part of the argument for coordination relates to perceived idiosyncrasies of some transport sectors such as an inelastic supply in face of a more volatile demand. Ruinous competition is commonly presented as a capacity problem linked to an inelastic supply, combined with a pricing problem linked to the difficulty to relate prices to costs due to a number of factors such as high fixed costs, joint costs, or a highly variable capacity utilization. These characteristics are then seen to lead – especially in periods of declining demand – to a situation where prices tend to drop to rather low variable costs, while capacity adaptation does not take place quickly due to the long life span of some transport vehicles and infrastructure. It can also lead to a situation where insufficient provision for vehicle replacement takes place.

The proponents of the need for coordination put this reasoning forward to illustrate the need for state intervention in the form of a limitation or abolition of the free market in the transport sector considered. This argument was most powerful in the 1930s, a period of depression. This reasoning has lost most of its proponents during the last two or three decades, with the success of deregulation and liberalization in several transport markets.

The traditional coordination argument also sees in intermodal competition another potential source of waste, and, consequently, a reason for intervention. This argument came about with the appearance of competing transport modes in the 1920s, characterized by different infrastructure regimes, such as buses and lorries entering into competition with trains. Much of the literature on coordination is devoted to this issue (e.g. Mance, 1941; Foster, 1963). The attractiveness of the railway system grows with the amount of traffic it carries, as fixed costs are spread over a higher load and/or a higher service frequency can be offered. The increasing bus and lorry competition took away loads from rail, providing perhaps direct advantages (price and route) to their own customers, but at what was perceived to be a disadvantage for the transport system as a whole. Willeke (1977), among others, analyzing the claim of the existence of ruinous competition between road and rail, concluded that there was no argument for state intervention in this case, and called for deregulation. Much of the argument is linked to the distorting effects of the historical and non-coordinated regulation of the various transport modes. The current policy of the EU since the Green Paper on fair and efficient pricing in transport (European Commission, 1995) now focuses on fairer competition between transport modes.

The existence of high fixed costs due to infrastructure is in most transport sectors a third issue of concern. The classic example is that of parallel railway lines, the building of which is rarely desirable due to the existence of economies of scale (the capacity of a four-track section being higher than the total capacity of two double-track sections). Coordination of infrastructure investments does not necessarily occur when left to the free initiative of the private sector under circumstances where the components of parallel infrastructure can be operated profitably. Proponents of the coordination argument see in transport infrastructure a case for state intervention in planning. Much of this remains undisputed, leaving transport infrastructure investment decisions to be taken, or at least coordinated, by the state. The case for international coordination has even been strengthened with the development of the EU, and a trans-European network program is now being developed to compensate for the lack of connection between national infrastructure programs.

2.2. *Market failures and networks*

A potential source of market failure in the transport sectors remains insufficiently covered: the likely existence of network effects in transport sectors. This issue applies particularly to scheduled transport services. Network economics is potentially an important field of theoretical research when analyzing issues relating to coordination. We have to distinguish between network effects for so-called "scheduled transport services" and network effects as analyzed in the general utilities literature (water, gas, and electricity distribution), which relate to infrastructure networks. Unfortunately, the analysis of the network economics of scheduled transport services is not sufficiently advanced, especially for local public transport and passenger rail networks. Sources within airline economics may provide additional information, but airline networks are substantially simpler, as competition takes place on point-to-point services, or bundles of such services in the context of hub-and-spoke networks. In public transport (both urban services and railway services) the analysis is considerably more complex as a single vehicle run provides services on numerous pairs of origins and destinations.

According to the literature on network economics (Economides, 1996), the main reason for the appearance of network externalities is the complementarity between the components of a network. A condition for complementarity is compatibility.

Complementarity

A direct network effect is the concept that the product's value for customers depends on the number of customers already using the products provided by the network. In a telephone network, the value of a connection for a subscriber – and therefore of the network as a whole – is dependent upon the number of other subscribers. Everybody in the network gains when more connections are added. The parallel in public transport is that passengers often consume connecting services to get to their destinations. Separate transport services within a network often have complementary demands. The direct network effect – a consequence of complementarity – is that service improvements in one segment (e.g. a higher frequency) will lead to an increase in demand for services in other segments used in combination with that segment.

A larger group of public transport users gives rise, in turn, to indirect network effects through the resulting reduced generalized costs for existing users due to higher frequencies or larger number of destinations served. This is also known as the Mohring effect (1972), according to which additional passengers on a link will in due course be followed by increases in frequency, which will attract additional ridership as higher frequencies mean higher service quality through reduced user costs linked with transfer time and excess waiting time at the destination. Some studies related to the Mohring effect emphasize its welfare economics aspects, and use it as an argument for subsidization.

Compatibility

A condition for complementarity between services is compatibility, and thus coordination becomes an issue. There are two parts to this issue of compatibility: the cost and revenue effects of coordination for transport operators, and the user cost effects of non-coordination for passengers. Compatibility is an absolute concept in telephony: communication is or is not possible. It is, however, a more graduated concept in passenger or freight transport: interchanges between public transport services are always possible, but the user costs of realizing such interchanges vary substantially. Relevant elements of user costs are the costs of searching for information on complementary services, the additional effort linked with the purchase of connecting tickets, and the disutility (i.e. time cost) of the excess interchange time when services are not well coordinated in planning or in realization (such as in the case of disruptions). From the passengers' perspective, optimal coordination leads to a reduction in generalized costs and makes the coordinated system (in terms of standards, information, billing, interchange, etc.) more attractive to customers than its non-coordinated parts.

A difficult question in practice is that of the optimal level of coordination. There is a tendency in the public transport world to value coordination above all else. A tendency noted early by Peterson (1930) relating to transport in general. Too much coordination may lead to passenger cost increases through longer travel times due to exaggerated bundling and additional transfers that are not compensated for by shorter interchange times. A proper recognition of the total social costs and benefits of coordinative measures is needed to identify a proper level of coordination. Coordination should be pursued up to the level at which the marginal social costs of coordination start to exceed the marginal social benefits.

An additional aspect is that the perceived quality of a chain of services will depend upon the quality of the weakest service along the chain. An important question is then whether independent companies operating on a pure profit-maximizing basis would take all of these aspects into consideration when determining the amount of compatibility (i.e. coordination) they want to achieve. This is an important issue in economic literature, but the implications for coordination in transport, and in public transport in particular, are unclear and underdeveloped. A number of conditions will determine whether firms strive toward compatibility when providing complementary services. Unfortunately, nothing can be said *a priori* on the location of the borderline between beneficial autonomous coordination and lack of incentive. The reason for this is that many situations can arise as to the costs to achieve compatibility, their repartition among the companies involved, and the size and division of the resulting additional revenues among participants. This may lead to low incentives to invest in compatibility with, furthermore, the danger of free-riding. In a number of cases, financial transfers between operators providing complementary services could appear autonomously, such as to realize coordination to their mutual benefit. But nothing can be said about the social optimality of such cooperation as it is based on a pure profit-maximizing approach, forgetting about the wider welfare implications. If operations are not profitable but subsidized – and this is the common situation in most of public transport – then incentives for cooperation may be further weakened. While coordination may result in additional ridership, in most cases this will also result in an additional need for subsidization even if the marginal customers require less subsidies than the existing customers. In this situation, the commercial incentives to coordinate services will depend not only upon the relative size of the additional costs and revenues of coordination but also on the details of the subsidizing arrangements, especially whether these allow subsidy growth.

Competition versus monopoly

The discussion so far assumes static conditions or that companies have their own territories. To be complete, however, competition has to be considered. The incentives to invest in compatibility (such as information, timetabling, and fares in public transport) can, in the presence of competition, become even lower. Compatibility becomes a strategic consideration in such a case. The stronger firm is likely to dislike compatibility, whereas the weaker one prefers it. Examples of such behavior can be found in the UK deregulated bus market, where dominant firms have exited from integrated ticketing systems.

The provision of all services by monopoly could solve, in principle, the network externalities problem. A recommendation for such an organizational form can often be heard in the public transport world. It was tried in the UK after the Second World War until 1963, with the creation of the British Transport Commission regrouping rail and road transport, the London passenger transport service, canals, ports, shipping, hotel and catering services, and travel agencies. A monopolistic organizational form is still common for most public transport in European cities. The current advice to introduce area-wide franchises by competitive tendering only seems to bring competition to this sector. In effect, central coordination and monopolistic provision are maintained by this new regime, while the competitive tendering process only solves the productive efficiency issues. A fundamental issue here is that of the planning efficiency in time of such monopolies and consequently of the design of appropriate regulatory measures to control them. The trade-off is between the perceived integration benefits of monopolistic planning versus the potential innovative gains that could be generated by the existence of operators in direct competition; this issue was considered in Peterson's definition of coordination dating from the 1930s.

3. Historical perspective on the coordination policy

Regulatory intervention in the transport sector varies considerably over time. This section provides a short overview of the main phases in transport policy over the last two centuries, after making a few remarks on the main factors responsible for changes and for differences between countries.

3.1. The shifting preferences for coordination instruments

A historical study of state intervention in the transport sectors reveals that the extent of intervention and the choice of instruments vary considerably over time. As far as coordination is concerned, some periods favored coordination by monopoly or strict regulation while others favored coordination by deregulated markets. The main aim of transport policy throughout the twentieth century remained to ensure that each means of transport fulfils the task for which it is best equipped, taking into account its technical and economic characteristics. What changed were the perceptions and possibilities as to the instruments that would facilitate the realization of that aim.

Two main factors are responsible for the changes in the form of the intervention over time. The first is the development of new transport techniques, which results in changes in the potential for intermodal competition and requires a rethinking of the intervention. The other factor is the changing collective preferences in terms of regulatory mechanisms. These preferences are influenced by prevailing social attitudes, by theoretical developments – such as the theory of contestable markets – and by successive phases of economic growth and recession.

Furthermore, differences in points of view and actions in terms of transport coordination can be observed from country to country. This is most visible when contemplating the differences that existed between western and eastern Europe during the Communist era. But it is more interesting to look at the substantial differences that existed and continue to exist within western Europe. Local economic circumstances and geographical circumstances have implications on potential intermodal competition. For example, the role played by inland waterways in the Netherlands created a different regulatory situation compared with Switzerland, where this modality has only limited relevance. The size of a country and the corresponding relevance of international trade also influence the main reference framework. The Netherlands, with its small country size, is, for example, traditionally more "liberal" in terms of international (freight) transport regulation than larger countries. Another determining factor is a country's general vision of the role of central government. This is in turn determined by many historical factors. The traditional French centralism and focus on Paris is indeed a very different tradition from the practices that can be observed in other countries.

3.2. Phases

The factors influencing these shifting preferences can be illustrated by presenting the changing visions for the need for state intervention and coordination in transport and some resulting changes in transport policy in western Europe.

Button (1993) provides a brief overview of the main phases of change in the transport policy that can be observed during the twentieth century in the UK. He distinguishes an anti-monopoly phase pre-1930, an anti-competition phase until 1945, and more central control and nationalization until 1951. This is followed by the appearance of a competitive framework until 1964, controlled competition until 1974, a search for more efficiency until 1980, and the age of regulatory reform, which is still ongoing. These phases illustrate the cyclical movement (competition/centralized control) of transport regulation. Such alternating phases of stronger coordination by regulation and of *laisser-faire* can also be observed over longer periods. This will be illustrated in the rest of this section, using Dutch transport policy in the nineteenth and twentieth centuries as the predominant example (based on Kuiler, 1949; Kuiler and Verhoeff, 1973).

The nineteenth century

The period before 1813 was, in the Netherlands, characterized by the dominance of transport by the waterways, organized in the form of local monopolies guaranteed by the issue of authorizations to transport operators and subject to timetable and fare regulation. The period up to 1880 was characterized by a decreasing level of regulation, made possible by the introduction of new transport techniques to inland waterways and the appearance of the railways. Intermodal competition became possible as the growth of transport demand resulting from economic growth provided opportunities for the new modes to challenge the dominance of the waterways.

Growing regulation

Intervention remained very limited until 1920, when the appearance of yet another modality in the form of easier road transport by bus and lorry led to a further increase in supply. At the same time, a sharp economic downturn pushed transport operators into difficulties due to a drop in demand. A new period of increasing regulation resulted, as the free market was no longer perceived to be the best way to guarantee the "general interest." The word "coordination" started to be used, though with many different interpretations.

As far as the road passenger transport sector is concerned, one can observe that the bus sector was essentially functioning on the free market until the 1930s, although the dates vary from country to country in western Europe. The operators called for regulation in this period of economic decline as they started to encounter problems. These problems may well have been the result of inadequate regulation of the competing modes, such as public service obligations imposed on the railways. The new public transport "coordination" regimes regulated the market by means of "traffic commissioners," charged to defend the general interest by coordinating supply. These regimes remained in place for the next 50 years in most of western Europe.

Coordination policy

After 1940, the resulting "coordination policy" had two forms: that of the coordination between services of the same mode of transport, and that of the coordination between services of the various transport modes. Markets were strictly delineated, and the number and production capacity of suppliers strictly controlled by entry regulation. The French legislation of 1949, for example, required local authorities to choose between bus and rail services in their transport plans and to avoid parallel services in order to rationalize networks. Railway companies were either nationalized before the war, as in the Netherlands, or just after, as in the UK. Further nationalizations also took place. Fare regulations were imposed. The freedom of choice of the passenger or shipper, however, was not infringed upon, as regulation concentrated on the supply side.

Much legislative effort was made to foster coordination, though some observers at the time stated that more actually took place by private agreement than under the various legislative measures taken (Klapper, 1958). The Treaty of Rome, which created the EEC, also paid specific attention to the so-called specific aspects of the transport sector (Article 75-1), while some forms of cooperation that were seen to be beneficial, such as shipping conferences and some air fare agreements, were excluded from the anti-monopoly dispositions (Articles 81–82), partially due to international regulatory implications.

The declared aims of the European policy were equality of treatment of suppliers and consumers, financial independence of transport operators, free access to the profession, and free choice of the shipper, but also coordination of the investments in infrastructure. The general philosophy of the common transport policy was to foster the provision of transport services at the lowest costs for the collective and with the most efficient technique of transport – which is a traditional coordination aim – while respecting the equal treatment between modes and setting an infrastructure usage price system guaranteeing budgetary balance. Much of this could not be realized until much later. Gwilliam et al. (1973) elaborate further on the issue of the coordination of the investments in infrastructure. They put forward that the coordination of investments in transport infrastructure is particularly important for several reasons. These are its contributions to a better allocation of resources and a better use of the already realized investments, to the setting up of a coherent transport policy and to the creation of a united and spatially integrated EU. An instrumental aspect of the transport policy comes to the fore with the last item.

The topic of coordination remained very much in dispute throughout the whole period. Tissot van Patot (1964) reported on the divergence of opinion in the

Netherlands: academics were divided about the existence of idiosyncrasies in the transport sectors – government and operators accepted them, while shippers rejected the argument. Ponsonby (1969) took a position that was against coordination by regulation in the UK, arguing in favor of variety under the condition of absence of cross-subsidization between services. Earlier questions related to freight transport by lorries had already been raised by Foster (1963) on grounds related to the development of effective labor laws and safety regulations after the 1930s. Willeke (1977) rejected the idea of cut-throat competition as an argument for state intervention in transport in Germany.

Opinions diverge on the question of whether the coordination regime functioned properly in the public transport sector. Cross-subsidization between profitable and loss-making services supplied by one company was often an essential part of the regime. Further problems arose around the 1960s in many countries as cross-subsidization between services became insufficient to maintain the existing networks. Direct subsidization became necessary. From then on coordination took a more dirigisme stance, though with differing forms in the various countries. Regional transport plans, public transport supply norms, and various subsidization regimes came in to guide or supplant evaluation by traffic commissioners. Many operators already were in state or municipal hands, and the amalgamation of operators played a growing role. Passenger transport authorities/passenger transport executives (PTAs/PTEs) were created in metropolitan areas in the UK, Verkehrsverbünde (transport unions) saw the light of day in Germany, and the merger of regional operators continued in the Netherlands. Fare integration was developed with enthusiasm in the Netherlands, with the introduction of national ticket and fare integration in 1980 in an attempt to facilitate coordination and the abolition of parallel services. Integrated ticket and fares systems were also implemented in other countries, such as the Carte Orange in Paris or the Travelcard in London.

Deregulation, liberalization and privatization

Fundamental changes in the functioning of the transport markets appeared internationally at the end of the 1980s. In the 1990s, the ideas of deregulation, liberalization, and privatization began to gain momentum in European transport policy. The concept, that of deregulation, referred to giving transport operators more freedom in determining their course of action in the market. Liberalization involved allowing new operators to access the market. Privatization included selling off state-owned companies to the private sector. Chapter 41 of this handbook provides a detailed treatment of the development of the European transport policy.

Developments during the last 15 years of the twentieth century meant that there was a marked change in prevailing opinions on the need for state intervention in the transport sectors. Experiences in several European countries with deregulation, liberalization, and privatization led to successful air, rail, and road transport sectors, and several of these free market policies were copied across the world. Japan had also followed the coordination path, and for the Japanese national railways this meant a strong reliance on cross-subsidization between services and overconfidence of the ability of the railways to stand up to the modes of transport that had arisen to challenge its transport monopoly. This coordination policy delayed the necessary updating of the railway regulation, and the Japanese national railways were in a difficult financial situation by the 1970s and 1980s (Saito, 1989). The national railways were later split into regional companies, and successfully privatized after 1987. Intermodal competition (rail–air and rail–road) has now been introduced in Japan, very much against the original coordination ideas.

At the same time, theoretical developments, such as the theory of contestable markets (Baumol et al., 1982), provided new theoretical explanations for the ability of some markets to function with less regulation than what was until then considered to be necessary. Finally, it also became clear that intervention itself had a cost in the form of non-market failures (Wolf, 1979). The consequences of state intervention – inefficiencies of state companies not submitted to competition, excessively comfortable labor agreements, and the personal agenda of bureaucrats including budget maximizing behaviors to mention but a few – may sometimes be more costly than the market failures the intervention was supposed to resolve.

Coordination by the market has now become dominant. Airlines, road freight, rail freight, and, in many cases, taxi markets, have been deregulated and liberalized. The usage of a *tour-de-rôle* in inland shipping in northern France, Belgium, and the Netherlands, another example of coordination, was abolished at the end of the 1990s. Numerous state-owned companies are well on their way to being privatized. Intra- and intermodal competition is now accepted in most sectors. With this general background, coordination seems to have disappeared from the list of major reasons for regulating the transport industries. This does not mean, however, that coordination has completely disappeared from the agenda. The focus on an "integrated transport policy" in the current UK transport policy proves the point, and infrastructure planning and investment remains a domain for state intervention.

The coordination discussion remains alive in two sectors, though. The liner shipping industry still uses "conferences" as a coordination mechanism between scheduled carriers on a particular trade route to restrict competition among themselves by setting mutually agreed freight rates and conditions of service (Gardner, 1997), with the aims of price stabilization and avoidance of cut-throat competition. A European regulation dating back to 1986 grants this industry exemption from antitrust regulations, but the European Commission has recently started a review of this sector. The other sector where coordination continues to play an important role is local public transport. This issue will be treated at greater length in the next section.

4. Coordination of public transport services

Public transport is a field of predilection for those advocating against coordination by the market. The issue of coordination in public transport – which is commonly called "integration" in that world – is a main concern for many public transport professionals. Urban transport planners and transport operators alike regard it as a desirable thing. The concept is also used in many related fields, always under the wording of "integration": integrated public transport, integrated transport policy, integrated urban infrastructure, etc. In most cases, the meaning of the concept is assumed to be understood and shared by the members of the community to which it is addressed. A deeper analysis is likely to reveal strong divergences of opinion, continuing the confusion referred to earlier.

This section reviews forms of coordination in public transport, introducing the concept of integration. It finishes with some remarks on the evolution of these forms during the last decades and relates these to the coordination debate.

4.1. From cooperation to integration

Forms of coordination in public transport vary in intensity. The lowest form of coordination is cooperation, defined as a willingness to work together to reach common objectives. Cooperation in public transport can often be observed between suppliers in free markets, between public monopolies, or between transport authorities. For example, there are mutual agreements that allow different companies to accept each other's tickets on some deregulated bus routes in the UK. Some public operators allow the booking of a taxi at the destination through the bus driver. Mutual agreements to provide timetable information, common route maps, or coordinated departure times at interchanges exist between transport operators. Such cooperation is more likely to occur when the participants' interests are clearly non-contradictory. But traditions of distrust or rivalries between neighboring jurisdictions may severely hamper such cooperation, and in this regard the role of personal relations and effort should not be underestimated – active promotion by a coordination "champion" (a trusted individual with good personal relations) is a main facilitating factor.

More formalized coordination between operators exist. A timetable conference is organized every 2 years on the basis of a legal article to ensure an integrated timetable throughout Switzerland. Coordination between operators in terms of planning (the German Verkehrsgemeinschaft) and/or fares (the German

Tarifgemeinschaft) are other examples of institutionalized coordination. Transport associations coordinating the services of individual operators on route corridors are also commonly encountered in developing countries. They do not always have a legal basis, and such coordination, often aimed at protecting the self-interest of the companies by precluding further entry, is not necessarily in the public interest.

Integration forms the highest level of coordination by bringing operators together by merging the service design and planning activities of participant transport companies into a single separate and common planning body. This can be observed on a wide scale in Germany (Verkehrsverbund). It started with the union of companies (Unternehmensverbund), such as the original example of Hamburg in the 1960s. Later integrations based on the union of authorities (Kommunalverbund) also appeared. When integration migrated from a "practice" to a legal requirement this became the next step in the evolution of the institutional framework.

4.2. Integration as the "good practice" of coordination in public transport

What transport professionals see as "good practice" can be categorized at four levels. The minimal form includes integrated information about non-coordinated services. An intermediate form includes integrated ticketing and fares. The third pertains to coordinated transport services. The most extensive form of integration includes coordination with other modes of transport and with other policies, such as land use planning.

Integrated information

Integrated information on routes (such as common maps), timetables, and fares (common leaflets) means that the system is to be perceived as a single entity, with a unified set of concepts and a common language in its communication to users (livery, logos, and marketing). Information integration is meant to increase ridership by increasing market transparency through a lowering of the search and information costs for the passengers for the system as a whole.

Ticket and fare integration

Ticketing and fare integration is meant to facilitate traveling from the perspective of the passenger. Ticketing integration is meant to facilitate payment. Fare integration has a more dubious economic basis, as it is meant to remove price differences between similar journeys (a situation that occurs when two similar trips can be accomplished by a one-legged journey for fare x, or by a two-legged journey with different operators for a fare of $y + z$ greater than x). The advent of new electronic ticketing systems has allowed fare disintegration and a closer relation between fares and costs, while keeping ticketing integration.

Network integration

Network integration is seen as a means to enhance the quality of existing services by the creation of a structure where each public transport mode fulfils a specific role within the system. The bundling of passenger streams to higher-ranking modes (from bus to trams, metros, and heavy rail) is explicitly aimed at. This is equivalent to the traditional definition of coordination where modes of transport are to be used in accordance with their relative advantages. Network integration further refers to several levels of coordination. The first is the coordination of investments in infrastructure and main interchanges. The second is the coordination of the transport services planned, avoiding waste and allowing the realization of network effects by providing attractive connections between services (both in terms of transfer time and transfer conditions). The third is the coordination of operations between operators to deliver the level of service promised. This includes guaranteed interchanges, procedures for behavior in the case of foreseen or unforeseen excess demand or service disruptions, adequate information, and/or remedial services such as taxis in the case of delays, and alternative services in the case of service failures such as buses replacing trains. Proper network integration also relates to the links between long-distance and local public transport networks.

Wider integration

This covers two main issues. Firstly, it covers the integration of investment, service, and operations planning with the wider transport system (private car, taxis, and bicycles) and with related non-transport services. Secondly, it means integration with urban planning and with environmental and social policies.

4.3. Transport planning

Transport planners regard coordination as an important topic. Examples of successful coordination are often reported at international professional congresses in the public transport field. Unfortunately, published studies are rather descriptive, see integration as an aim in itself, and concentrate more on ridership successes than on a thorough analysis of all the social costs and benefits of integrative measures to determine what an optimal level of integration would be.

Central planning and the need for integration are present right from the start in transport planning considerations. Other than a reference to "the destructive competitive practices" of the 1930s and a brief assessment of the current functioning of the deregulated UK bus markets, the transport planning approach pays no further attention to market dysfunctions (market "failures") in the argument for intervention. Indeed, much of the current discussion in this area is similar to the transport coordination debate dating back to the period around the Second World War.

This perspective on optimality in public transport is dominant in most places in Europe, and this has much to do with the history of the sector and its "engineering" tradition. Arguments used refer to an efficient usage of resources and produce an "optimal" combination of public transport systems. Elements used are, among others, the system-wide efficiency of vehicle rostering or the ability to provide higher-ranking transport systems (metros instead of trams, trams instead of buses) by bundling of streams through "efficient planning." The focus is often more on hard aspects of service design such as routing and timetables, and less on softer aspects such as vehicle comfort, attractiveness of vehicles and stops, quality of information, and fare systems. While the issues treated are in many respects similar to those of hub-and-spoke airline operations, considerations of total producer costs and of the various time components of user costs often rank much lower in the public transport planner's agenda. In the absence of the competitive pressures of the airline world, this may result in an exaggerated belief in the need for uniformity without economic demand and cost considerations to counter-balance it. At the extreme, the outcome may be a rigid definition of fares and service concepts (regional coverage, average speed, vehicle characteristics, etc.) without adequate inclusion of the reality of local demand on marketing actions. The national fare integration introduced in the Netherlands since 1980 has certainly created such a problem.

Would optimal coordination appear out of the market process instead? This question has not been addressed by the traditional planning approach, but it is central to the microeconomic perspective, and network externalities do play a major role here. Unfortunately, it is difficult to draw clear and general policy implications from the theoretical debate as many questions remain unanswered. To make things worse, the issue is made more complex by its combination with the subsidization issue and state ownership, blurring the effect of many incentives.

4.4. *Reforms*

The public transport sector is the transport sector that currently witnesses the most extensive forms of state intervention in its coordination. But the past two decades have also witnessed substantial reforms in its institutional setting in Europe. This, perhaps, opens a new chapter in the coordination debate, as the path chosen is substantially different from that chosen for most other transport sectors. While other transport sectors were substantially deregulated, liberalized, and privatized, most of public transport continues – even after the reforms – to be subject to a high degree of central coordination. A common feature of many reforms was and is the introduction of some competitive element, but this was done through two completely opposing movements in the public transport sector (van de Velde, 1999).

More coordination by the market

A movement toward coordination by the market has been observed in some longer-distance bus markets, such as in the UK and Sweden. Of these, the most interesting is the 1986 reform of the UK bus markets outside London. Coordination by traffic commissioners and integration by metropolitan providers were abolished and replaced with coordination by the market. "Cooperation" was further prohibited as fare and timetable agreements were from then on seen as unlawful practices that had to be controlled by the UK Office of Fair Trading. This movement fits with the general drive toward coordination by the market that can be witnessed in the other transport sectors.

Although ridership dropped and fares increased in the UK bus market, this experience was also combined with a sharp reduction in subsidization. While the general perception of the experience is negative, Romilly (2001) mitigates this conclusion by concluding that the positive effects of deregulation *per se* on fares and passenger journeys are broadly cancelled out by negative effects of subsidy reduction. The main challenge is to identify the true shortcomings and performances of the current version of the UK deregulated regime compared with the continental monopolistic high-subsidization regimes. An example such as the city of Oxford shows that bus deregulation can indeed be compatible with a substantial ridership growth and a drop in subsidization when taking place in the context of a sensible long-term and wider integration policy addressing transport and land use issues (Parkhurst and Dudley, 2004); and such a context exists in many other European cities.

Another possible conclusion may be that the abolition of coordination has gone too far in the UK. The fair trading regulations should perhaps be abolished, to allow some new form of fare and timetable coordination, through, for example, a traffic commissioner. Such a conclusion would reopen the coordination debate.

A continued coordination tradition

In the other movement the traditional form of coordination remained. Integral competitive tendering regimes were introduced but under a continued central service planning by an authority-linked organization. This regime of outsourcing of the production by route or by a bundle of routes was, introduced in among other places, London, Copenhagen, and Stockholm. Such networks are characterized by a very high level of integration in all dimensions (information, network, fares, operations, infrastructure investment). In fact, the transformation of the former authority-owned transport operator into a central planning and procurement agency has in most cases allowed for an increased integration of services and a closer integration with other transport policies. The London public transport system now falls under the same organization as that responsible for the congestion charging policy. And, in Copenhagen, transport planning and land use planning are consolidated within the same organization.

Areas functioning upon the basis of tendered networks (concessions), as in the French urban areas and, to a growing extent, in public transport in the Netherlands, also reject coordination by the market. Coordination is guaranteed here by choosing a monopoly to run all services within an urban area for a specific period. In contrast to the London example, service planning is in principle delegated to the operator in these cases, but highly specified contracts often leave the operator little leeway to autonomously amend services. In high-population-density countries this has led to new types of coordination problem. New institutional arrangements now have to be created to solve cross-boundary services between neighboring concession areas.

Pending reforms

Many areas continue to function under the old non-competitive, non-tendered regime, and many of the operators in these areas defend their anti-competitive stance as a way to support coordination. While public transport services indeed reach a very high level of coordination in most urban areas in Germany, there are also numerous other cases where public operators coexist without integration. At the same time, London, Copenhagen, Stockholm, and other areas that have followed the same route of market freedom show that a multitude of operators is perfectly compatible with a level of coordination that can be higher than in places with publicly run monopolies.

5. Conclusion

During the last two decades, deregulation, liberalization, and privatization have had tremendous impacts on the organization, conduct, and performance of the various transport sectors. This is clearly visible in, for example, the EU, where measures continue to be developed to foster freer market access to modes of transport and a more effective competition both within and between modes. Except for infrastructure planning and investment, coordination has largely disappeared from the list of major concerns for regulation in the transport industries, i.e. the argument of cut-throat competition is almost dead, and the market mechanism is trusted as a coordinating device.

The main exception to date is public transport, which is still mostly coordinated by planning and regulation. Central planning with competitive tendering is a recent addition, and is a modern competition-based instrument of coordination for public transport. Integral competitive tendering of transport services ("franchising") by transport authorities replaces the free market, and abolishes autonomous market entry and behavior. It should therefore not be confused with coordination by the market. Full deregulation, as introduced in the UK bus sector outside of London in 1986, is the main alternative and only large-scale example in western Europe of coordination by the market in public transport. While this example clearly has some shortcomings, further studies are required to judge its true merits compared with traditional coordinated regimes (with or without competitive tendering).

The main issue from the perspective of a legislator or regulator is that of the design of institutions that will maximize the appearance of an optimal level of coordination through the market and/or by direction. A related issue is the implementation of an adequate institutional reform to facilitate this goal. As optimal coordination is a moving target, owing to technical and social changes, the institutional design should be flexible enough to accommodate future needs without repeated legal changes. Unfortunately, no universal solution exists to guarantee adequate coordination and avoid dynamic inefficiencies (see also Carlton and Klamer, 1983). To lead to optimality, any regulatory framework will have to induce a continuous optimizing process, and this is perhaps the greatest challenge. This means that the social costs and benefits of existing and additional coordinative measures need to be constantly re-examined. Adequate action needs to be induced both to amend or cancel existing measures that have adverse effects and to develop additional actions that have positive balances.

Two main routes remain open in public transport for those who reject the theoretical idea that optimal coordination will result from the market process. The first is to devise an adequate set of "rules." Such incentives, through laws, regulations, and contracts, would have to solve the problem of market failure and generate adequate coordination. Within this setting, operators would have to behave according to their own interests, where the "rules" bias their own interests toward the general interest. Whether such "rules" can be designed remains debatable. The second is to impose a predetermined form of integration through laws or regulations, where an authority takes the lead by making specific decisions in the various dimensions of integration (routing, frequencies, fare setting, etc.), even if some residual leeway is left to the operators. It is likely that a mixture of both instruments will be needed, depending upon the specific issues at stake and the related transaction costs. The controversy between market-based and regulation-based coordination remains alive, at least in the local passenger transport sector.

References

Baumol, W.J., J.C. Panzar and R.D. Willig (1982) "Contestable markets: an uprising in the theory of industry structure," *American Economic Review*, 73:491–496.

Button, K.J. (1993) *Transport economics*, 2nd edn. Aldershot: Edward Elgar.

Carlton, D.W. and J.M. Klamer (1983) "The need for coordination among firms, with special reference to network industries," *University of Chicago Law Review*, 50:446–465.

Economides, N. (1996) "The economics of networks," *International Journal of Industrial Organization*, 14:673–699.

European Commission (1995) *Green paper: towards fair and efficient pricing in transport*, DG VII. Luxembourg: Office for Official Publications of the European Communities.

Foster, C.D. (1963) *The transport problem*. London: Blackie.

Gardner, B. (1997) "EU competition policy and liner shipping conferences," *Journal of Transport Economics and Policy*, 31:317–324.

Gwilliam, K.M., S. Petriccione, F. Voigt and J.A. Zighera (1973) *Coordination of investments in transport infrastructure: analysis–recommendation–procedures. Studies, transport series,* No. 3. Luxembourg: Office for Official Publications of the European Communities.

Klapper, C.F. (1958) *Transport co-ordination in Great-Britain. Monografie No. 18*. 's-Gravenhage: Nederlands Verkeersinstituut.

Kuiler, H.C. (1949) *Verkeer en vervoer in Nederland schets eener ontwikkeling sinds 1815*. Utrecht: Oosthoek.

Kuiler, H.C. and J.M. Verhoeff (1973) *Inleiding tot de vervoers- en haveneconomie: een selectie uit de geschriften van Prof.dr H.C. Kuiler.* Rotterdam: Universitaire pers.

Mance, H.O. (1941) *The road and rail transport problem*. London: Pitman.

Mohring, H. (1972) "Optimization and scale economies in urban bus transportation," *American Economic Review*, 62:591–604.

Parkhurst, G. and G. Dudley (2004) "Bussing between hegemonies: the dominant 'frame' in Oxford's transport policies," *Transport Policy*, 11:1–16.

Peterson, G.S. (1930) "Transport co-ordination: meaning and purpose," *Journal of Political Economy*, 38:660–681.

Ponsonby, G.J. (1969) *Transport policy, co-ordination through competition*, Hobart Paper 49. London: Institute of Economic Affairs.

Romilly, P. (2001) "Subsidy and local bus deregulation in Britain: a re-evaluation," *Journal of Transport Economics and Policy*, 35:161–194.

Saito, T. (1989) "Transport coordination debate and Japanese national railways problem in postwar Japan," *Transportation Research A*, 23:13–18.

Tissot van Patot, J.P.B. (1938) *Coördinatie: nota omtrent Engelsche coördinatiepogingen in het goederenvervoer.* Amsterdam: Vereniging Binnenscheepvaart-Congres.

Tissot van Patot, J.P.B. (1964) *Enkele beschouwingen over motieven en doeleinden van coördinatie-politiek*. Assen: Born.

van de Velde, D.M. (1999) "Organisational forms and entrepreneurship in public transport (Part 1: classifying organisational forms)," *Transport Policy*, 6:147–157.

Willeke, R. (1977) ""Ruinöse Konkurrenz" als verkehrspolitisches Argument," *ORDO Jarhrbuch für die Ordnung von Wirtschaft und Gesellschaft*, 28:155–170.

Wolf, C.J. (1979) "A theory of nonmarket failure: framework for implementation analysis," *Journal of Law and Economics*, 22:107–139.

Chapter 9

INTEGRATED TRANSPORT SYSTEMS: PUBLIC–PRIVATE INTERFACES

JOSÉ M. VIEGAS
CESUR – Instituto Superior Técnico, Lisbon

1. Transport systems as integrators

In the sense that is normally used in transport, integration allows components of a system to perform well together; that is, to have small transition costs when the user of one component moves to another component. Accepting this concept of integration leads, however, to discussion one level upstream, to the integration between social activities, developed in different locations, that are connected by the transport system.

Integration between land use and transport is especially important because the lowest transition costs between activities occur when there is little or no need to travel from the location of one activity to that of the next one undertaken by the same person or using the same material goods. There are clear advantages in such situations, both for the individual – less time and cost spent on relocating oneself – and for society – less allocation of space to support transport activity and less environmental impact.

A wider vision may lead to a definition of various scales of integration of transport systems, from the functional scale to the environmental, economic, and policy scales, as defined in Potter and Skinner (2000). In their definition, concerning transport and planning (land use) integration, we must still consider social integration, whereby the needs of all those who have a stake in transport are considered in the definition of the transport system (Viegas and Macário, 2003), and environmental, economic, and transport policy integration, by which a holistic view of all the relevant policies is pursued, as demonstrated in a report by the European Commission (1995).

When a transport system is designed with a strong hierarchical structure, it is relatively easy to specify where, in relation to that structure, each type of land use should be located, depending on the geographical range. Equally, when the major generators of travel demand in one region or agglomeration are settled, it is relatively easy to integrate their needs into the planning of a public transport network.

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

Over the long term, however, this integration is made difficult by issues associated with the inflexibility of transport infrastructure and economic activities:

- Over the life cycle of a piece of transport infrastructure there will be significant changes to some of the activities occupying land parcels served by it, and changes to the way those activities are organized, including the relations they establish with other activities in order to improve performance. Therefore, when transport services are provided over dedicated links of long duration, patterns of demand will vary. In the early stages the majority of connections may be direct; however, these will later change to involve many different origin–destination pairs, most of them involving transfers.
- With regard to economic activities, it is very difficult to ensure that what was once an efficient location from the point of view of transportation needs remains so over the lifetime of that activity. However, the relocation costs for economic activities are normally so high that economic agents prefer to endure some loss of transport efficiency rather than search for the ideal location. With the systematic reduction of transport costs over recent decades, the quantity of transport being used is certainly much larger than would be preferable in ideal planning circumstances.

In addition to inflexibility over time, there are also difficulties on the political front, since the rationale of the agents involved in land use decisions (investors, planners, and local politicians) is geared toward different targets from those involved in transport planning and policy. Location, with its implicit link to transportation, is always referred to as a critical factor in site selection, with a focus on creating favorable conditions for the activity and attraction of the corresponding investment, including providing good access to the higher-scale transport networks; however, there is little concern for the aggregate consequences of the mobility thus generated on the transport networks. If fact, whereas in most countries land use changes are a result of private initiative, subject to public permission, transport systems develop mostly based on public initiative, partly as a response to pressure from the private side.

In the short term, apparently good practices of spatial integration of activities occur mostly because economic agents recognize the existence of significant agglomeration economies, both of scale and of scope, around certain locations. Regional concentrations of particular sectors provide critical masses of skilled workers, allow specialized outsourcing and an overall environment of competitive pressure, and sometimes even have the presence of dedicated specialized transport links. However, many examples have occurred in the recent past in which this spatial economic specialization, or monoculture, has led to severe economic depression when the dominant sector of activity goes through a phase of decline, either through loss of competitiveness or simply through lack of demand.

In the long term, therefore, less specialization may be healthier for the regional economy; even if that implies that most of the time the quantity of transport generated will be much larger than is strictly necessary. The best compromise between specialization and diversity may be impossible to define, since the success and survival rates of the various activities in the end always largely depend on the skills and entrepreneurship of individual agents.

2. The features of integrated transport systems

Here we deal with the scale of functional integration as defined in Potter and Skinner (2000); that is, with the internal integration of the components of a transport system, taking its design as a given. The issue of network re-design will, however, be addressed where it is relevant to the question of internal integration.

For the purpose of this discussion, it is useful to define what we mean by a well-integrated transport system. In very general terms this question may be easily answered: a transport system is well integrated when all (or the large majority of) its users consider that it does not impose unjustified transition costs on them, i.e. when their transport need implies service by two or more links, the process of changing from one link to the next does not constitute a significant cost in terms of time, price paid, physical energy, loss of comfort, etc.

These problems are generally considered in the context of public transport services, because the road system is capable of providing very good levels of integration overall.

The key feature is interoperability; this means that the vehicle, or more generally the set of production resources, including the vehicle, driver, and corresponding licenses to operate, enjoys seamless transition from one subsystem to the next. Whereas interoperability is focused on the supply side of transport, integration is more general, and reflects the view from the users' perspective. Of course, a lack of interoperability is by itself a source of integration costs.

Even on the road system there are still many examples of poor integration when crossing borders from one country to the next, although the sources of the transition costs in these cases are normally found outside the transport sector. Recently, interoperability problems have emerged for heavy goods vehicles within the EU, as various incompatible systems are being adopted for electronic toll collection across the different countries.

Problems of poor interoperability are more widespread across national borders in the railway system, with differences in gauge, electric power, and signaling systems, as well as in drivers' accreditation, with each of these contributing to an essentially non-interoperable system.

Whenever interoperability between two networks is missing, any service between one node in one of those networks and a node in the other requires one transfer, with its associated transition costs.

Public transport systems, urban or interurban, by land, water, or air, with their variety of modes and underlying technologies, are a domain where non-interoperability is very common. Where direct connections are not available, the transfer process should be as smooth as possible, not only in time, energy, and additional cost for the user but also in uncertainties related to the next part of the trip.

Seen from the users' side, this is normally described in three dimensions of integration (ISOTOPE, 1997):

- Physical integration, which means that the transfer process from one link to the next must be such that the user does not have to expend much time or effort on it. This normally requires well-designed interchange stations and properly coordinated timetables. When luggage is involved, adequate facilities and operations must be in place.
- Tariff integration, which means that travelers who would like to be served by a direct link but are not should be able to buy the transport titles at a single location and time, and should not have to bear a tariff penalty on top of the physical penalty that is already inflicted upon them by the need to transfer. In a well-integrated system, inasmuch as transport prices are a function of distances traveled, prices of composite services (two or more links used with a connection) should reflect what would have been the price of a direct service, not the sum of the prices of the various components.
- Logical integration, which means that users should be able to perceive the service as a whole unit, with available information about the whole chain of services as if it was a single through service, with special emphasis on services that allow the reduction of the uncertainties and risks associated with any transfer process.

In order to achieve these user-side objectives, significant efforts must be made on the supply side. In addition to the attention given to the design and organization of interchanges so that good physical integration is possible, legal and contractual arrangements must also be in place for the tariff and logical integration to be sound and stable. A special case occurs when transport titles are issued for specific services and a connection is missed due to delays in the upstream transport service. Good integration implies that the traveler must be served as soon and as swiftly as possible, with no increase in price.

In this process, a critical role is played by transport interchanges. Interchanges should be transformed from a necessary evil in the journeys of travelers who do not enjoy a direct connection into an element of added value to the area where they are located, owing to the opportunities deriving from the artificial concentration of demand generated by those transfers, frequently on top of a rather high local/regional demand. In this perspective, an interchange should not only provide smooth transport connections for the transit passengers but also allow them positive experiences of leisure and consumption. In parallel, the large numbers of transit passengers can make a greater variety of shopping, entertainment, and leisure opportunities viable, which could in turn be made accessible to a wider public in the area, not necessarily in connection with the transport system that justified the interchange in the first place. This type of evolution became very visible during the 1990s, the best known of which are Schiphol airport in Amsterdam and Leipzig central railway station in Germany.

Although the design and operation of integrated transport systems imply additional costs for the economic agents involved, these costs can be socially justified by two types of benefit: market growth and distributive equity. Lack of integration implies much higher costs for those clients located outside the main poles or axes of demand, where most direct services start or end, and thus a smaller mobility of markets overall as well as a very skewed distribution of mobility. From a collective point of view there is also an interest in avoiding this skewed distribution of mobility by public transport, because those who feel badly served will tend to solve their problems by returning to individual transport, thus aggravating congestion problems.

3. The difficulties and processes of transport system integration

On the supply side, integration has significant costs of the following types:

- network planning, so that transition costs from one system component to the next are as low as possible for the client and for the company, and supply patterns can be regularly adjusted to make full use of the additional demand potential;
- preparation of contractual agreements between the companies involved in the integrated system and potentially with some public authorities driving it;
- information systems that allow each supplier in the system to know about the timing and flow quantity it will receive from each upstream component in good time, by keeping track of all flows and incidents in the actual operations, and through billing other integration partners accordingly;
- contingency planning, so that clients are adequately treated in case of a disruption to the planned flow of events;
- operations staff that are dedicated to the transfer clients, both in normal and in disturbed circumstances.

The costs outlined above are real, and grow as the size and complexity of the networks increase, even if all transport suppliers are in the public sector. So, integration must be questioned. No user is negatively affected by the presence and proper functioning of an integrated transport system, except if only a direct connection is needed, and the user is somehow forced to contribute financially, through the ticket price, to these integration costs, i.e. cross-subsidizes other users.

Given the existence of these additional costs, we need to consider the following fundamental questions:

- Does transport system integration occur without state intervention, and if so in what circumstances?
- If transport system integration does not occur very frequently without state intervention, what are the reasons that could justify state intervention promoting transport systems integration?
- If the state decides to intervene, what is the best way to do it?

On the first question, it should be noted that good examples of integration always occur when there is a strong concentration of operations in one company (no problems of demand reassignment or cost allocation), or when there is strong leadership of the integration process by a public agency that sets clear rules and includes the interconnection costs in the compensation it pays to operators. In the second scenario there is a clear case for state intervention; however, the first may occur in many cases of market domination by a single private company.

In both scenarios we may speak of the existence of strong "property rights" in the Coasian context (Coase, 1960) that is over passengers, in the sense that a single public or private entity largely determines their mobility options and the prices they are required to pay for them.

Although the main focus of this chapter is on systems with state intervention, some inspection of integration practices in settings without that intervention may be helpful in understanding the forces favoring or fighting integration in the markets. When there is no state push toward integration, private companies may find it in their interests to adhere to integration, normally in association with acquisition or preservation of strong market positions, and rarely in a totally open scheme. Integration will be available only for some combinations of services, with either low marginal production costs or with high potential for client capture (a stronger form of loyalty).

One example in this area is found in the air passenger transport sector. The concept of interlining is one form of integration that has been applied for several decades. This provides passengers on complex trips involving different flights with physical integration (check-through) for their luggage a particular type of tariff integration, namely protection against delay in one leg of the trip, by ensuring that in such cases passengers will be transported on the next available flight, possibly through another company and route.

In preparation for the gradual opening of the markets (now fully realized in the USA and in the EU, albeit only for domestic companies), airlines have tried to maximize their benefit in these integration processes, by adopting hub-and-spoke network designs and charging much cheaper prices when the passenger flies all the legs of the trip using the same company (or companies in the same alliance), in comparison with the prices charged when the passenger "jumps out" of that supply chain and makes use of one flight by another company or alliance. As airline alliances reach a global scale, using a company outside the alliance on which the journey started will become so expensive that a passenger would only undertake this in exceptional circumstances. So voluntary capture is achieved through tariff discrimination in integration.

In parallel to this practice, a competing concept to these alliances has been growing very strongly: point-to-point services, with no integration whatsoever with other services. Often provided by low-cost carriers (LCCs) for passengers who do not require integration, and especially for those who also do not require special assistance at terminals or on-board services, price reductions can frequently be as high as 80% for these services. This has the double effect of increasing market size and also of reducing the flow of passengers in the alliance services on the links served by LCCs, thus creating problems for their cost coverage and possibly forcing them to reduce service frequencies, one of the core features of the hub-and-spoke concept.

Traditional airlines are trying to fight back with a variety of tactics; for example, by selling some of their point-to-point seats at discounted prices. The final result of this competitive process is not yet clear, and will probably not be the same in all regions. In some cases the main effect of the new supply by LCCs will be to increase the market dimension, leaving the traditional companies with enough demand to allow preservation of their production scheme. In other cases, however, the market may not grow enough to let this happen, and the loss of market share by the traditional companies may imply a reduction of their service frequencies with a corresponding deterioration in the conditions of smooth travel between pairs of secondary or tertiary airports for which there is no direct connection.

Some Internet-based services have begun to offer combination searches of low-price flights for these kinds of connections, even if in such cases travelers must take full care of their own luggage and bear all the risks of delayed flights without protection for the next legs of the journey (e.g. Gooflight – www.gooflight.com).

A noteworthy case, in the direction of non-occurrence of integration, is that of intermodal freight transport in Europe. In spite of all the effort put in by the European Commission in research and development projects in this area, as well as in its political promotion, it still accounts for very little market share, even considering only those markets where the distance is long enough to make it viable. Recently there have been calls for the creation of the profession of "freight integrator" within Europe; however, it must be questioned what features of the market have to change for this integration to succeed without massive subsidy from the state (European Commission, 2001).

The explanation for this impasse is simple. In open markets where intervention by states is limited (at least through allocation of subsidies to particular modes or combinations thereof), none of the key players have found it to be to their advantage to promote integration, at least under the current conditions of general prices and of access to the infrastructure. Only if access to road infrastructure for heavy goods vehicles starts to become overly expensive, which is not an obvious scenario, will some coordinated movements begin to be implemented, involving large operators of the heavier modes, enjoying advantageous positions at key transfer points, and large haulage companies, aiming at dominating positions in a market that will be less fragmented than the current haulage market.

We can therefore conclude that some integration may occur without state intervention but only when it is an instrument of reinforcement of the dominant positions of private companies, not because of a global design of the system. In principle that does not diminish the value of integration for consumers, but there may be a thin line dividing these integrative practices from abuse of dominant positions. Thus, the state must still monitor companies through competition authorities.

On the question of why the state should intervene when integration does not occur through the direct action of private agents, we should start by looking at the most common example of this setting: large urban areas. In these areas the complexity of travel patterns is such that a public transport network with poor integration would certainly reduce the mobility of poorer citizens and give away a good part of its market potential to the private car.

Therefore the rationale for integration is that it generates consumer benefits which are larger than the additional producer costs it entails:

- There are two levels of X-efficiency involved: directly, as the proper integration of the same public transport input clearly reduces the costs for consumers using these modes; and indirectly, as the much greater efficiency of public transport over the private car in the use of urban space, and the very high cost of that space reduces the need for this precious input as long as public transport service is good enough to run with high-occupancy levels and can attract car-owning clients.
- Social cohesion arguments can also be added, in defense of residents in less dense areas or with lower purchasing power, who could be severely discriminated against in the absence of adequate levels of integration.

However, at least in local markets, it is politically difficult to charge the benefits of integration to those consumers who directly benefit from them. Each citizen does not see it as their fault that the trip they want to make is not served by a direct connection, so why should a second penalty in the form of price be imposed on top of the physical and time cost associated with the transfer? In the end it is much cheaper to provide accessibility through integration than through provision of direct connections between all relevant origins and destinations, and to divide the additional costs of that integration among all clients.

There are some good examples of integration in planned networks in interurban transport, for instance the Rail+Bus 2000 concept that was introduced in the 1990s by the Swiss Railways, after favorable referenda held in the late 1980s. Given the size of the country and the fragmentation of the population across it, the favored option was to design a network with very efficient physical integration – good frequencies, matching timetables to optimize transfers – instead of betting on high speed in a few connections (Andersen, 1993). But here again we are dealing with a state vision, enacted by a publicly owned company.

In the longer term, good integration of the transport system also brings benefits by reducing the polarization of the territory on the main transportation axes, thereby improving the distribution of accessibility, which is instrumental to the distribution of many factors of welfare: jobs, public facilities, social contact, etc.

The role of transportation in modern societies is so critical for most activities that the preservation and stability of its functionality is paramount. Not only in normal circumstances but also during maintenance and upgrade operations, the supplier of each transport component and its clients should be able to count on the supplier of the components upstream and downstream. In such systems, allocation of supply roles to the various agents inevitably has a great inertia, so that possible changes at this level do not cause loss of the stability and corresponding integration.

These attributes of stability and integration require strong management, which can only be delivered by the state (as is the case in most public transport networks in urban areas) or by a cohesive alliance/conference of powerful players who may be competitors in some domains but have a vital interest in the stability of the system, as is the case in the global air transport and intercontinental shipping businesses.

4. The involvement of private agents in integrated transport systems: advantages and difficulties

Even in systems so strongly driven by the state, the presence of private players may be advantageous, for four main reasons:

(1) they are more subject to the risks of poor economic performance, which makes them more determined and capable in the pursuit of economic efficiency;
(2) they can mobilize funds in private capital markets, avoiding formal constraints on public borrowing and spending, thus allowing faster completion of investment programs and earlier entry of new transport components into service;
(3) they tend to search frequently for improvements in their position in the market – one way of doing this is through innovation in the services they supply or in the ways to produce them;
(4) they limit irrational choices of key supply variables, as they are less susceptible to direct interference by politicians who frequently feel tempted in electoral periods to offer additional services or reduced prices for public transport.

Two more reasons are less frequently considered: the involvement of private companies in the production of services of general interest, as opposed to direct production by state-owned companies, seems to increase the willingness of citizens to pay a fair price (Gomez-Ibanez and Meyer, 1993); and the involvement of private companies also favors the transparency of the deals when some form of public subsidy is present. Although not all of these arguments are always applicable, there is an overall favorable assessment of this type of involvement (Spackman, 2002).

There are also costs and difficulties related to the introduction and presence of private suppliers in those markets with direct state supervision:

- capital costs are higher, in direct relation with the remuneration of private capital, the commercial risk of the project and the lack of guarantees given by the state for the payback;
- there are significant transaction costs, related to the tendering and negotiation processes, in which an attempt at quasi-definitive allocation of risks between the public and the private side is frequently made;
- there will naturally be different objectives on the public and on the private sides, which implies the need to design a balanced and effective set of contractual incentives so that those objectives can come closer. This balance is especially delicate in integrated transport systems, where the performance of the agent responsible for each component may be affected by the performance of other agents.

In the recent past, the dominant reasons for the opening of previously regulated transport markets to private operators have been those connected with productive efficiency and the availability of capital outside the formal limitations of public funding.

Integrated transport systems may include some forms of competition among operators, but only in such ways that it does not cause instability (Viegas, 2003). Thus, periodic tendering for parts of the market poses no problems, as well as substitution of one operator by another in a part of the integral supply. Abandonment of a previously existing service, however, or the introduction of a new one (or even more simply, a change of timetable or stopping patterns) may cause unacceptable disruption. The limits to these changes are normally fixed by those who control the integration attributes of the system, i.e. the states or the alliances of dominant operators.

Although this requirement of stability is positive in the short term (favoring integration, it helps consumption efficiency), it does, however, pose risks for the longer term. The system functions as a sequence of time-limited sets of spatial monopolies, where gains of market share are always very difficult, since the market is divided and allocated to specific operators for each of the time periods.

In the traditional forms of these integrated networks, operators can only improve their situation by gains of productive efficiency, not by redesigning their supply. This is a strong disadvantage, both directly by prohibiting innovation, and indirectly because it incorporates this prohibition into a mindset, transforming those operators into simple vehicle pushers and cost-cutters.

How detrimental to the evolution of the transport system this joint accommodation of authority and operators to current practices may be is occasionally made notorious by the unexpected entry into the market of new players who find ways to circumvent the regulations and offer new types of services that clients adhere to but that were not included in the planning.

Good design of an integrated transport system must therefore include some openings for innovation through the initiative of private operators and not by the specification of the authority.

In an integrated system, intervention from private companies must not be limited to transport operators. The public authority has to exercise the political will in the definition of its priorities and the allocation of public resources, but this does not imply that the multiple technical tasks required for good integration, besides transport production, have to be performed by public agencies. The following is a list of the most significant tasks:

- designing the network and timetable;
- managing the interchange stations;
- creating and managing the information systems that facilitate the use of the transport system by its potential clients;
- collecting tariffs from users and controlling proper payment for the services received;
- monitoring the performance of the system and its components and transfers, as well as the quality as perceived by its users;

- remunerating the operators for the services provided;
- allocating bonuses and penalties to service providers, based on their performance;
- handling complaints and suggestions by users of the system as well as by impacted third parties;
- informing the public at large (tax payers) about the allocation of public money and the results obtained.

None of these functions is susceptible to being left to an open market, as they all require a stable relationship between providers and the authority. Private companies, however, normally selected by tendering, could deliver all of them. In several cases, at least in large systems, there could even be more than one provider, by dividing the system into a few spatial domains. This may increase the burden of contract negotiations and management, but allows direct comparison between performances of parallel providers, thus creating an element of permanent competitive pressure on them, instead of just at the moment of contract renewal through another round of tendering. It also constitutes an interesting element for induction and quicker diffusion of innovation.

Some of these functions grow in importance with the level of integration required on the system, and these are the ones that require greater care in specification, contracting, and management. In general, for prevention of conflicts of interest or bias in delivery, a function that is related to integration should not be allocated to a company that is responsible for one of the services being integrated in a wider chain, unless it also provides all the other services in the chain. The evaluation of the performance of the integrating functions should distinguish between a good end result (e.g. an efficient transfer) that occurs simply because all components were as expected and one that is achieved in spite of misaligned components, and between a bad end result that occurs owing to poor performance of the components and one that occurs because of a failure of the integrator.

5. The instruments for integration and their deployment

We can now address the third question: if the state decides to intervene in support of integration, how should it be done?

Just as the integration of transport services in a cohesive network is not easy to design or maintain, so the integration of public and private partners in the specification and delivery of such a network requires good design and management efforts. In the end, for the visible integration of transport services to function properly, there has to be an invisible integration of public and private agents supporting it.

The essential requirement for a good start of this process is clarity of purpose on the public side, where the driving power resides. Public authorities must specify what level of integration of transport services they require, what instruments and what resources can be mobilized for this, and what role private agents are expected to play in the delivery of such a system.

The next step is to develop a good institutional design for the implementation and management of the process, recognizing that, even on the public side, there may be diverging interests, for instance related to different territorial levels of administration, or simply to different circumscriptions of the same level. Once all the functions required for a well-functioning system are listed and described, with explicit mention of the aspects of each one that may raise difficulties, thorough analysis of the legal competencies, the financial capabilities and managerial skills of the various parties – public and private – to carry out each of those functions must be performed. Finally, cross-checking of potential allocations of functions to parties has to be undertaken, in order to prevent conflicts of interest as well as excessive concentration of power. It must be noted that the list of functions should include not only the production and marketing fronts but also monitoring of performances and auditing of accounts.

The first two steps are the structural ones, corresponding to the mission and organization driving the integrated transport system. Once these are taken, decisions have then to be made concerning ways to engage other parties besides the agency leading the initiative. This will be translated through a set of agreements and contracts, involving public–public as well as public–private relationships.

Good design of relationships between different public institutions is not a trivial matter, and is very much dependent on the constitutional organization of each country, i.e. allocation of roles between central, regional, and local levels and hierarchical relations between them.

For the engagement of private parties in an integrated transport system by a public agency a formal contract is desirable, covering at least the purpose of the relationship (specification of the obligations of both parties), its normal duration and conditions for abnormal termination, the remuneration formulas, and the allocation of risks between the parties.

The preparation and management of such contracts have several special difficulties in the case of integrated transport systems. An initial difficulty is related to the level of competence required on the public side:

- Contracts should be designed so that they impose effective pressure on the service suppliers to provide high-quality services. Those private companies who are competent enough to provide these services will also be expected to be competent to negotiate the corresponding contracts, which requires that on the public side there will also be competent, highly skilled, motivated staff, which is not always easily achieved.

- Proper management of these contracts over their lifetime requires technical, legal, and managerial skills on the public side, in line with the skills available on the private side.
- The public sector must retain its ability to procure efficiently. The experience of some cities who went from provision of public transport exclusively in the public sector to exclusively in the private sector shows that such a complete transfer may raise a delicate problem, namely that after a few years the public sector may lose its ability to procure services effectively and efficiently, by being left with no-one who really knows by direct experience how to produce those services. One solution is to leave parts of each of the services to be carried out by a small public agency, or by regular hiring of staff from the private sector, to renew the skills in the public agency. However, the first option is only viable in large systems, and this coexistence of two rationales may be difficult to manage over time.

A potential answer to these problems has been found in the Netherlands, with the creation in 1999 of the CVOV – the Center for Innovation in Public Transport. This is an advisory institution at the state level that acts first as the issuer of a set of guidelines for these procedures and a contractor/disseminator of best practice, and then as a "consultant" to the various departments and agencies, at the regional or municipal level, who are responsible for the actual tendering and contracting, as they act on specific cases. Such agencies regularly hire people from the private sector to ensure good procurement capacity, thus relieving each local authority from a similar need, and can later also be used as a technical adviser to a public sector inspectorate or in a court case, or to assist in the auditing of the performance of public agencies.

Contract duration and the degree of specification are decisions of great significance. In the case of participation of private agents with significant mobilization of private equity (including dedicated infrastructure), it is usual to have franchise contracts with a long duration and very rigid clauses, to allow full amortization of the private investment under what could be called a complete contract. In these contracts the concessionaire is responsible for the investment, construction, maintenance, and operation of the system, and collects the revenues from its users. In addition to this revenue there may be an additional value paid by the state, which can be paid either at the beginning of the contract, serving as a reduction of the investment supported by the private company, or during the contract, as a fixed value per time unit, or more frequently as a top-up fee to complement the charges paid by the users.

Together with the pressure on the concessionaire for overall economic efficiency in construction, in procurement, and in service provision, this type of contract has the significant advantage of avoiding classification of the investment supported by the private company as public debt. It does, however, require great care in the specification, design, and performance of the elements of the transport system surrounding the subsystem that are put into concession, so that they do not constitute bottlenecks preventing adequate feeding of that subsystem or, on the flip side, overtly competitive alternatives that reduce the market potential for the subsystem under concession.

The option for a complete contract may create in both parties the satisfaction of closure, but it may raise severe difficulties during its life, as the framework of mobility requirements and solutions may change, opening the way for changes in associated policy goals, and with them the aim of that component of the transport system.

If the initial contract is too rigidly specified, then contract renegotiation or break-up is then necessary. In fact, a return to shorter and incomplete contracts is preferable (Viegas and Colaço, 1998; Segal, 1999), which explicitly allow for a residual value to be paid to the concessionaire, but frame the contract in a time horizon up to which there is a reasonable idea of what the priorities will be. Renegotiating or breaking complete contracts is always possible, but generally very costly to the state, and it is not uncommon to see authorities living with less-than-optimal formulas for transport provision because of the correction costs.

If a franchise contract with shorter duration were signed, a new one would likely follow on its termination, but there would then be the possibility of redefining the objectives and conditions for engagement of the private parties, in line with new traffic patterns, public policies, and priorities.

As well as franchise contracts associated with significant private investments, there are also many cases of participation of private companies in the provision of public transport services, which may be accompanied by more limited investments, for example in rolling stock.

In this case, it would seem that in the name of efficiency and ingenuity it would be appropriate to allocate industrial and commercial risks to the concessionaire, adopting what is normally called a net cost contract. Under this type of contract the operator is responsible for all production costs, and collects the revenue from the users of the system, possibly complemented by a contribution or subsidy from the authority side, which can be of a fixed value per time unit or be based on some indicator of the level of demand.

Experience has shown that there are strong disadvantages associated with this type of net cost contract in the area of complex urban mobility systems.

Difficulties arise from integration itself, particularly when multiple operators are present in the same network:

- In a well-integrated transport system a large proportion of passengers will be traveling with integrated tickets valid over multiple routes, possibly operated by different companies. Sharing those revenues among these companies based on passenger-kilometers transported is possible but requires complex surveys or sophisticated electronic tools to check each passenger in and out of every leg of the trip.

- The modal option for public transport is dependent on the quality of service perceived by the users, and this depends on the quality perceived on each component of the trips made, thus making each operator commercially dependent on the quality supplied not only by itself but also by others.
- Numerous elements in the urban transport system affect the commercial results of a surface public transport line or group of lines; for example, its connections to the rest of the public transport network, the types and levels of priority of public transport over general traffic (dedicated lanes, active light control at intersections, etc.), and the level of supply and price regimes of car parking in the city center. Changes to any of these elements may be desirable for general or specific reasons of mobility management but can have negative consequences for the commercial results of a group of lines under concession, without the concessionaire bearing any responsibility. To avoid the natural claims in such cases, urban transport policy and management may be captured by the concession contract.

It would seem that the answer would be to have public transport operations in each agglomeration/network contracted with one single operator, but this creates more problems than it solves, as shown by Toner (2001) to be the case in London:

- Particularly in relation to large networks or parts of a network, net cost contracts raise the problem of reducing contestability of the market, as the levels of risk associated are normally perceived as too high for small operators. As a consequence, the gains expected from the commercial pressure put on the operator by the type of contract may well be lost through the reduced number of bids, and possibly even collusion among a small number of potential bidders for a series of tenders.
- Another serious problem with this type of contract is related to the second and later cycles of the tendering process, as there is a strong asymmetry of information between the incumbent operator and all other candidates in relation to the actual costs and revenues of the various services included in the package. Not only are lower bids by new entrants unlikely, but, if the case arises, the authority is faced with the dilemma of opting for that bid with a reasonable risk of instability in supply due to miscalculation of the costs and revenues by the new entrant, or staying with the incumbent.

For all these reasons, the preferred alternative for the engagement of private companies in integrated transport systems is generally to use gross cost contracts. In these contracts, the tariffs collected from the users belong to the authority, and the operators are remunerated by their production according to a network and timetable defined by the authority, with a correction through incentives related to their patronage and to the quality with which their service is produced and perceived by the clients. In the pure case (no incentives) the private party still has a production risk but no commercial risk.

As mentioned above, integration of services provided by different operators makes their performance and attractiveness for clients mutually interdependent. In gross cost contracts this will impact on the variables used for computation of incentives, albeit at a much lower scale than for net cost contracts, as the basic remuneration level is based on the volume of production.

Having only one operator for the whole integrated transport system, which would be the natural choice in small- to medium-sized agglomerations anyway, may solve this particular difficulty. However, in larger agglomerations there are also advantages connected to the existence of several sub-networks with different operators, including direct comparability and thus permanent pressure on the performance of each of them, and regular tendering opportunities. In such cases, the incentives written into operators' contracts must be specified in terms of indicators of quality delivered – which can be measured for each transport service – and restricting the indicators of quality as perceived by the clients and of patronage to segments of the network with little or no integration.

Another problem that is common with gross cost contracts, particularly in integrated systems, is how to break the apparent dilemma between regulated integration and innovation in the design of services. Not only is it the rule that in these contracts the authority specifies and the operators provides as specified but the requirements of integration leave less room for experimentation with new patterns of service.

The innovation area of concern here is not new vehicles or technologies – which are "pushed" into the systems by persuasion applied to the operators or to the authorities by the respective manufacturers – but innovation in the suite of services available to clients, as mobility wishes evolve and become more segmented. It can be argued (Viegas, 1999) that one of the reasons for the systematic loss of market share by public transport in the developed world in recent decades has been the lack of market segmentation by continuing to provide a single type of service, designed only for those who do not have a car available. Some examples of possible services for emerging requirements are suburban antennas feeding railway stations with smaller vehicles and a much lower density of stops, the introduction of reserved seats in a separate part of the standard vehicle, and contacting clients in their offices or homes as "their" bus approaches the boarding point.

Leaving innovation only to the initiative of the authorities, who would specify new types of services at the moment of launching a new tender, seems a poor choice, not only because authorities in general tend to be less inventive but also because there is a clear asymmetry of information. Operators know the market and the evolving wishes of clients much better than authorities because they are in daily contact with them, and they better understand the costs of providing the various types of services under different circumstances.

This is a common problem in systems with principal-agent relationships (Agrell et al., 2002), but in our case the need for integration of services reduces the degrees of freedom available for the agents. Therefore, in this setting, a good alternative seems to be to let operators suggest new types of services, with or without special attention to integration – but without pre-empting or allowing encroachment on territory served by other operators. These suggestions should be subject to a positive preliminary assessment of their potential by the authorities specifically aimed at avoiding predatory practices over other operators, followed by their introduction for a trial period of some months, with full risk on the side of the operator.

The level of commercial success of these new services should be monitored and, depending on the level of success, the authority would partly or wholly cover the production costs of the experimental period. In the case of a high level of success, greater effort toward integration of these new services into the rest of the network should be made, and the operator would also have its concession package enlarged with the new services that resulted from its initiative, under remuneration conditions similar to those defined in the base contract.

Contract duration is related to the number of sub-networks being tendered. These durations should be a compromise between the time needed by the operators to conceive, test, and amortize new ideas and the risk for the authority and the public of being tied to an operator, without knowing whether a better option would be offered in the event of a new tender. With a higher number of sub-networks subject to regular tendering, the frequency of tenders increases, and the duration of each contract can be somewhat longer. The ideal dimensions of a sub-network could be defined by the optimum levels of productive efficiency, which for the case of buses have been found to be between 100 and 200 buses (ISOTOPE, 1997).

An attitude favoring of innovation by private companies may have a secondary effect of great importance, which is to make the contract an instrument for synergy between public and private parties toward better service for the population and not a basis for permanent adversarial positions between the players. Private companies, through the knowledge, skills, and capital they provide, should be seen and treated as partners, and not simply as suppliers. Both sides should remember that the integrated transport system is justified because it has an added value for the population, and serving that population well should be a success both for the authority and for the operators. At the institutional level, integration is also preferable to antagonism.

6. Conclusions

We have seen that transport systems serve as integrators of social activities and that better spatial integration of these activities would require less transport, thus increasing overall efficiency. We recognize, however, that such spatial integration is very difficult to sustain in the long term given the different life cycles of economic activities and transport infrastructure, and the high costs of relocation.

Integration within the transport system is more achievable, and has at least one quite successful example in the road component, where the vehicles produced are basically compatible with the infrastructure on a world scale, as are the driving licenses issued for the drivers. On a single map we can not only see roads of various administrative levels in one country but also roads from neighboring countries, and all shown to the user as an integrated network. The railways are probably at the other extreme of integration, at least in Europe, with multiple "islands" of barely connectable sub-networks scattered around the continent.

Integration is especially important at the level of transport services, possibly across modes, at all geographical scales. Integration provides a much wider reach for those who are located outside the main poles or axes, and thus constitutes an essential element of distributive justice of mobility across the territorial space. This element of distributive justice, together with the contribution to widening the labor and consumption markets, justifies public interest in the provision and proper functioning of integrated transport networks.

For all the agents engaged, integration means additional costs in planning, control, risk coverage, and, sometimes, also in additional components of service. In purely private terms, this can only be justified by global market expansion, improvement of the market position of the operator, or by additional revenue collected from the user or from the authority. Some examples have been given in this chapter of selective integration occurring without the push from the state, normally associated with access to more powerful market positions by the operators.

When integration occurs at the request of and under the control of a state authority there are both costs and advantages to introducing private companies to supply transportation and other related services. The complexity of the system can easily become quite significant, so it is essential to have clarity of purpose (What are the objectives of integration and how do we measure the level of success in reaching them?), and to start with a good institutional design for the distribution of the multiple tasks that must be fulfilled at the various stages.

Once these foundations are well laid, the engagement of private companies must be made on the basis of contracts, of which multiple forms are available, and considerable experience already exists on their relative merits. A point frequently forgotten is that more stringent contracts on the private side require more competent organizations on the public side to design the contracts, negotiate them, and manage their application through their life cycle.

Some critical points in these processes have been discussed in this chapter, and suggestions presented on how to address them. In the end it becomes clear that good integration of transport systems requires not only the integration of the transport services but the integration of the public and private parties engaged in the design and functioning of the system.

References

Agrell, P.J., P. Bogetoft and J. Tind (2002) "Incentive plans for productive efficiency, innovation and learning," *International Journal of Production Economics*, 78:1–11.

Andersen, B. (1993) "A survey of the Swiss public transport system and policy," *Transport Reviews,* 13:61–81.

Coase, R.H. (1960) "The problem of social cost," *Journal of Law and Economics*, 3:1–44.

European Commission (1995) *The citizens' network – fulfilling the potential of public passenger transport in Europe*, COM(95)601. Luxembourg: Office for Official Publications of the European Communities.

European Commission (2001) *European transport policy for 2010: time to decide*, White Paper, COM (2001)370. Luxembourg: Office for Official Publications of the European Communities.

Gomez-Ibanez, J.A. and J.R. Meyer (1993) *Going private*. Washington, DC: Brookings Institution.

ISOTOPE (1997) *Improved structure and organisation for urban transport operation in Europe*, Transport Research Fourth Framework Programme, Urban Transport, VII-51. Luxembourg: Office for Official Publications of the European Communities.

Potter, S. and M.J. Skinner (2000) "On transport integration: a contribution to better understanding," *Futures*, 32:275–287.

Segal, I. (1999) "Complexity and renegotiation: a foundation for incomplete contracts," *Review of Economic Studies*, 66:57–82.

Spackman, M. (2002) "Public–private partnerships: lessons from the British approach," *Economic Systems*, 26:283–301.

Toner, J.P. (2001) "The London bus tendering regime – principles and practice," in: *7th International Conference on Competition and Ownership in Land Passenger Transport*. Molde.

Viegas, J.M. (1999) "Public transport in the sustainable urban transport policy package: taking an integral policy approach," in: *ECMT-OECD Workshop on Implementing Strategies to Improve Public Transport for Sustainable Urban Travel*. Athens.

Viegas, J.M. (2003) "Competition and regulation in the transport sector: a recurrent game and some pending issues," in: *Proceedings of the 16th International Symposium on Theory and Practice in Transport Economics*. Paris: ECMT.

Viegas, J.M. and V. Colaço (1998) "Le "project finance" pour des autoroutes: recensement des risques principaux et encadrement des études de trafic," in: *Actes Symposium International sur le Financement de la Route*. Paris: ENPC.

Viegas, J.M. and R. Macário (2003) "Involving stakeholders in the evaluation of transport pricing," in: A. Pearman, P. Mackie, J. Nellthorp and L. Giorgi, eds, *Transport projects, programmes and policies: evaluation needs and capabilities*. Aldershot: Ashgate.

Part 2

PLANNING PERSPECTIVES

This Page Intentionally Left Blank

Chapter 10

THE HISTORY OF TRANSPORT PLANNING

MAX G. LAY
Royal Automobile Club of Victoria, Melbourne

1. Before mechanical power

Transport planning is a relatively new component of transport policy. For most of recorded history, transport has been an individual process, self-motivated and self-managed. Some group travel arose when people found it useful to bond together into tribal units and freight travel occurred when man used domesticated animals to carry marketable goods. As armies grew in size, mass travel by troops became significant, and their need for supplies, reinforcements, and messages created the need for regular transport services. The first transport planning probably occurred within the logistics component of these armies. Once the armies had completed their task, conquered lands required transport for administration and the return of booty and taxation.

As a technology and as a service, transport developed slowly and was rarely able to perform significantly better than the private efforts of individuals. Indeed, despite the romantic views in novels and film, until the twentieth century most transport traveled at an overall speed that was little better than walking pace. For example, in the nineteenth century, sailing ships covered the 17 000 km from London to Melbourne in 3 months or 200 km per day – a steady 8 km/h. On land, many travelers were delighted to cover 30 km in a day. Transport planning was more about the logistics of feeding man and beast than any concern with schedules and interchanges.

Once the technology of the Industrial Revolution made coaches possible, those who could afford it could travel in fast (16 km/h) stage coaches – in daylight hours – and might manage 60 km in a day. In 1832 – at the dawn of steam rail – the new mail coaches were managing the 180 km journey from London to Bristol in 12 hours, i.e. at an average speed of 15 km/h. Over time, running by humans outsped even the horse, with Irish "footmen" in England managing 120 km in a day. Meanwhile, land freight with its relatively small loads still could not even match walking speed. Before the advent of steam rail the canal system was developed to allow large loads to be moved on land with assurance, albeit still at walking speed.

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

It could be said that transport planning began at the beginning of the nineteenth century, when the burgeoning Industrial Revolution produced increased demands for the movement of road freight. Road improvements were clearly needed. The first transport planning practitioner was probably the US Secretary of Treasury Albert Galatin, whose 1808 *Report on Internal Improvements* clearly enunciated the basis for an economic assessment of roads and led to a comprehensive plan for a national system of roads and canals in the USA.

In England, Thomas Telford attacked this road investment issue in a systematic way, realizing that the inherently slow-moving traffic required good pavements and relatively flat roads. To assess how smooth and flat roads should be, Telford and his colleagues conducted a series of experiments to measure the rolling resistance. From this data, they calculated the maximum slopes that should be used with particular surfaces, for a given number of hauling horses. The underlying analysis was similar to a modern benefit–cost analysis

At 6 km/h, transport decisions could be contemplated, debated, implemented, and revised in the real time associated with the journey. At the time of the emergence of steam, the biggest transport planning need would have been associated with the distribution of mail. Stephenson and his Rocket then created a whole new transport paradigm, both in terms of speed and on-land freight capacity.

In freight terms these mid-nineteenth century needs were to move heavy or high tonne-km products from a known producer to a known distribution center. It was basically a problem in engineering and logistics, and was well catered for by the new steam power – on rail, on canals, and at sea. In urban transport terms, the low-throughput horse-bus was succeeded by the city train and tram, once iron-makers were able to make rails suitable for street use. The installation and operation of these systems were basically problems in logistics, urban management, and engineering. As initially installed, they catered for middle-class passengers.

In seventeenth century London the fleet of Hackney carriages serving middle-class travelers became so large as to destroy competition and cause severe traffic congestion. King Charles I introduced laws to limit their "general and promiscuous use." The laws had little effect. In 1814 Hackneys were required to carry a registration number, and in 1838 their drivers had to be licensed. Throughout this time, the problems associated with the Hackney carriage continued unabated.

Such transport services were introduced by operators seeking to make an immediate profit, and required little capital other than the cost of the vehicles and their haulage animals.

In the pre-steam portion of the nineteenth century, most significant roads and bridges were toll roads. The tolls were certainly a way of keeping the roads maintained – particularly after McAdam – but were also often a rewarding investment. Traffic volumes were known from existing usage, and planning decisions were related to route diversion as toll charges were applied or raised. In 1844 Arsène Dupuit in France conducted a study of toll bridges which led to him proposing "marginal" costs as the appropriate basis for tolling existing structures. This concept remains at the core of current tolling theory.

Dupuit also measured and quantified the degradation of roads under traffic, another key parameter in road pricing and road planning. The key initial debate was over the rate at which deterioration would occur. Dupuit assumed that it would be linear with respect to the tonnage hauled, and this convenient core assumption was confirmed by de Saint-Hardouin in 1877.

One prescient example of transport planning occurred in 1815 when the British Parliament established the Holyhead Road Commission to plan, build, and operate the road from London to Holyhead, the terminal for the Irish ferry. The commission worked long and hard, and when the road was completed in 1830 the mail time on the route was reduced from 41 hours to 28 hours. Nevertheless, by 1837 all London to Holyhead mail had been transferred to the newly invented steam rail.

2. The age of steam

John Burgoyne was Chairman of the Irish Board of Works. He conducted the first comprehensive traffic survey in 1837, measuring traffic flows, vehicles used, cargo hauled, and origins and destinations. This was shown using variable-width traffic-flow charts of the type still used today. The work was conducted to help him plan the Irish rail network, and correctly predicted a dramatic decline in road use following the introduction of rail. Thus, Burgoyne took transport planning beyond considerations of mobility to the larger concern of land use/transport (LUT) interaction.

In 1857 Henry Carey gave the basis for most predictions of the quality and quantity of the travel produced by various land uses when he suggested that it could be predicted by an inverse square law by analogy to Newtonian physics and the "great law of molecular gravitation." The first empirical confirmation of the gravity model was obtained by studies of migration patterns. It was first used in the 1927 Boston Transportation Study, but the seminal application was by Alan Voorhees in 1955 (Voorhees, 1956). The gravity model was also the basis of the Lowry model introduced in 1964, which directly linked transport to employment, services, and residential location (Lowry, 1964).

The coming of steam power early in the nineteenth century saw much greater needs for capital consumption and therefore for credible planning predictions of travel demand. Initially, coal haulage was the major load, but the attractions of passenger transport soon became apparent. Steam power readily filled an inter-city role and after Burgoyne's origination–destination work, the planning problems largely related to the problems of finding routes for the trains to enter urban areas. In cities such as London and Paris, lines continued until they met insurmountable inner-city land prices.

By and large, steam power was too cumbersome to cater for urban transport needs. The introduction of electrically powered trams and streetcars resulted in a significant drop in fares and raised travel speeds to 20 km/h. Public transport was now a facility for the urban masses. Planning largely related to selecting routes that had sufficient patronage demand to meet the capital and operating costs.

As city populations grew to meet the demographic demands of the time, it soon became evident that there was much private and public advantage in developing land at the city fringes to meet those demands. Although the motives were largely related to land speculation, transport planning was required to ensure that the fringe land had adequate apparent accessibility. Predictions of patronage levels were not required. A typical pattern would be that the owners of the fringe land would provide the capital to install rail or tram to their land. Operating costs were to be met from the fare box. Given that the fringe developments were usually associated with low-density housing, few transport systems have ever had the patronage to achieve this. Thus, after a short term, most of the new urban systems were either abandoned or heavily subsidized by government. With government ownership, many of the systems were consolidated into larger systems. The remainder were soon abandoned.

The railways also forced transport planning to the higher strategic level of considering what to provide and with what priority and at what cost (although the canal system had required some thinking of this nature, the suite of potential solutions was too small to make the task intellectually challenging).

The idea of conceptualizing what travelers might do in order to plan systems for them was taken a stage further by the Prussian military late in the nineteenth century, when they developed war games using detailed transport and logistics components (Wilson, 1968). War games are, of course, preferable to war. Similarly, transport models were developed because we can rarely experiment with a live transport system to discover the effect of a future policy change or level of service alteration.

In the largest, most developed, and most prosperous cities, such as Paris, London, and New York, underground transport systems developed to meet the needs of the traveling public without the need for major land acquisition. Such systems developed incrementally and competitively. New York had seven street railways in operation by 1858, and these formed the basis of the subsequent subway system. By 1860 there were 14 horse-railroads. The first elevated route opened in 1871. Transport planning was clearly via the market place. Perpetual franchises were granted by a government commission.

Thus, by the second half of the nineteenth century, transport planning could be seen as occurring in the four stages that remain in evidence today:

(1) the political stage, which considers broad social and economic goals, planning at a conceptual and policy level;
(2) the development stage, which considers such factors as land use and transport levels of service, planning at a strategic and systems level;
(3) the transport stage, which considers transport demand management and infrastructure provision options, planning at a corridor level;
(4) the facility stage, which considers track capacity and track design, planning at an implementation level.

In the nineteenth century, when there was strong central government and a concentration of capital, stage 1 was usually critical. Today, transport goals are diverse and far from clear. Thus the remainder of this review mainly focuses on stages 2 and 3.

Nevertheless, many key transport planning decisions continue to be made at a political level. For example, over recent decades the US federal government has introduced a number of legislative moves to encourage public transport development in US cities.

3. After the car

The impact of the motor car on transport began in the years immediately following the First World War. Boston was one of a number of cities that began transport studies in the 1920s, as a result of the ominous burgeoning influence of the motor car. In addition to trip origins and destinations, the studies investigated road capacities, traffic systems, and traffic forecasts. The leader of the thrust was Herbert Fairbank at the Bureau of Public Roads in Washington.

In retrospect, many of the unfulfilled dreams of the 1920s planners were best left unfulfilled – they included triple-deck urban streets and suburbs with streets on a 60° diamond grid to minimize travel distances.

In a far-sighted move the US Congress, beginning with the Federal Aid Highway Act of 1934, provided 1.5% of road funds for surveys, plans, and investigations. By 1940 all US states were participating in such activities. The surveys included road inventories, traffic counts, vehicle usage surveys, and car ownership trends.

Another significant tool that came out of the US Department of Agriculture in the 1930s was the use of regression analyses of large data sets to find relationships between data. The methods were then applied to transport modeling in particular and econometrics in general.

By 1960 computers were able to find up to 100 "explanatory" variables from a data stack, using stepwise regression and without the need for any phenomenological model. Wonderful relationships between variables were found, and the approach was (and still is) widely applied, often with scant recognition of the statistical risks involved. Unrelated variables have been forced into close relationships and counter-intuitive results promoted as physical fact. Consequently, many modeling predictions based on data regression must be treated with great caution. In recent times the author has pointed out three such cases in the field of pavement systems analysis leading to suggested pavement investment and maintenance strategies by life-cycle costing (Lay, 1999). One of the cases involved a World Bank model widely used around the world.

Urban travel studies began in the USA in 1944, when the Highway Act of that year made federal funds available for urban projects. The work showed that there was vastly inadequate data available on trip origins and destinations, and on the basic factors affecting travel (Holmes and Lynch, 1957). The work involved both home-based and roadside interviews, and it would appear that tools were not then in place to analyze adequately the data collected.

Lessons learnt from the 1930s and 1950s were that there was a great deal of data to collect and analyze, and that processes needed to be put into place to make that collection and analysis routine. As a first step, in 1961 the US Bureau of the Census began collecting transport information.

It also became evident that much better means were needed to predict the benefits of investments in transport. Difficult unanswered questions had arisen, particularly (Holmes and Lynch, 1957):

- How and why did operating costs vary?
- What was the value of time?
- How did property values vary with transport provision?
- How were projects to be prioritized (e.g. by sufficiency analysis)?

Following the last point, by about this time the method of benefit–cost analysis was being applied to transport schemes. It had been developed early in the twentieth century for the assessment of proposals for ports, harbors, and irrigation. It was brought to the transport sector by a major report issued by the American Association of State Highway and Transportation Officials in 1960. In his landmark report, *Traffic in Towns*, Colin Buchanan noted that benefit–cost analysis had first been used for transport studies in Britain in 1960 (Buchanan, 1963).

Many grandiose urban transport plans were produced in the USA in the 1950s and 1960s – and much of the rest of the world soon followed suit.[a] In retrospect, it can now be seen that these plans were predicated on the simplistic assumptions that:

- providing extra road space would satisfy urban travel demands (we now know that this can only apply if all suppressed demand has been met);
- the diversion of urban space to roads was of first priority;

[a]For example, US planners rapidly replicated their plans – warts and all – in all Australian cities.

- the funds existed to put the full plans into practice;
- simple built solutions will solve complex and interactive urban problems.

While these assumptions were often valid for small cities (under 1 million people), for large cities they frequently floundered in all four areas, and the partial solutions resulting from these unfulfilled dreams left many undesired legacies. By the late 1950s there was growing concern over the impact of urban freeways on US cities. In 1964, New York freeway builder Robert Moses explained big city planning in the context that "when you operate in an over-built metropolis you have to hack your way with a meat axe." In recognition of the problem, the 1962 Federal Aid Highway Act created a federal mandate for urban transport planning.

Throughout the world, transport was presenting civilized society with major dilemmas. The demand for travel appeared insatiable, but the accepted charging mechanisms would not generate the funds to satisfy that demand. And when solutions were found, their effect on the environment often made them unacceptable. How did the increasingly numerate discipline of transport planning react to these challenges? The following case study will illustrate a typical, but in retrospect inappropriate, response in which civil engineering solutions predominate.

4. Melbourne case study

The author has recently had the opportunity to review the impact of a sequence of such transport plans produced for the city of Melbourne (current population about 3 million). The results are summarized below (Lay, 2003), and as a case study they typify the development of transport planning during much of the twentieth century.

4.1. 1929 Melbourne Plan of General Development

The plan reacted positively to the motor car age, but in retrospect its qualitative and intuitive basis led to a curate's egg of far-sighted recommendations and untested fantasies. While many regard the 1929 plan as visionary, most of its proposals were far too original to lead to their ready implementation, particularly as the plan had the misfortune to be released during the Great Depression. Similar plans in the USA had proposed solving traffic problems by building triple-decked arterial roads.

Such visionary plans were probably a necessary stage in the development of transport planning.

4.2. MMBW 1954 Town Plan

Following the hiatus of the Depression and the Second World War, planning began again in the early 1950s. It was a cautious "trend" plan, underpinned by intuition. Using the LUTS[a] methods then available, the plan under-predicted Melbourne's growth, with the city reaching its design population in 1971. Atkins (1987) notes that accurately predicting such a fundamental factor as population change had proven to be a major flaw in many plans. After its presentation in 1954, public comments and appeals meant that the plan was not finalized until 1959.

The lesson from 1929 had not been learnt, and Melbourne again had a road plan that was too grandiose, too inappropriate, and too expensive to be readily implemented – for instance, it proposed some 460 km of largely unattainable freeways. Some of these proposals exemplified the worst excesses of the freeway age and would have ravaged inner Melbourne. The city was fortunate that the plan remained a plan. The plan would not have passed the sustainability constraints used in today's planning processes.

4.3. MTS 1969 Transportation Plan

This conventional LUTS-based plan used 1985 as its horizon year, and was based on optimistic population assumptions (e.g. a Melbourne population in 1985 of 3.6 million that has yet to be reached in 2003). The plan had been produced by a local team in association with a group of US transportation engineers, who concentrated on the surveys and modeling and supplied the service standards, which would have been a 50% improvement over those then being experienced in Melbourne. At the time, similar plans were in the process of being imposed on many cities throughout the world. Generally, the freeways were placed on a 5 km grid in the belief that a grid would avoid the congestion associated with the center of a radial system of freeways, dispersing users over a wider area. This argument was consistent with the then-current depopulation of urban cores and outward spread of the suburbs.

The plan recommended 510 km of urban freeways as its main transport "solution." Melbourne then had some 80 km of freeways in operation or under construction. The planners had assumed that a complete reconstruction of inner Melbourne was desired. Significant public transport solutions were also advocated.

In hindsight, the plan was unacceptable in terms of its cost (it would have meant a trebling of the then current expenditure rate on freeways), the immensity of the

[a]The "S" added to LUT is taken to stand for "system" or "study."

system proposed, its effect on the environment and the urban form, and its general inappropriateness for a city such as Melbourne that already had a good arterial road system. Today, most citizens are prepared to accept lower levels of road service in a trade-off with the urban character they desire for their city. Strong community opposition arose to such redevelopment after the release of the plan.

The plan forced the community to articulate its desires for the city and its future form. For instance, there was also a recognition that the purpose of any inner-city freeways should be to allow traffic to avoid the downtown area, not to feed more traffic into it.

4.4. *From 1973 to date*

In March 1973 a new political leadership announced the end of all dreams of a Melbourne enmeshed in a web of freeways, and cut the planned network in half on sociological and environmental grounds. Transport planning since 1974 has been an incremental process, with each increment arising without a strong public context, but rather from the internal workings of the various transport ministries.

In subconscious support of this incremental planning, a series of nine separate micro-plans were issued over the 24 years between 1978 and 2002. The underlying transport planning principles in this series appear in retrospect to have been:

- future growth in car travel must be curtailed;
- public transport usage must be increased;
- residential densities should be increased.

The principles have appeared in many well-intentioned attempts by legislatures around the world to achieve transport planning by parliamentary decree; the author is unaware of any examples of where such decrees have been successful.

A major defect in the city's transport planning has been minimal use of freight modeling and little recognition of the role of regional economic development in the transport planning for a major city within a region. Despite this, recent transport developments have had major impacts on freight movement and on factory/warehouse location.

Nevertheless, or perhaps as a consequence, transport in Melbourne is relatively effective and the metropolitan area is pleasant and livable.

5. Model applications

Up to the 1960s, transport planning in the USA had largely – and in many instances totally – concerned the car and the truck. A breakthrough came with the

Urban Mass Transportation Act of 1964, which encouraged planning for area-wide urban mass transportation.

Meanwhile, comprehensive transportation modeling had begun in the USA with the Detroit Metropolitan Area Traffic Study (1953–1955) and the Chicago Area Transportation Study (1959–1962). These studies were part of a common thrust in a number of areas of social and economic enquiry to apply mathematical modeling techniques and systems approaches to previously insurmountable problems (Lewis et al., 1990). Detroit in particular was the first major study in which computer power had made it possible to study the land-use/transport (LUT) interaction in a systematic way. British models commenced in 1962 with the London Travel Survey.

The LUT models have been applied at six levels (Lewis et al., 1990):

(1) urban areas, as with the Detroit and London models mentioned above;
(2) metropolitan areas, following level 1 but with larger zones, as with the Melbourne studies discussed above;
(3) major travel corridors, e.g. in the initial studies for the Channel Tunnel;
(4) regions, e.g. for land use planning and resource allocation;
(5) national, e.g. to determine transportation and taxation policies;
(6) international, e.g. European Community studies on trans-national transport operations.

The models have always held out great promise, particularly from a policy viewpoint, permitting policy-makers to test the transport and land use outcomes of proposed policy changes, without the need for public experiment. Likewise, the models potentially allow operators to explore the service consequences of proposed operational changes. Unfortunately, the LUTS models frequently failed to fulfil these expectations. Atkins (1987) noted that these early model applications were closer to acts of benevolent paternalism.

6. Computing power

It is worth lingering here to comment on the influence of computing power. In mathematical terms the problems involved matrices of data, which for even the simplest of practical cases were large (e.g. 200×200), and a need to find optimal solutions to problems with many mathematically acceptable solutions. The Second World War and its immediate aftermath had a major impact on these two problems.[a] Firstly, desk calculators capable of (albeit slowly) tackling the

[a]For example, much transport theory arose from the Berlin Air Lift.

calculating problems came into common use.[a] Secondly, logistics theory, and in particular linear programming, produced the tools needed for tackling such issues as travel desire lines, shortest paths, and traffic assignment.

In the 1950s, access to large computers was rare. When access became more commonplace in the 1960s, programming was in machine-related codes conducive to many errors that were minor in commission but which could each stop the program, and date entry was fragile, tedious, and rife with accidental mistakes. The author can personally vouch for the fact that to achieve one successful multi-hour run of one program with one data set was a rare event that deserved celebration. The lucky person rarely tempted fate by running the program again or altering the data. The problem was compounded by the enthusiasm with which data were collected. Overwhelmed by the output of that enthusiasm, the task and the data set were frequently massaged to fit the computer.

Thus, if the traffic on a route exceeded the capacity of the route, this was merely noted as "congestion." It was not feasible to rerun the program to redistribute the traffic once capacity had been reached. Processes and results from the 1950s through to the 1980s must always be seen in this context of limited computing capabilities. Thus, iterations were necessarily rare, and data changes were relatively small.

The transition during this period has been well described by Richardson (1990), who wrote:

> Prior to the early 1950s, transport planning was seen as a subset of town planning and was performed using predominantly qualitative and codified methods. With the advent of computers, however, it was realised that it was possible to process significant amounts of data, thus enabling quantitatively oriented planners to analyse traffic movement patterns in a more systematic manner. The mathematical modelling of the transport system thus became an important feature of transport planning. At times, in fact, it seemed that the modelling became more important than the planning, with substantial amounts of time and money being spent on the collection and analysis of data, and relatively little being spent on the generation of alternative scenarios and the evaluation of the impact of the alternatives.

7. The four-step LUTS

Returning to the LUTS models, these developed into the "four-step" models. The four steps were:

[a]The author recalls stories of one of the early freight studies by Lazlo Foldvary of the Australian Road Research Board that involved a large room full of desk calculators and their operators, each performing a single arithmetic process, before passing the result over their shoulder to the next calculator. Two independent streams of calculators were being used to avoid error.

Demand side:

(1) data collection and forecasting for land use and for trip production at land use origins, and trip attractions to land use destinations;
(2) trip distribution between land use, origins, and destinations, commonly using the gravity model.

Supply side:

(3) splitting of the trips from an origin to a destination between the existing and forecast transport modes;
(4) assigning the trips to particular modal route.

While it is easy to now display the four steps, it was a major conceptual breakthrough to separate the tasks into these four discrete components. To its detriment, this was seen as a one-way task leading to a solution at the end of step 4. It is now obvious that many iterations are needed between the steps.

Forecasting changes in land use in step 1 was largely based on Stouffer's (1940) intervening opportunities model. During the 1970s the use of the gravity model in step 2 was further developed, seeing the number of trips between an origin and destination predicted by a model which replaces the inverse square of the separating distance by a function of travel cost. Using entropy maximization theory this function was chosen as the negative exponential of generalized travel cost. The constants in the equation were obtained by using the known trips in the base case. Other applications used the intervening opportunities model.

Mode and route choice decisions in steps 3 and 4 were initially based on the assumption that travelers would minimize their trip costs. This was first formally articulated by Wardrop in 1952 in his two principles describing the "game" played between users and operators:

(1) users selfishly choose a route on which the trip times are at least as good as on any other route;
(2) operators seek a system where the total trip time is minimized.

While transport fares and tolls are straightforward, the need to establish useful trip costs has required a great deal of work to be devoted to estimating the value travelers placed on their travel times and to assessing vehicle operating costs (Hensher, 1989). In addition, all these parameters had to be soundly and consistently based on economic theory, and this resulted in some major theoretical contributions. The concept of generalized costs provided a conceptual breakthrough. Costs were seen as the sum of actual costs and perceived costs of such factors as walking, waiting, and comfort. Some of these could be expressed in units of time, and this then raised again the question of the value of time. There were many significant contributions to this debate, particularly from Hensher and his colleagues. By and large, time is now valued at the same rate as is used by the travelers' employers and the freight's owners.

The best known example of the four-step LUTS model has probably been the well-used version within the US Government's Urban Public Transport Planning System (UTPS), produced by the Urban Mass Transportation Administration (UMTA).

Initial public disillusionment with the transport models has been discussed in the Melbourne case study given above. Even the relatively simple task of forecasting population growth had proved an impossible task. In the 1980s there was also a period of considerable professional disillusionment with the output of such models. Peer critics described them with words such as "inadequate," "policy insensitive," "cumbersome," "biased," and "useless" (Atkins, 1977). Atkins (1987) quoted John Wootton writing in 1980:

> Twenty years' experience when using transportation models to predict future travel demands has convinced me of the inadequacy of existing models and the need to treat their results with caution.

Perhaps the most damning review was that produced by Mackinder and Evans in 1981:

> an assumption of zero change from the base year would not have produced larger forecast errors ... for trips the errors would have been considerably less ... [than those obtained from a transport model].

The problem did not go away. One author in 1990 gave a typical view when he wrote (d'Este, 1990):

> Models of transport systems have a reputation of being incomprehensible black boxes that consume vast amounts of money and resources and produce mountains of printout in a largely unintelligible form.

8. Further developments

An important explanatory concept that arose during the 1970s was Zahavi's (1974) "time budget" statement, which pointed out that the 24 hour day meant that there was only a limited time that people could spend in traveling, particularly on the journey to and from work. Furthermore, it appeared that that time (summed for both trips) was universally relatively constant at about 1 hour. One obvious manifestation of this "law" was that in cases where travel speed was increased, e.g. with a radial route, people located their homes further out from the city center.

Another important development in the 1970s and 1980s was a set of strong explanatory links between population, car ownership, and car usage (Mogridge, 1983).

The four-step LUTS model was followed in the early 1970s by the second-generation or "disaggregate" LUTS model, and then in the late 1970s by the third-generation or "behavioral" LUTS models (Hensher and Stopher, 1979). These data-hungry advances focused on steps 3 and 4 of the model, and tried to predict the mode and route choice decisions. Disaggregate models are based on market segmentation of the aggregate population at an origin or destination. Behavioral models then attempt to predict the behavior of the market segments in particular circumstances. Often the behavioral predictions require a further set of surveys to be undertaken. In the 1980s, activity models were created that treated transport behavior as just one part of the behavior of a family or firm. Hensher (1978) gives a useful review of this period. In recent times, market research and consumer choice principles have been used to better predict traveler behavior in changed circumstances.

Environmental features began to be added to the models in the 1970s. This has never been a simple process. There are two main approaches. Firstly, environmental constraints can be imposed to prevent some options from being modeled and considered (e.g. traffic banned from some routes). Secondly, the model outputs can be utilized to predict various environmental outputs such as noise levels, noxious emissions, or land value impacts. The tools to predict these outputs improved dramatically during the 1980s.

Inspired in part by work in the merchandising/advertising industry, survey methodology for data collection advanced dramatically in the 1980s (Ampt et al., 1990). Of course, much of this change went hand-in-hand with the move from aggregate to disaggregate to behavioral to activity models discussed above. To the traditional counts of traffic past a point had to be added sociological and logistical data on the internal functioning of families, supermarkets, and factories.

A further model development in the 1980s responded to the fact that as more travelers use a route, travel times begin to increase, and the route becomes less desirable. Indeed, most routes have a maximum capacity that they can carry in any circumstances. This problem can be resolved by iterative methods – e.g. by rerunning the choice model as conditions change, or by running the models over time and responding incrementally as traffic levels increase. These approaches had to wait until sufficient computing power was available.

Sensitive situations are now being tackled by models taken from behavioral studies and traffic engineering and which simulate the behavior of individual travelers. First introduced for air traffic control in the 1960s, in the early days such models were poorly calibrated and slow to run. This is no longer the case in a number of prime applications. The current limit of these applications is possibly in intelligent transport systems, where motorists are beginning to request real-time advice. For example, if it is suggested that a motorist divert to another route because of congestion ahead, if the advice is to be given to many motorists, it will be necessary to model the situation after they take the advice, before giving that advice. Predictions will need to be available in advance of real time.

The other importance of simulation has arisen from the realization that many grand schemes have been halted by local issues. Behavioral predictions and associated policy advice that do not cover these issues is seriously flawed, socially and politically.

9. Problems and solutions

Perhaps because the transport modeling tools are so often used for contentious policy-making, they have continued to be the subject of public debate. Often the core of the criticism has lain more in the process than in the model. There have been two basic problems. The first has been an unwillingness by model users to acknowledge and publicize the limitations of the model data and the modeling process. The second has been an unwillingness to see the models in context. For example, much of the debate in the UK late in the last century over the predictions of models used for examining motorway proposals hinged on the initial failure of those applying the models to fully consider the traffic generated by the proposal. And yet predicting traffic generated is a fundamental component of the four-step LUTS model.

Other longer-term problems with the LUTS models had been associated with their approach to data. From the beginning, traffic data were relatively easy to collect, and the studies sometimes degenerated into dead-end exercises in data collection. To manage the data, the regions being analyzed were divided into zones, and because computing power still loomed large as a problem, the zones were small in number and large in size. In retrospect, this was a major contributor to some of the poor modeling results. When the author was involved with the bidding stage of the large Melbourne City Link BOOT Project in the mid-1990s, the combination of bidders, banks, government, and assessors meant that about six transport models were applied to the same problem. It was found that the main way to achieve compatible predictions from the various models was to reduce zone sizes. Richardson (1990) has previously noted that the definition of "zone" could dominate model output. Indeed, it would seem that only simulation (see above) will produce usable results at an operational level.

Another long-term problem has been associated with model calibration. Data collection provides inputs and outputs from the current LUT configuration. Hence, it has been argued, LUTS models can have their internal "constants" adjusted to bring the initial outputs into line. This causes a potential problem whenever there is no phenomenological basis for checking the adjustments to the "constants." The problem has been heightened by the fact (mentioned above) that many of the internal behavioral responses in the models are based on regression analyses of data, and their adjustment requires a great deal of statistical caution. Further, some models were so calibrated that they were no more than mathematical representations of the present, drained of any useful predictive power.

A broader problem with the earlier models was that they were so large and cumbersome that they were seen as ends in themselves. They produced a master plan with a single solution. As in the Melbourne case study given above, those solutions were often – with good reason – indigestible politically and socially. Moreover, while grand visions are important – in 1907 Chicago planner Daniel Burnham admonished "Make no little plans, they have no magic to stir men's blood" – life proceeds incrementally, and needs and opinions are continually changing. Thus, grand visions are essential, but plans and planning are continuous processes. Current technology permits this, and thus the post-1974 aspect of the Melbourne case study is perhaps not surprising.

Understandable public and political enthusiasm for public transport as the solution to many transport issues has also demonstrated how prone models can be to subconscious political and community pressures. There are now many examples of large public transport investments where the patronage experienced has fallen far short of that seriously predicted at the proposal stage. In retrospect, over-enthusiastic forecasts are easily detected. And yet these lessons seem slowly learned. A one-sided review of this situation in a US context is given in O'Toole (1999).

To end on a positive note, there have been many highly useful recent advances. Much data collection is now automated, detailed, accurate, and easily processed. There is no longer any reason to use aging databases. Microcomputers have provided more-than-adequate power to all desktops, so that the need to fully validate a model (rather than simply calibrating it), to rerun scenarios, to use other data, and to explore other options are no longer impediments to proper decision-making. Perhaps most importantly, inputs and outputs can now be presented in a variety of ways that place no restrictions on the detection of model inconsistencies and allow the full understanding of model outputs.

The successes of transport planning in the past have been associated with understanding traveler behavior and modeling and predicting that behavior. This has resulted in good incremental developments and increasingly wise policy decisions. The transport planning process, once the data are collected and the model predictions become available, is a search for possible solutions and a prioritizing of those solutions. This stage requires a great deal of skill and judgement. In addition, the process requires the planners to cajole the community and its legislators into determining and then articulating their vision for the community they want, and to define how they will measure success and failure in seeking that vision.

Often the appropriate solutions focus on level of service (e.g. safety and comfort), simple ticketing, and easy interchanging rather than on new infrastructure provision. In this context, and a number of others, transport planning has become much more market driven in recent times.

Many past failures of transport planning have arisen because the basic problem being addressed has been poorly defined or articulated. Fortunately, the community is now beginning to articulate high-level principles to be followed such as respect for the environment, sustainability, the triple-bottom-line, positive demand-management measures, the long-term and system-wide nature of public transport investments, and the total unacceptability of allowing profligate current trends to continue unabated. There are clear signs that these principles have filtered down through the transport planning processes and are now influencing individual planning recommendations.

In my personal view, the one historical lesson yet to be learnt by transport planners is that in the end market forces will exert more influence than government policy.

References

American Association of State Highway and Transportation Officials (1960) *Road user benefit analyses for highway improvements*. Washington, DC: AASHTO.

Ampt, E., A. Richardson and A. Meyburg (1990) *Selected readings in transport survey methodology*. Melbourne: Eucalyptus.

Atkins, S. (1977) "Transportation planning models – what the papers say," *Traffic Engineering and Control*, 27:460–467.

Atkins, S. (1987) "The crisis for transportation planning modelling," *Transport Reviews*, 7:307–325.

Buchanan, C. (1963) *Traffic in towns*. London: HMSO.

Chicago Area Transportation Study (1959–1962) Vol. 1, *Study findings*. Vol. 2, *Data projections*. Vol. 3, *Transportation plan*. Chicago: Harrison Lithography.

d'Este, G. (1990) "Meeting user needs in transport modelling," in: *Proceedings of the Australian Transportation Research Forum*. Sydney.

Detroit Metropolitan Area Traffic Study (1955–1956) Part I, *Data summary and interpretation*. Part II, *Future traffic and a long range expressway plan*. Lancing: Speaker-Hines and Thomas.

Hensher, D. (1978) "Thoughts on the merits of transport planning packages," *Environment and Planning A*, 10:1155–1169.

Hensher, D. (1989) "The behavioural and resource value of travel time savings," *Australian Road Research*, 19:223–229.

Hensher, D. and P. Stopher (1979) *Behavioural travel modelling*. London: Croom Helm.

Holmes, E. and J. Lynch (1957) "Highway planning: past, present and future," *Proceedings of the American Society of Civil Engineers*, 83:1298-1-3.

Lay, M.G (1999) "One of our models is missing," *Road and Transport Research*, 8:96–98. (Response by J.R. McLean and T.C. Martin: *ibid*., 8:99–110.)

Lay, M.G (2003) *Melbourne miles*. Melbourne: Australian Scholarly Press.

Lewis, S., P. Cook and M. Minc (1990) "Comprehensive transportation models: past, present and future," *Transportation Quarterly*, 44:24–65.

Lowry, I.S. (1964) *A model of metropolis*. Santa Monica: Rand.

Mackinder, I. and S. Evans (1981) *The predictive accuracies of British transport studies in urban areas*, TRRL Report SR 699, Crowthorne: Transport and Road Research Laboratory.

Mogridge, M. (1983) *The car market*. London: Dion.
O'Toole, R. (1999) *The vanishing automobile and other urban myths*. Portland: Thoreau Institute.
Richardson, A. (1990) "Transport planning and modelling – a twenty year perspective," *Australian Road Research*, 20:9–21.
Stouffer, S. (1940) "Intervening opportunities – a theory relating mobility and distance," *American Sociology Review*, 5:845–867.
Voorhees, A.M. (1956) *A general theory of traffic movement. 1955 proceedings*. New Haven: Institute of Traffic Engineers.
Wardrop, J. (1952) "Some theoretical aspects of road traffic research," *Proceedings of the Institution of Civil Engineers, Part II*, 1:325–362.
Wilson, A. (1968) *The bomb and the computer.* New York: Delacorte Press.
Zahavi, Y. (1974) *Travel time budgets and mobility in urban areas*, FHWA Report (8183). Washington, DC: US Federal Highway Administration.

Chapter 11

THE EVOLUTION OF TRANSPORT NETWORKS

DAVID LEVINSON
University of Minnesota, Minneapolis, MN

1. Introduction

Between 1900 and 2000, the length of paved roads in the USA increased from 240 km to 6 400 000 km (Peat, 2002), with virtually 100% of the US population having almost immediate access to paved roadways. Similarly, in 1830 there were 37 km of railroad in the USA, but by 1920 the total track had increased more than 10 000 times to 416 000 km; however, since then, the railroads have shrunk to about 272 000 km (Garrison, 1996). This picture is also familiar to other countries. The growth (and decline) of transport networks obviously affects the social and economic activities that a region can support; yet the dynamics of how such growth occurs is one of the least understood areas in transport, geography, and regional science. This is revealed time and again in the long-range planning efforts of metropolitan planning organizations (MPOs), where transport network changes are treated exclusively as the result of top-down decision-making. Changes to the transport network are rather the result of numerous small decisions (and some large ones) by property owners, firms, developers, towns, cities, counties, local transport departments, and MPOs in response to market conditions and policy initiatives. Understanding how markets and policies translate into facilities on the ground is essential for scientific understanding and improving forecasting, planning, policy-making, and evaluation.

Charles Darwin laid out the idea we refer to as evolution, survival of the fittest, or natural selection in his *Origin of Species*. He constructed natural selection in analogy to the artificial selection that animal breeders use to create new varieties. The idea that species randomly change over time, and the fitter variations are more likely to survive and propagate than the less fit variations, has now become commonplace. Darwinian evolution has been used as a metaphor in statistical analysis (the genetic algorithm) (Holland, 1975), artificial intelligence (Minsky, 1986), and in brain development (neural Darwinism) (Edelman, 1987). In particular, the logic of brain development seems particularly relevant, as the neural connections in the brain (the neural network) are analogous to other

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

networks such as transport. The neural Darwinism argument suggests that at birth there are many connections in the brain, some of which are more useful than others. More useful connections are reinforced, while less useful connections are deprecated. These approaches contrast with creationist arguments, that species (along with their ecological niches) were designed, or that we are born hard-wired, that are no-longer favored in the biological community.

Without becoming theological, it is clear that the idea that planning, engineering, and the intentions of decision-makers drives the topology of networks is a top-down creationist viewpoint, in contrast to a model which suggests that networks evolve, with successful facilities being expanded, and less successful transport sections allowed to wither. This chapter considers the theory and evidence surrounding network evolution models. The aim of network evolution models is to describe reality rather than optimality; there is no obligation to maximize welfare directly. This is in contrast to the long line of research on thenetwork design problem (Abdulaal and LeBlanc, 1979; Friesz et al., 1998; Huangand Bell, 1999). However, the deviation of actual decisions from welfare-maximizing decisions are worth noting.

The evolutionary perspective considers how species come to be; how ecological niches are filled. Using this metaphor, one could think of links as being the equivalent of species, and the links they are connected to as being higher or lower on the food chain (traffic being consumed in a predator–prey relationship). Alternatively, we can consider the network as an individual organism that develops over time. An analogy for that circumstance is to what extent "nature" or the genetic programming drives development of an organism (measured in various ways: physical structure, intelligence, personality, language, etc.), in contrast to "nurture" or the influence of relatives (especially parents), friends, and peers, as well as the availability of resources (food, clean air and water, education, etc.) on that same development. If we apply the nature versus nurture argument to networks, the question is to what extent simple rules (the rules by which travelers choose to use certain links over others, the rules by which resources for network expansion are obtained, the rules which give us the cost of network expansion, and the rules by which investments are made) drive development, in contrast to decisions being made for political or other circumstances determining where the network will be expanded and contracted.

As with the nature versus nurture argument in human development, this may be a false dichotomy. It is clear that without genetic programming, intelligence would be impossible. But without resources, education, and care, intelligence would also be impossible. At best we can assess the marginal contribution of each to something that does not vary too much across the general population.

This chapter considers macroscopic and microscopic perspectives on network evolution in turn. The macroscopic perspective examines S curves and changes in investment patterns over time. The microscopic perspective problematizes node formation, link formation, and link expansion and reviews each. The chapter concludes with some notes about future research and applications.

2. A macroscopic perspective

The macroscopic perspective on network evolution has been examined in a great deal of research, especially at the Institute for Applied Systems Analyses (IASA) in Austria (Grübler, 1990). What has been most noted is the emergence of "S curves," which relate time to deploy a network (or any technology) with market saturation. For a period of time (the growth phase), as knowledge of a technology (a mode) and realization of its benefits spreads, the rate of adoption increases. Each project acts as a demonstration to potential new users. Furthermore, the advantages to adoption may increase with the number of users if there are network or inter-firm scale, scope, or sequence economies. As the technology diffuses, those who expect to attain the most benefit adopt it first. After a point, diminishing marginal returns set in. It is expected that, after complete exposure, technology is adopted by those who gain the most, and then by those who gain less and less from it, until it is fully deployed.

Diminishing marginal returns limit growth, but there is also the issue of decline. Canals, for instance, were made obsolete by railroads, illustrating that the life of a technology may be cut short by competition. Alternatively, as in the case of plank roads, a technology may collapse because a technological problem is discovered shortly after deployment (wooden planks deteriorated much sooner than expected). Figure 1 illustrates S curves for a number of transport technologies in the USA.

One would expect S-shaped curves, because transport is a product like others; it enters and floods a market. Observation suggests that it takes 60–70 years for a transport system to run its growth cycle from launch to saturation in the first major market. However, in places where the technology is adopted later, growth occurs faster because the course is well trodden and learning can be cut-short by simpler copying.

To aid in understanding the behavior of a system as it runs its S curve, the life cycle metaphor may help. Broadly there are three main phases: birth, growth, and maturity. There may be subsequent phases of decline and death, as the S curve runs backwards (perhaps it should be called a Z curve). The period of birth is one of possibility: many new technologies and networks are conceived, yet few are realized; there is an explosion of possible evolutionary paths, yet only one is taken. To illustrate, consider automobile technologies. At the onset of the twentieth century, the form of the auto was unclear. Loosely, it would be similar in size to the horse and carriage, but would it be powered by steam, electricity, gasoline, or otherwise? Would it serve business travel or leisure travel predominantly (it

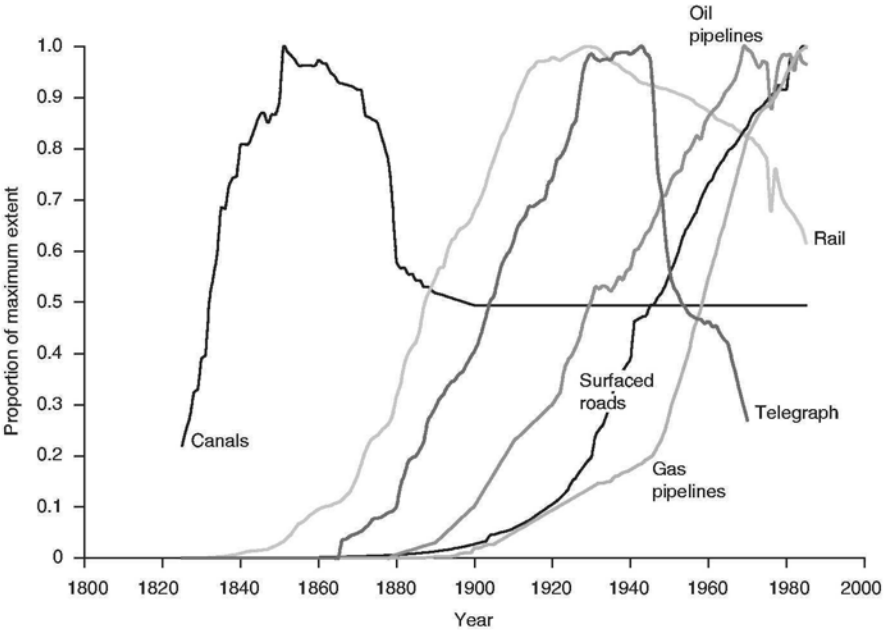


Figure 1. US networks as a proportion of maximum extent.

started as a toy for weekend trips, but later took on a more serious role)? For various reasons, a particular technical path is selected, and ultimately locked in; in the case of the auto that path was the gasoline-powered engine. The more gasoline engines that have been purchased, the more valuable it is for new autos to have gasoline engines, since the gasoline distribution network is already in place. These network externalities are established by the point of transition from birth to growth. The growth phase sees process improvements and technological honing; economies of scale and scope take hold, and costs drop while benefits increase. In the third phase, the network has saturated its market niche, and the focus switches from deployment to management, from growth to maintenance. Because the system is mature, there is little additional benefit to be gained (by the existing system) from technological advances; advances will rock the boat and move the status quo, with which most players are comfortable. Of course change does eventually occur, but it is often the "next new thing" rather than a modification of the network in place.

Systems seem to face diminishing investment in new technology as a system moves through its life cycle. It would be only rational to make the most cost-effective improvements in a process first, and then the next best, and so on.

While previous improvement may open new opportunities, it is likely that each improvement is slightly less effective than the previous, at least after some point. For instance, the fan jet engines used by jet aircraft are approaching the limits on the thrust that can be obtained from them; the Otto cycle engine is just about as fuel efficient as it can be made to be. Investment in product development is generally high during the early days of the life cycle, but as time passes, more and more attention is paid to processes of production.

The relationship to these S curves for dominant technologies and the economy as a whole has not gone unnoticed. Researchers have associated "long waves" in the economy of prosperity, recession, and depression (Kondratieff cycles) with waves of innovation, often waves of transport and communication technologies. Mensch (1979) argues for waves of innovations that trigger investment and jobs, but as those technologies begin to age there is recession and then depression. Recovery begins as another wave of innovations begins. Yet this process leaves open the question of what causes waves of innovation. Garrison and Souylerette (1996) present the companion–innovation hypothesis, which says that opportunities created when transport – and communication – systems are innovated and deployed trigger waves of innovations. In other words, transport – and communication – systems create new opportunities in other economic sectors, allowing people not only to do the same things better (faster, cheaper), but to do new things that were previously inconceivable. Those innovations drive further changes.

A line of research examines how transport investment affects the economy at large, but tends to treat transport (e.g. highways) as a black box, and makes no distinctions between different kinds of transport investment (Aschauer, 1989; Gramlich, 1994; Ishaq Nadiri et al., 1996; Button, 1998). The input is investment in transport (or infrastructure), and output is gross domestic product. The research has shown that investment in transport has been declining in the USA, and that the effectiveness of transport investments (the rate of return) has also been declining over the last 40 years of the twentieth century. Those observations are consistent with transport (particularly highways) being at a mature stage. Just as transport investment affects the economy, the economy affects investment patterns. Carruthers and Ulfarsson (2001) find that various public service expenditures such as roadways are influenced by demographic and political characteristics. The New Jersey Office of State Planning (1996) also finds a similar pattern in roadways expenditure.

Miyao (1981) developed macroscopic models to take transport improvements as either an endogenous effect of urban economy or as an exogenous effect on the economy. Endogenous growth theory suggests that economic growth is a two-way interaction between the economy and technology; technological research transforms the economy that finances it (Aghion and Howitt, 1998). The companion–innovation hypothesis of Garrison argues that transport is not only

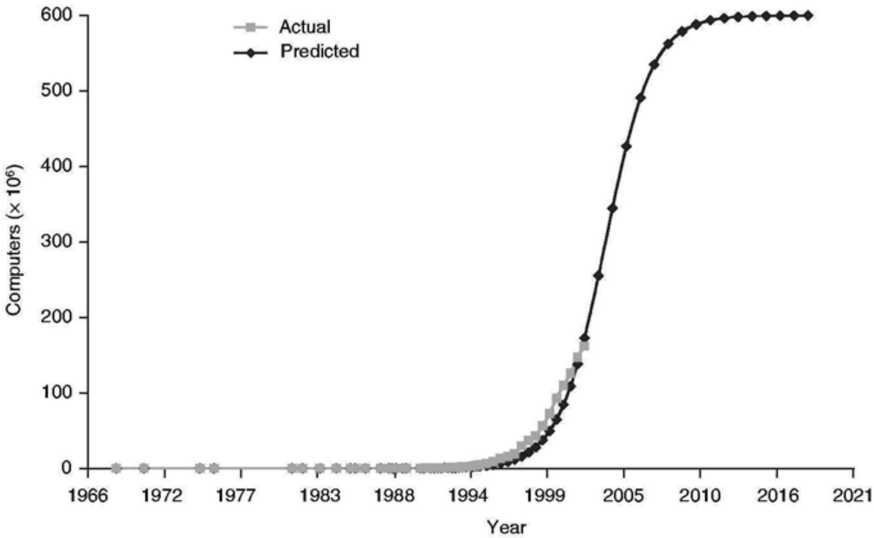


Figure 2. Internet host computers.

unlikely to be an exception but may be the most important transformation agent in the economy: through revenue sources such as the gas tax, transport investment drives the growth that funds it.

The life cycle model can be represented by the following equation:

$$\frac{f}{1-f} = e^{at+b}, \tag{1}$$

where f is the fractional share of technology (technology's share of the final market share), t is time, and a and b are model parameters. Such a tool may help in forecasting, because if the final size of the market can be assessed, and some deployment has already taken place, the pace of future growth can be understood broadly. On the other hand, this cannot tell what the microscopic decisions are that will situate a network in space, or will indicate small upturns and downturns. Further, for many technologies, the final size of the market is unclear until after the fact. For instance, how large will the Internet be? Figure 2 suggests an alternative (the best fit on the available data using equation (1) suggests a maximum of some 600 million Internet host computers).

Macroscopic models allow us to identify a general process for describing how technologies are deployed. But they do not really help us understand the underlying individual decisions on deployment, except to the extent that networks are deployed faster (in percentage terms) the younger they are.

3. Microscopic models

Few researchers have considered the process of transport network growth at the microscopic level, highlighting the importance of this research. Taaffe et al. (1963) study the economic, political and social forces behind infrastructure expansion in underdeveloped countries, finding that initial roads are developed to connect regions of economic activity and lateral roads are built around these initial roads. A positive feedback between infrastructure supply and population was also observed. Barker and Robbins (1975) investigated the London Underground's growth, but did not develop a theoretical framework.

The network evolution question at the microscopic level divides itself into several related problems. The first considers the location of network nodes. The second considers the connection of nodes with links. The third considers the sizing of links and the hierarchy of roads.

3.1. The node location problem

The location of network nodes reminds us of the geographer's central place question (Christaller, 1966). Christaller's central place theory arose in response to the question of how urban settlements are spaced, or, more specifically, what rules determine the size, number, and distribution of towns. The question of network evolution is in many respects similar, but we may think of it as the question of what rules determine the size, number, and distribution of links (or nodes). Christaller's model made a number of idealizing assumptions, especially regarding the ubiquity of transport services, in essence, assuming the network problem away. His world was a largely undifferentiated plain (purchasing power was spread equally in all directions), with central places (market towns) that served local needs. The plain was demarcated with a series of hexagons (which approximated circles without gaps or overlaps), the center of which would be a central place. However, some central places were more important than others because those central places had more activities. Some activities (goods and services) would be located nearer consumers, and have small market areas (e.g. a convenience store); others would have larger market areas to achieve economies of scale (such as warehouses).

The preceding paragraph does not do justice to Christaller, but his research has been extended by geographers and regional scientists (Lösch, 1938; Heilbrun, 1987). Models developed by Batty and Longley (1985), Krugman (1996), and Waddell (2001) all use newer modeling ideas, treating the problem in a decentralized fashion, and consider land use dynamics, allowing central places to emerge. However, those models, too, take the network as given.

In a more empirical sense, observation suggests that nodes emerge for a variety of reasons. Nodes occur at points of resource extraction (e.g. a mining town). Nodes occur at points of energy extraction (a waterfall) where a natural energy source can be exploited. Nodes also occur at points of trans-shipment, where nature's links (rivers and oceans) can be exploited. Nodes may also be located for military advantage (to protect an area against incursion by other forces). These nodes may then be connected with each other by links. Links may then cross, creating new nodes with high levels of accessibility. Table 1 shows the 20 largest US cities and their principal natural feature that was exploited.

If this argument is believed, it was in a sense inevitable that there would be nodes at these places. However, what may not have been inevitable was the size of the node, that is, there are a number of places which may have had equivalent natural bounty but never became a top 20 city. However, investigation of the history of these cities suggests that transport was the dominant reason for their existence, be it the movement of goods, energy, or water for irrigation. In most cases, waterways acted as conduits of travel, in several (e.g. Detroit and Dallas), the waterway was largely a barrier that was narrowest at these points compared with others, and in Phoenix, it was the water that was the resource. That said, it further argues that it is geographical asymmetries that drive the growth of nodes (a node has some advantage over all places that were not selected as nodes).

3.2. The link formation problem

The construction of new links can be modeled in several ways. In the first set of considerations, we assume we know the location of nodes. We could assume that all (or a very large number of) nodes are connected, but at some very low speed, and then use a network expansion model (such as discussed below) to reinforce selected links and allow others to wither, much as neurons develop in the brain of an infant (Edelman, 1987) or a neural network model learns. In contrast to this Darwinian selection process, we could assume that, for every node, there is a set of possible nodes it can connect with (neighbors within a certain radius that it does not already connect with directly). Constraints about crossing existing links could be established. We could incorporate forecasts or expected demand were the link to be built. Then a choice model, based on accessibility added or expected volume served based on the traffic of existing links using that node, would select new links to be added. In this vein, Garrison and Marble (1965) observed connections to the nearest large neighbor that explained the sequence of rail network growth in Ireland.

Alternatively, we could specify a process for simultaneous link formation and generation of new nodes. It is reasonable that nodes are generated by one of three processes: the location of natural features (e.g. harbors and waterfalls), the

Table 1
Important US nodes: largest metropolitan areas

City	Feature in dominant city
New York–Northern New Jersey–Long Island, NY–NJ–CT–PA	Harbor
Los Angeles–Riverside–Orange County, CA	Harbor
Chicago–Gary–Kenosha, IL–IN–WI	Harbor, river/canal connections to Mississippi
Washington–Baltimore, DC–MD–VA–WV	Harbor (Baltimore), capital (Washington)
San Francisco–Oakland–San Jose, CA	Harbor
Philadelphia–Wilmington–Atlantic City, PA–NJ–DE–MD	Harbor
Boston–Worcester–Lawrence, MA–NH–ME–CT	Harbor
Detroit–Ann Arbor–Flint, MI	Strategic crossing
Dallas–Fort Worth, TX	Trading post/crossing of Trinity River
Houston–Galveston–Brazoria, TX	Harbor
Atlanta, GA	Rail terminus
Miami–Fort Lauderdale, FL	Rail terminus, resort
Seattle–Tacoma–Bremerton, WA	Harbor
Cleveland–Akron, OH	River/canal terminus, Great Lakes port
Minneapolis–St Paul, MN–WI	St Anthony Falls on Mississippi River, most northerly navigable location
Phoenix–Mesa, AZ	Site of an ancient Native American irrigation system, Salt River
San Diego, CA	Harbor
St Louis, MO–IL	Confluence of Missouri and Mississippi rivers
Pittsburgh, PA	Confluence of Allegheny and Monongahela rivers with Ohio river
Denver–Boulder–Greeley, CO	Gold discovery at the confluence of Cherry Creek and the South Platte River (resource extraction)

location of artificial features (e.g. the intersection of two roads connecting different places), or explicit design (the nodes shall be in a grid spaced every 2 km). The first two are most interesting for exploring network evolution. A more general process then can be formulated: first, each step adds a node (networks are assembled one node at a time); second, attach each node to two other nodes with two links. The rules for attachment then become critical. As Barabasi et al. (1999) note, if nodes were more likely to connect with already well-connected nodes, we would have a scale-free network (resembling airline hubbing). But if nodes connected randomly to neighbors, we would have a random network that more resembles highways. Scale-free networks follow a power rule in the distribution of node connectedness. However while the connection structure of highways is limited to nodes being connected to usually at most four others, the links that connect them have attributes that differentiate them. Just as hubs are hierarchically organized in a node-based system, some links are more important in the hierarchy than others: they are faster, wider, and carry more traffic. Yamins et al. (2003) develop a simulation that grows urban roads using simple connectivity rules proportional to the activity at locations.

To some extent both the Darwinian and the choice process reflect aspects of the growth of real networks. The Darwinian process is probably best suited to undeveloped areas being opened, with access provided and cost of construction being the main offsetting factors. In particular we can think of the Darwinian process being appropriate in areas without developed transport planning, where animal trails are adopted by humans on foot, and later roads are built over those trails. The choice process better reflects the more "sophisticated" process of building a new link in already developed areas (or areas adjacent to developed areas). Even rural areas that already have a road network can be considered developed from this perspective.

3.3. The link expansion (contraction) problem

When a transport facility is built or expanded, travel increases on that facility both due to re-routing and re-scheduling and due to what is often called induced or latent demand, a finding confirmed at both the macroscopic level (e.g. counties) (Noland, 1999; Strathman et al., 2000; Fulton et al., 2000) and at the microscopic level (individual links) (Parthasarathi et al., 2003). As travel costs for commuters are lowered, the number of trips and their length increase. In market sectors of the economy, as population grows and preferences shift, leading to higher demand, suppliers produce more of a good. While surface transport decisions are often made in the political arena rather than the market, politicians and officials also respond to their customers – the voter and taxpayer. Although transport supply is relatively inelastic over the short run, it varies in the long run. However, it is not known to what extent changes in travel demand, population, income, and demography drive these long-run changes in supply. Answering this induced supply question in transport is a critical step in understanding the long-term evolution of transport networks.

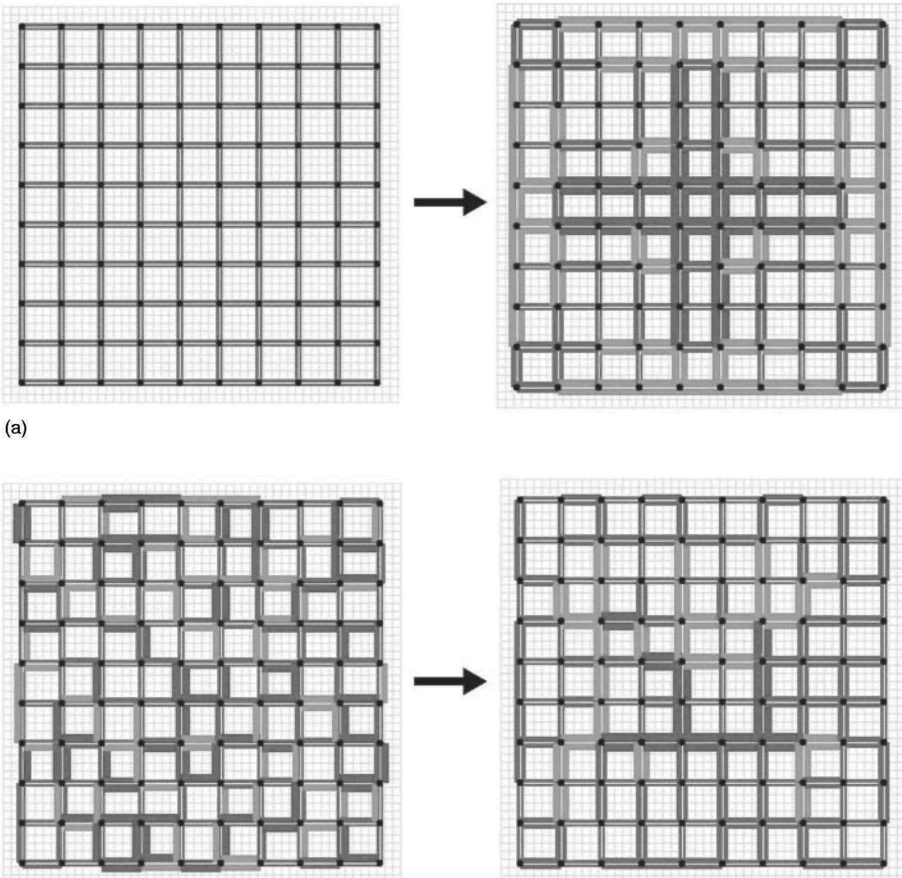
Observation suggests the hypothesis that decisions to expand transport networks are largely myopic in both time and space, usually ignoring non-immediate and non-local effects. This myopic decision process, when applied sequentially, tends to improve the relative speeds and capacities of links that are already the most widely used, and thereby expand their use. The rate and extent of this process is constrained by the cost of those improvements and limited budgets. The full ramification of network expansion on future infrastructure decisions is seldom considered. Improving one link will cause complementary (upstream and downstream) links to have greater demand, and competitors (parallel links) to have lesser demand (and be less likely to be improved). These network effects both complicate the problem and may suggest a structure for analysis.

In particular, the phenomenon of network hierarchy is an important issue. For instance, roads are classified in a way that designates most roads as relatively low-speed, low-volume links. Only a few links on the hierarchy of roads carry the bulk of traffic. Although planners and engineers design for the hierarchy of roads, those designs are constrained by previous decisions. In many respects, the hierarchy of roads is the network analog of geography's central place theory, which seeks to explain how hierarchies of places develop.

Yerra and Levinson (2004) simulate the link expansion problem, showing how a network can differentiate into a hierarchical network from either a random or a uniform network. The network, like observed networks, exhibits a power rule type of behavior – a few very fast links, some moderate speed links, and many slower links. Yerra and Levinson observe that the hierarchical structure of a network emerges as a function of induced demand (travelers take advantage of additional capacity by making longer trips as well as re-routing), cost functions with certain economies of scale, revenue proportionate to demand, an investment rule that embeds a "rich get richer" (but not "winner take all") logic, reflecting that important links get reinforced, and an underlying network structure (grid, radial, etc.). This model contained no comprehensive master plan applied to a *tabula rasa*. In brief, the hierarchy of roads would exist even if no planners or engineers intended it.

We can think of each link as an agent that chooses its speed based on preferences and constraints. There are several exogenous inputs: the base network, the distribution of land uses and demographics, and user-specified events. There is also a travel demand model that translates land use data into traffic flows and speeds on network links. Those traffic flows and speeds inform the network investment model. Those flows also determine the revenue and costs of maintaining and improving the link. When each link has exhausted its resources, the time period is incremented, population grows, land uses are updated, the travel demand is recomputed on the new network, and the process repeats.

Figure 3 shows the evolution of link speeds on two stylized grid networks using the above formulation. The thickness and shading of the line indicates the speed



(a)

(b)

Figure 3. Link expansion: the evolution of an abstract grid network (Yerra and Levinson, 2002). (a) Case 1: 10 × 10 grid, uniform initial speed = 1, uniform land use. (b) Case 2: 10 × 10 grid, random initial speed (distributed between 1 and 5), uniform land use.

of the link, divided into four categories with the widest bars being fastest. As can be seen, the network with an initial uniform distribution of link speeds develops into a hierarchical network with some faster links attracting more traffic and slower links serving local land uses and less traffic. Clearly the degree to which the network differentiates depends on particular model parameters. The 10 × 10 grid with an even number of streets in the north–south and the east–west directions maintains symmetry with two major north–south and east–west streets. (A 15 × 15 grid, with an odd number of streets in the north–south and east–west directions, and thus an equal number of streets to both the left and right of the center street, has only one.) Though the network is symmetric, it has boundaries, and it is those boundary conditions (the links at the edge serve traffic differently to links in the interior, as traffic at the edge tends to go inward) that allow differentiation to occur even with a uniform land use and equally spaced links of identical initial speed. The economies of scale (additional flow increases costs less than linearly) along with increasing costs for higher speed links drive the system to the equilibrium where link costs equal link revenue. The second case, showing a random initial distribution of link speeds, illustrates a more complex hierarchical emergence, one that is not symmetric. It is one where initial conditions strongly determine final results. Research shows that other types of initial differences (for instance a higher density downtown, natural topographical features, or random events or random distribution of other initial conditions) further differentiate the system.

Karamalaputi and Levinson (2003) empirically estimate choice models of highway network expansion and new construction. The network expansion model aims to predict which links will be expanded based on empirical factors (anticipated cost, length, traffic flow on the link, a parallel link, and upstream and downstream links, year, budget constraint, type of road, and population of neighboring areas). Those variables were generally important in prediction. Levinson and Karamalaputi's (2003) empirical prediction of new construction is by its very nature more difficult, since the choice set is not limited to existing links but rather to all potential links. Determining the set of plausible highway links requires developing rules (maximum length of link, connections to existing nodes, etc.) that make prediction feasible. This is a harder problem and will require significant further study.

4. Conclusions

The physical network as an object of study is a relatively new endeavor that complements studies of land use, traffic flows, and social networks (Watts, 1999). The development of network evolution models at the macroscopic and microscopic perspectives offers new insights into processes that previously were thought to result from the visible hand of planners, engineers, and politicians. While there is of course a residual to the analysis that can be explained as the product of conscientious decision-makers, there is a large part of network growth that is driven by the underlying geography, economy, and technology. Both the estimation of individual component models and their integration into a simulation of network growth (and decline) will increase our heretofore limited understanding of network evolution processes. This new understanding will have broad impacts on transport planning practice, and ultimately on the shape of cities and regions. In particular, it will provide a tool to illustrate the implications of current decisions on the future shape of the network, a consideration that is lacking in most planning and engineering studies. By providing a glimpse into a suggested future network configuration, cognizant policy can aim to redirect investment to produce an alternative set of preferred future investments.

Incorporating explicit measures of network externalities in decision-making will lead to better plans, network routing decisions, and implementation strategies. Understanding and illustrating how decisions at one point of time affect future choices should help guide planners and decision-makers desiring to shape the future. The long-term consequences of incremental changes will be assessed. This will help decision-makers assess the effects of changed policies, expanding existing facilities or routes, or building in new rights-of-way or offering new services. This improved understanding of long-term network dynamics would lead to better planning and design of road networks to exploit network externalities and maximize future choice for decision makers.

It is worth speculating about the implications of network evolution on the future of transportation technology. Transportation and telecommunications serve as both complements and substitutes. The capacity of telecommunications networks to provide "virtual" presence in contrast to the "physical" presence provided by the transportation network may affect the long-term demand for, and thus provision of, new transportation networks. Figure 2 illustrated the potential growth of the Internet, which is but one of many emerging communications networks. The increasing use of communications networks will undoubtedly change individual daily activity patterns as it has already changed business. Whether we will see a pattern of the expansion of communications resulting in a diminished investment in transportation (much as steamboats replaced sail or the airplane replaced intercity rail), or an expansion (the telegraph enabling the long-distance railroad) remains to be seen, but merits continued monitoring.

One of the key applications of communications technology in the transportation domain has been in the intelligent transportation systems arena, which has absorbed a great deal of research effort over the past two decades. While many of the investments are responses to maturity (traveler information to slightly improve the quality of a trip, or ramp meters to slightly improve capacity), the long-term goal is in the direction of vehicles that can drive themselves. This may be seen as an effort to give birth to a new mode rather than an incremental response to an existing mode.

References

Abdulaal, M. and L. LeBlanc (1979) "Continuous equilibrium network design models," *Transportation Research B*, 13:19–32.
Aghion, P. and P. Howitt (1998) *Endogenous growth theory*. Cambridge: MIT Press.

Aschauer, D. (1989) "Is public expenditure productive?" *Journal of Monetary Economics*, 23:177–200.
Barabasi, A., R. Albert and H. Jeong (1999) "Scale-free characteristics of random networks: the topology of the World Wide Web," *Physica A*, 272:173–187.
Barker, T.C. and M. Robbins (1975) *A history of London transport*, Vols 1 and 2. London: Allen and Unwin.
Batty, M. and P. Longley (1985) *The fractal simulation of urban structure. Papers in planning research 92*. Cardiff: Department of Town Planning, University of Wales Institute of Science and Technology.
Button, K. (1998) "Infrastructure investment, endogenous growth and economic convergence," *Annals of Regional Science*, 32:145–162.
Carruthers, J.L. and G.F. Ulfarsson (2001) "Public science expeditions: the influence of density and other characteristics of urban development," in: *The Pacific Regional Science Conference Organization*. Portland.
Christaller, W. (1966) *Central places in southern Germany* (trans. C.W. Baskin). Englewood Cliffs: Prentice Hall.
Edelman, G. (1987) *Neural Darwinism*. New York: Basic Books.
Friesz, T.L., S. Shah and D. Bernstein (1998) "Disequilibrium network design: a new paradigm for transportation planning and control," in: T.J. Kim, L. Lundqvist and L.-G. Mattson, eds, *Network infrastructure and the urban environment*. Berlin: Springer-Verlag.
Fulton, L.M., R.B., Noland, D.J. Meszler and J.V. Thomas (2000) "A statistical analysis of the induced travel effects in the U.S. mid-Atlantic region," *Journal of Transportation and Statistics*, 3:1–14.
Garrison, W.L. (1996) *CE250 course notes*. Berkeley: University of California.
Garrison, W.L. (2000) "Innovation and transportation's technologies," *Journal of Advanced Transportation*, 32:31–63.
Garrison, W.L. and D.F. Marble (1965) *A prolegomenon to the forecasting of transportation development. Technical report*. Washington, DC: Office of Technical Services, US Department of Commerce, US Army Aviation Material Laboratories.
Garrison, W.L. and R.R. Souleyrette (1996) "Transportation, innovation, and development: the companion innovation hypothesis," *Logistics and Transportation Review*, 32:5–37.
Gramlich E.M. (1994) "Infrastructure investment: a review essay," *Journal of Economic Literature*, 32:1176–1196.
Grübler, A. (1990) *The rise and fall of infrastructures: dynamics of evolution and technological change in transport*. Heidelberg: Physica-Verlag.
Heilbrun, J. (1987) *Urban economics and public policy*, 3rd edn. New York: St Martin's Press.
Holland, J.H. (1975) *Adaptation in natural and artificial system*. Ann Arbor: University of Michigan Press.
Huang, H.J. and M. Bell (1999) "Continuous equilibrium network design problem with elastic demand: derivative free solution methods," in: M.G.H. Bell, ed., *Transportation networks: recent methodological advances. Selected proceedings of the 4th Euro Transportation Meeting*. Amsterdam: Elsevier.
Ishaq Nadiri, M. and T. Mamuneas (1996) *Contribution of highway capital to industry and national productivity growth*. Washington, DC: US Federal Highway Administration.
Karamalaputi, R. and D. Levinson (2003) "Induced supply: a microscopic analysis of network growth," *Journal of Transport Economics and Policy*, 37:297–318.
Krugman, P.R. (1996) *The self-organizing economy*. Cambridge: Blackwell.
Levinson, D. and R. Karamalaputi (2003) "Predicting the construction of new highway links," *Journal of Transportation and Statistics*, 5:91–89.
Lösch, A. (1938) "The nature of economic regions," *Southern Economic Journal*, 5:71–78.
Mensch, G. (1979) *Stalemate in technology*. Cambridge: Ballinger.
Minsky, M. (1986) *The society of mind*. New York: Simon and Schuster.
Miyao, T. (1981) *Dynamic analysis of the urban economy*. New York: Academic Press.
New Jersey Office of State Planning (1996) *Projecting municipal road costs under various growth scenarios*. Document 109. Trenton: New Jersey Office of State Planning.
Noland, R.B. (1999) "Relationships between highway capacity and induced vehicle travel," in: *Transportation Research Board 78th Annual Meeting Preprint CD-ROM*. Washington, DC: Transportation Research Board, National Research Council.

Parthasarathi, P., D. Levinson and R. Karamalaputi (2003) "Induced demand: a microscopic perspective," *Urban Studies*, 40:1335–1351.

Peat, F.D. (2002) *From certainty to uncertainty: the story of science and ideas in the twentieth century*. Washington, DC: National Academy Press.

Strathman, J.G., K.J. Dueker, T. Sanchez, J. Zhang and A.E. Riis (2000) *Analysis of induced travel in the 1995 NPTS*. Portland: Center for Urban Studies, Portland State University.

Taaffe, E.J., R.L. Morrill and P.R. Gould (1963) "Transport expansion in underdeveloped countries," *Geographical Review*, 53:503–529.

Waddell, P. (2001) "UrbanSim: modeling urban development for land use, transportation, and environmental planning," in: *17th Pacific Conference of the Regional Science Association*. Portland.

Watts, D. (1999) *Small worlds*. Princeton: Princeton University Press.

Yamins, D., S. Rasmussen and D. Fogel (2003) "Growing urban roads," *Networks and Spatial Economics*, 3:69–85.

Yerra, B. and D. Levinson (2004) "The emergence of hierarchy in transportation networks: how networks grow," in: *Western Regional Science Association Conference*. Porta Rico

Chapter 12

TRANSPORT AND REGIONAL GROWTH

CHRIS JENSEN-BUTLER
University of St Andrews

BJARNE MADSEN
Institute of Local Government Studies (AKF), Copenhagen

1. Introduction

It is well established that there is a strong correlation between economic growth and demand for transport, and that this is the case both for freight and passenger transport (Baum and Korte, 2002). Passenger traffic (passenger-km) has an income elasticity of demand a little more than unity, and freight traffic (tonne-km) an income elasticity of demand of about unity. What is perhaps less clear is the direction of causality in this relationship. On the one hand, it is possible to view transport essentially as demand that is derived from economic activity, or as Vickerman (2002) points out, economic growth requires trade, and trade requires transport. Economic growth implies greater division of labor and spatial specialization, the development of new technologies requiring transport such as just-in-time, growth in commuting and business travel, and growth in the quantity of goods and services (essentially people) to be transported. At the same time, rising household incomes generate increased demand for travel, both leisure activities and for shopping.

The demand for transport, however, cannot be treated simply as derived demand. Increasing mobility is a precondition for increased productivity, and growth and improvements in transport may in themselves promote growth. The UK Standing Advisory Committee on Trunk Road Assessment (1999) identifies six mechanisms by which this can occur:

- reorganization and rationalization of production, distribution, and land use;
- reducing labor costs by expanding catchment areas;
- increases in output through lower production costs;
- stimulation of inward investment;
- unlocking inaccessible sites for development;
- triggering cumulative growth.

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

This seems to be even more the case in developing countries, where the World Bank (1994) claims a 21% rate of return on its transport infrastructure investment projects in the 1980s and as much as 29% in the case of highway projects. Also in the advanced economies, investment in transport and transport infrastructure has become a major policy instrument designed to promote economic development, typically at the regional level. For example, the EU's Regional Development Fund invests heavily in transport infrastructure projects as a regional policy instrument.

As transport policy becomes more integrated with policy designed to promote economic growth, questions are being raised about the nature of this relationship: Do transport system improvements always have a positive economic benefit? What is the spatial distribution of such benefits? What can be done to maximize the benefits of investment in transport? How should such investments be financed? What are the appropriate methodologies that can be used to assess the benefits of investments in transport or changes in transport pricing? The last of these questions is the main focus here, and has been the focus of a number of studies (Rietveld and Nijkamp, 2000).

It is clear, however, that there is a two-way interaction between growth in economic activity and growth in transport activity, and any modeling approach designed to assess benefits of policy-based investments in transport or other changes in the transport system should encompass this two-way relationship. The importance of being able to model and evaluate the nature of the relationship has grown, as other features of increased traffic flows have begun to affect profoundly the societies and economies in which we live. The negative externalities arising from transport activity – environmental pollution, congestion, accidents, and noise – now constitute serious problems that any transport improvement project must address. The usual methodology for addressing these issues is cost–benefit analysis (European Conference of Ministers of Transport, 2001), the efficacy of which is increasingly being questioned (Standing Advisory Committee on Trunk Road Assessment, 1999). Transport policy can be directed either at reducing transport costs, as is the case with improved infrastructure, which promotes interaction and traffic, or at reducing traffic flows, using, for example, road pricing or fuel taxation.

1.1. Definitions

Transport moves goods or persons between physical locations. The changes in the transport system considered here are those that alter the costs of transport. They include changes in transport infrastructure provision, changes related to changing transport technology, for example modal change, and changes related directly to transport pricing, for example road pricing or fuel taxation. A variety of indicators has been used to assess the economic effects of changes in the transport system, including changes in employment, in income or income per capita, in productivity, and in new firm formation.

Externalities affect the value of transport system improvements. An externality is an unpriced effect on the utility or production function of a third party, caused by a transaction between the direct users. Externalities can be positive (reducing costs) or negative (increasing costs), as in the case of congestion or pollution. A distinction is made between technological or locative externalities (Bannister and Berechman, 2000), for which no compensatory payment is made (such as traffic congestion or knowledge spillovers), and pecuniary externalities, which refer to the way in which changes in transport costs alter prices in other markets. Pecuniary externalities are transmitted through markets, for example changing patterns of land rents as accessibility changes, and represent transfers between agents, and should therefore not be counted as benefits. However, they can have differential regional effects, which merits their consideration. Network externalities are benefits conferred upon network users other than those immediately affected by the change.

1.2. Problems

Analysis of the relationship between transport and economic growth is, partly because causal mechanisms are difficult to specify and the economic and spatial systems involved are themselves inherently complex, involving many feedback mechanisms. These difficulties arise for a number of reasons. First, transport plays very different roles in an economy. Passenger trips can be classified broadly into three journeys to work, shopping, and tourism/recreation, while freight transport can be classified as trade flows. Second, these different types of traffic operate at very different spatial levels. Commuting and shopping are mainly sub-regional or local, tourism can be at sub-regional, regional, national and international levels, and freight transport operates at all levels. Any change in the transport system (e.g. a new motorway) typically affects all of these spatial levels, and any modeling approach must be versatile enough to encompass all of these levels, which suggests that sub-regional economic modeling really is the appropriate modeling base. Third, there are different types of actor (households, firms, government) demanding transport in different types of location. Fourth, in conceptual terms, in an economy there are places of production, places of residence, and places of market demand, all of which are both defined and linked by traffic flows. Fifth, changes in the transport system affect different markets (e.g. product, factor markets, housing markets, commercial property markets) in different ways, and the different markets take different time periods to adjust to a new equilibrium.

Sixth, the effects of any change in the transport system on the economy depend upon the assumptions made about both the transport system and the economy. In particular, as will be shown below, the assumption of perfect competition in product and factor markets for transport-using sectors can give different results as compared with imperfect competition. This also applies to the transport sector itself: in the face of market failure, as for example when externalities are present, the competitive equilibrium will not be Pareto optimal. Seventh, there are substantial problems concerning data availability, partly because of the scarcity of economic and traffic flow data at regional and sub-regional levels, and partly because for transport-using sectors some of the transport demand is met by supply from the sector itself (own vehicles) and part is purchased from the transport sector. In the absence of satellite accounts for transport in national or regional accounts, this creates substantial data problems for modelers. Finally, there are a number of different general approaches to the analysis of the relationship between transport and economic growth, providing varied results. These different approaches are discussed below.

1.3. Spatial issues

Much of the work on the economic impacts of changes in the transport system have dealt with one spatial unit, such as a country or a large region. The impacts upon different regions and even sub-regions, however, is becoming a question of increasing importance, for modelers and policy-makers alike. There are the different regional economic impacts of transport system changes, arising from the way in which firms and consumers react to changes in transport costs in different locations, including organizational change. Second, there are the consequences for spatial competition, relocation, and agglomeration, which is the field of study of the "new economic geography" (Krugman, 1998; Brakman et al., 2001). Spatial outcomes in this approach can be regarded as the result of interactions between, on the one hand, scale economies and market size, and on the other transport costs. To this can be added agglomeration economies, of localization and urbanization (Fujita et al., 1999). Third, there are effects on the labor market. Transport system improvements will generally increase the geographical size of labor markets and therefore increase labor productivity (Prud'homme, 2002). Finally, transport has an effect on property and housing markets: it is well-known that transport has a profound effect on urban growth and prices of housing. Further, house prices affect location decisions of households, and hence commuting flows, as well as decisions to migrate. Road pricing in cities affects land and housing prices, as markets adjust.

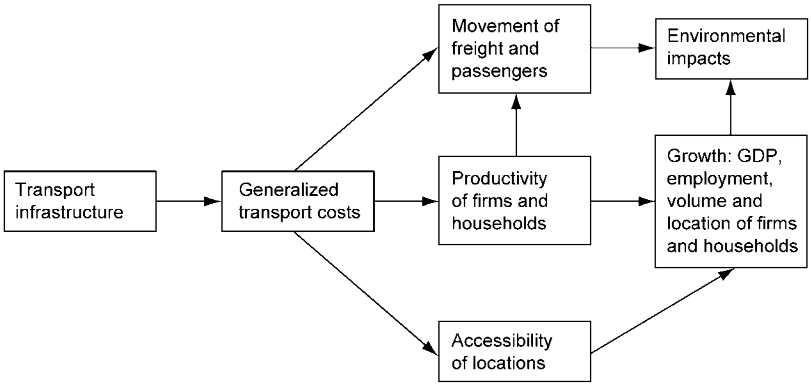


Figure 1. Transport infrastructure and its economic effects (Rietveld, 1996). GDP, gross domestic product.

2. Approaches to the analysis of interactions between transport and the economy

From a theoretical perspective, there are a number of different approaches to the analysis of the impact of transport system changes on economic activity that have been proposed (Nijkamp and Blass, 1994; Rietveld and Nijkamp, 2000).

2.1. Traditional approaches

Banister and Berechman (2000) identify the traditional approach as one where transport infrastructure investment has both direct and indirect effects. The direct effects, related to changes in accessibility have consequences for consumer and producer surplus; cost savings in production and transactions; and effects on location decisions. The indirect effects are externality effects (such as congestion and environmental effects) and multiplier effects associated with the construction phase of the project.

Rietveld (1996) proposes a framework for assessment of the regional growth effects of transport infrastructure investment, as shown in Figure 1. This scheme includes one type of externality, but like the previous approach is essentially unidirectional in its causal structure. Here, reductions in transport costs affect primarily firms' productivity and lower costs for households, which creates regional growth.

2.2. *Approaches based upon the analysis of externalities*

Banister and Berechman (2000) argue that the traditional approaches are inadequate because of the pervasive effects of externalities. They propose an explanatory framework, as shown in Figure 2. In addition to the primary benefits, which result in welfare gains through cost reductions and productivity increases, there are a number of effects that they term allocative (cost reducing, technological) externalities, affecting regional growth, but which do not operate through the price system. These are agglomeration economies, or inter-firm externalities arising from increased accessibility; reduction of labor market imperfections, network externalities, where an improvement to a part of the network raises transport activity in the network as a whole; and congestion and environmental externalities.

What is common to these approaches is that the effects are unidirectional, there being no explicit feedback. Furthermore, the differential spatial (regional and sub-regional) economic effects of transport system change are difficult to identify. These approaches lend themselves to analyses based upon reduced-form single-equation models, as discussed below.

2.3. *Social-accounting-matrix-based approaches*

An approach to developing a two-way interaction is through the use of inter-regional social-accounting-matrix (SAM)-based analyses (Round, 1988; Hewings and Madden, 1995), where the input–output approach has been of prime interest. In this approach, regional economies are modeled primarily using input–output techniques, and space and transport costs enter into the modeling framework, which can include a household sector, where a number of variables are introduced to determine the private consumption component of final demand. A distinction is made between the region of residence (and demand) and the region of production. Space and transport enter into the model in the determination of inter-regional trade flows to satisfy final and intermediate demand. The approach usually involves a high level of detail and segmentation.

The two main types of effect arising from a transport system change that can be estimated using this approach are distribution effects, or the way in which regional economic activity changes as trade flows change after a transport system change, and generation effects, or changes in trade flows caused by an overall increase or decrease in inter-regional trade. This type of approach has a number of advantages, including feedback effects and the explicit modeling of linked regional economies, but involves the usual problems associated with input–output approaches, including being demand driven, using fixed coefficients with no scale economies, and no supply-side constraints. In addition, data requirements are substantial. This type of approach is illustrated in Figure 3 (Jensen-Butler and Madsen, 1996).

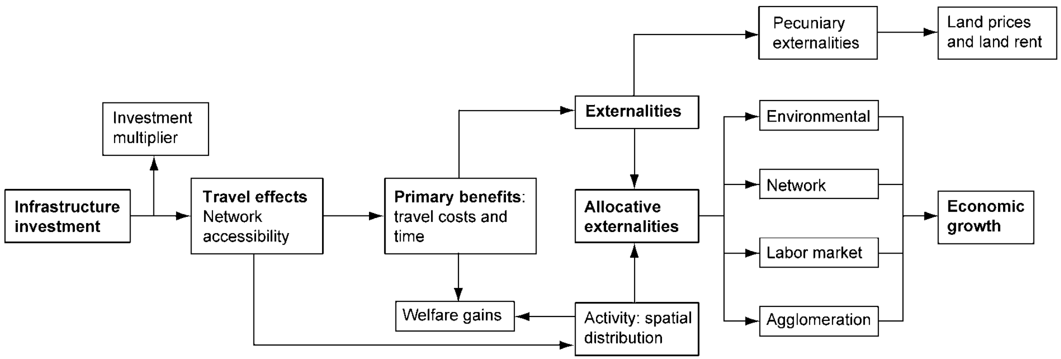


Figure 2. Transport infrastructure and its economic effects, including externalities (Banister and Berechman, 2000).

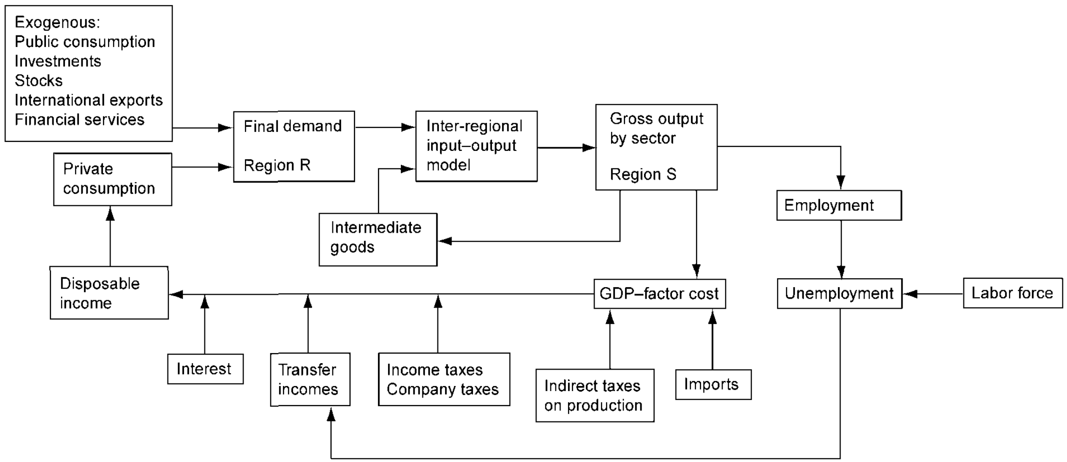


Figure 3. SAM/input–output approach.

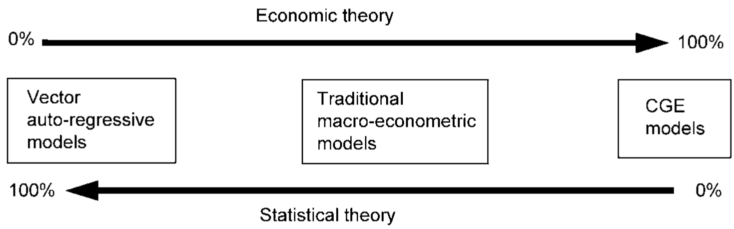


Figure 4. Economic models.

2.4. Computable general equilibrium models

The deficiencies of the input–output and other approaches (McGregor et al., 1998) have in recent years led to the construction of spatial computable general equilibrium (CGE) models.[a] They are usually of the comparative static variety, where household and firm behavior is modeled explicitly, using microeconomic theory, where the equilibrium of supply and demand in all markets is determined by price signals, and it is assumed that markets clear, and, typically, all flows in an economy are modeled using an SAM. In the non-spatial versions, in the basic form there are markets for intermediate and final commodities, and, for factors there are both the labor and capital markets. At equilibrium there is no excess demand or supply, and there is a set of unique non-negative prices that determine the equilibrium in all markets simultaneously. Consumer equilibrium is based upon utility maximization subject to income constraints, and producer equilibrium is based upon profit maximization subject to technological constraints. Any exogenous price change makes all markets adjust simultaneously. Petersen (1997) has described these models in relation to a continuum, as shown in Figure 4.

At the one extreme vector autoregressive (VAR) models attempt to explain a vector of variables as a function of their own lagged variables, their theory content being very limited, while CGE modeling starts with theory and then finds data that fit the model, their statistical content being limited. Traditional econometric models occupy the middle ground. The modeling process (Shoven and Whalley, 1992) is as shown in Figure 5.

A data set representing the economy in equilibrium is required initially. Then a set of parameter values are determined (calibration), often from a variety of sources, which can replicate the benchmark equilibrium. Change is then introduced into the model, which produces alternative (counterfactual) equilibria, which can

[a]The general principles involved in the construction of (non-spatial) CGE models are discussed in Shoven and Whalley (1992) and Kehoe and Kehoe (1994), and for regional models in Partridge and Rickman (1998).

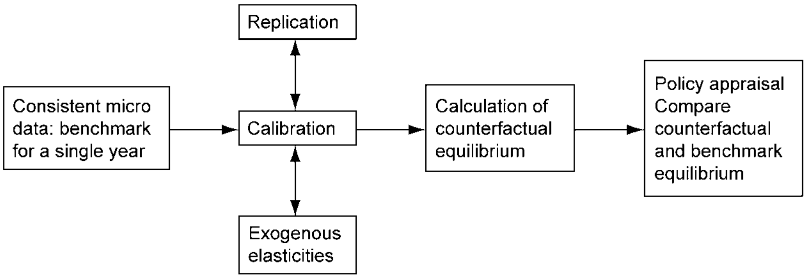


Figure 5. The CGE modeling process.

be compared to the benchmark. Input–output models are a specific type of CGE model, with no supply side constraints.

Spatial CGE models have been used to examine the relationship between changes in transport systems and economic growth. They usually incorporate inter-regional SAMs, and input–output components inside their framework (Haddad et al., 2002).[a]

There are three main avenues to expand the non-spatial CGE framework in the direction of spatial models. The first is by using a two-region model, which means that there is no need to model transport costs explicitly (McGregor et al., 1995). Second, there is a multiregional extension with fixed inter-regional transport costs, and here there are spatial equilibria, with no excess demand or supply in any region. This means that the delivered price in the importing region must equal the producer price plus transport costs. These types of model can incorporate multiple products, modeled separately. The third extension involves the explicit modeling of the transport sector, so that transportation costs between regions are no longer fixed, equilibrium in the market for transport is modeled explicitly, and pricing in this market is efficient. There are still regional equilibria for each product, and consumer and producer equilibrium conditions apply. In both of these cases, any change in the costs related to the transport system will disturb the initial equilibrium, and the model undertakes the necessary adjustments to restore equilibrium, providing a variety of information on spatial impacts, including changes in income, employment and welfare measured for example by equivalent variations. In addition, CGE models permit evaluation of welfare changes and their distribution. Modeling national and regional economies using CGE approaches also involves the use of inter-regional social accounting matrices. The general characteristics of these types of model are that:

[a]Only a brief discussion of spatial CGE models can be presented here; for a more extensive review, see Van den Bergh et al. (1996), and for an introduction see Isard and Azis (1998).

- they are usually based on the classical Walrasian market clearing assumptions (Arrow and Hahn, 1971);
- they are solved numerically rather than analytically;
- many of the relationships in the model are inherently non-linear;
- they are general, rather than partial (all markets are modeled);
- the parameters of the model are not usually estimated econometrically but are loaned from other studies or are based upon economic theory;
- they are applied models in the sense that much early general equilibrium modeling was directed at proving the existence of an equilibrium, whereas in these types of model the equilibrium itself is determined;
- they permit analysis of changes in welfare, between institutional groups, and, in the spatial versions, across space.

CGE models can model explicitly imperfect competition, including downward-sloping market demand curves facing firms (which means that price does not equal marginal cost), the existence of scale economies in production, and the treatment of externalities generated by the transport sector. The Dixit and Stiglitz (1977) monopolistic competition framework is commonly used in these types of model, to allow for variety as an element in utility functions and to incorporate imperfect competition among producers. The use of non-linear relationships permits modeling of substitution between inputs by firms and between consumption goods for households.

The real strength of CGE models is their solid foundation in microeconomic theory. These types of model, however, also have substantial downsides. They are inherently very complex, and often only operate at high levels of spatial and sectoral aggregation (few regions and few sectors) as well as one type of production factor and one type of institution (household). Thus, in empirical terms, they are very abstract, and distant from the real world, providing low levels of detail. There are a number of theoretical and mathematical complexities. The general equilibrium solution for economies with more than one industry subject to monopolistic competition is impossible to obtain for models with fixed population distributions, as there are multiple equilibria. The number of possible equilibria increases with the number of regions and industries. The different types of interaction can only be combined inside one model by increasing its complexity substantially. Frequently, the transport sector is not modeled explicitly, rather the Samuelson (1954) iceberg principle is used, where transport costs are treated as a loss in quantity of delivered goods, which raises questions of theoretical adequacy. While resting on a strong theoretical base, the empirical foundation of CGE models is often weak. For example, the consistent estimation of key parameters, such as substitution elasticities at the regional level, remains a problem. The results obtained are usually treated as indicators of the likely direction of effects of a transport system change, rather than providing forecasts as such. Equally problematic is that in the real world it is not immediately obvious in which regions and sectors imperfect competition exists, so that, in a sense, perfect competition remains the benchmark.

3. Illustrating the central modeling issues

In this section the issues involved in modeling the transport–economy relationship are brought together illustrated using a Danish model, LINE (Madsen et al., 2001a,b; Madsen and Jensen-Butler, 2004). It incorporates general equilibrium modeling principles while at the same time being anchored firmly in empirical data and operating at low levels of aggregation and high levels of detail, with several actors and dimensions. The model attempts to capture the complexity of the relationships involved. Discussion of the model serves to illustrate the key issues that arise when modeling interactions between the transport system and the economy.

3.1. The LINE model

As noted, there is a two-way relationship between developments in the transport system and changes in economic activity. On the one hand, changes in economic activity affect the demand for transport, creating changes in traffic flows, and on the other hand, changes in the transport system typically affect prices and costs, which feed directly into regional economies, causing changes in economic activity. Underlying this fundamental two-way relationship is the basic role of space. Changes in economic activity create changes in demand for transport and new patterns of traffic flows in space. Changes in the transport sector transform spatial relationships, creating new patterns of economic activity and spatial interaction. Here the model structure embodied in LINE is used to examine the fundamental relationship between transport and economic activity. This structure is used to establish basic definitions and to identify central issues and relationships and to provide a framework into which different studies can be placed.

A number of key dimensions are to be found in the model. First, regional economic activity is classified according to its relationship with three fundamental types of locality: place of production, place of residence, and place of demand.[a] These distinctions are not usually clearly made in many modeling approaches.

[a]Ideally, a fourth type of locality could be included: location of the factor market. While unusual, it can in fact be found in reality, for example hiring casual labor at the docks.

Second, different types of actors (based upon SAM principles) are identified: activities (industries), factors of production, and institutions (households, firms, governmental organizations), as well as commodities. The use of commodities rather than gross output, as in the case of input–output based approaches, is necessary if the model is to relate to underlying microeconomic theory. This conceptual framework can be used to identify what any specific study attempts to do in terms of location, geographical scale, and type of transport activity, related to the first dimension, and the what are the effects of changes in transport on production and distribution using the second dimension. A third dimension involved in modeling of transport costs and commodity prices on the one hand, and regional economic activity on the other, is the definition of price concepts used. Fixed prices are used to represent physical flows, and current prices represent the prices used by economic agents as the basis for their economic decisions. A fourth dimension relates to externalities, dealing with both pecuniary and technological externality effects between producers, between consumers, and between producers and consumers, relating in turn to the nature of interaction between agents.

A simplified graphical presentation of a conceptual model of the relations between transport and the economy is found in Figures 6 and 7, whose structure reflects the four dimensional framework described above.[b] Figure 7 can be superimposed on Figure 6.

The economy–transport model is based on the idea that regional economic activity can be described using two interrelated circles: a real Keynesian circuit and a dual cost–price circle. Figure 6 shows the real circle, involving regional economic activity, and Figure 7 shows the way in which transport is related to this activity. There is a corresponding counter-clockwise cost–price circle, and a corresponding transport cost–price circle, which for reasons of space are not shown here. In Figures 6 and 7 the horizontal dimension is spatial: place of work (denoted R), place of residence (T), and place of demand (S). Production activity is related to place of work. Factor rewards and income to institutions are related to place of residence, and demand for commodities is assigned to place of demand. The vertical dimension is more detailed, and follows with its fourfold division the general structure of an SAM model. Production is related to activities; factor incomes are related to activities by sector (E); factors of production (G) (with labor classified by sex, age, and education); institutions (H) (households by type); and demand for commodities (V) (related both to final demand and intermediate consumption).

[b]The full and operational model LINE model is described in detail in Madsen et al. (2001a). The data used in the model, together with the inter-regional SAM, are described in Madsen and Jensen-Butler (2002) and Madsen et al. (2001b).

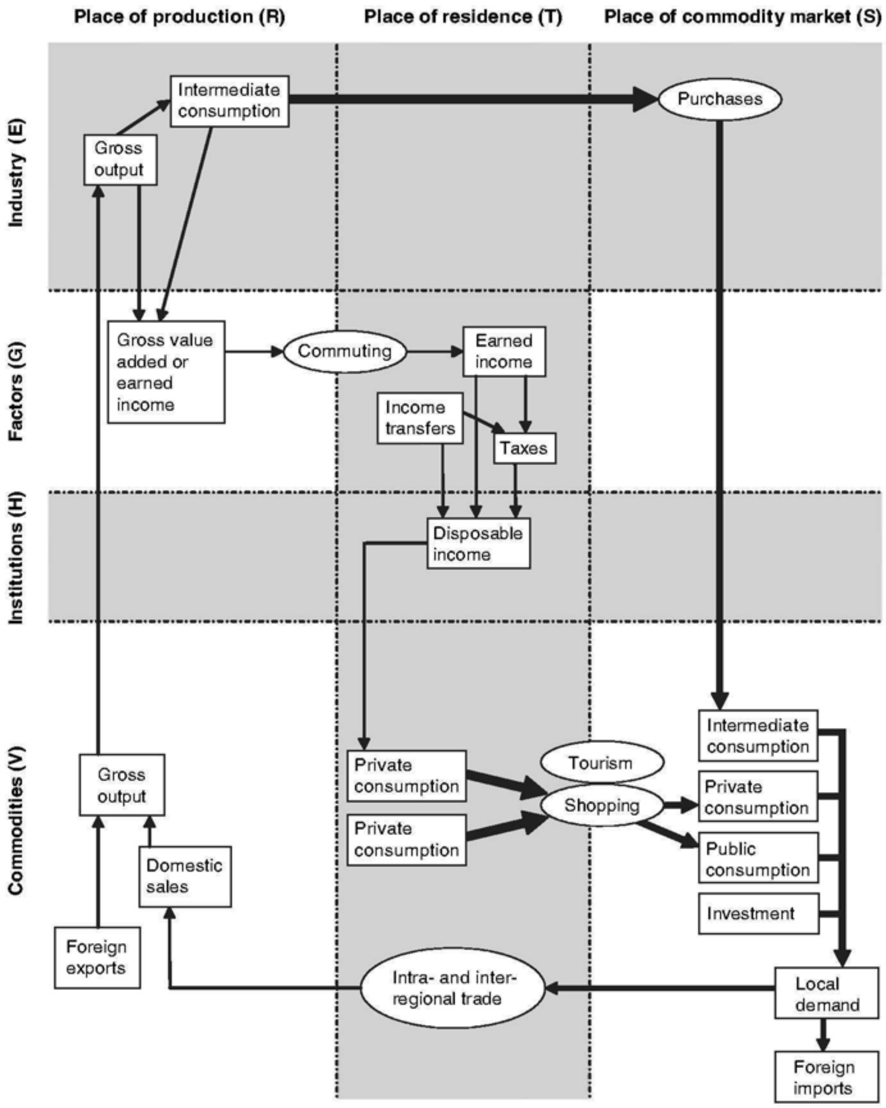


Figure 6. The real circuit in LINE.

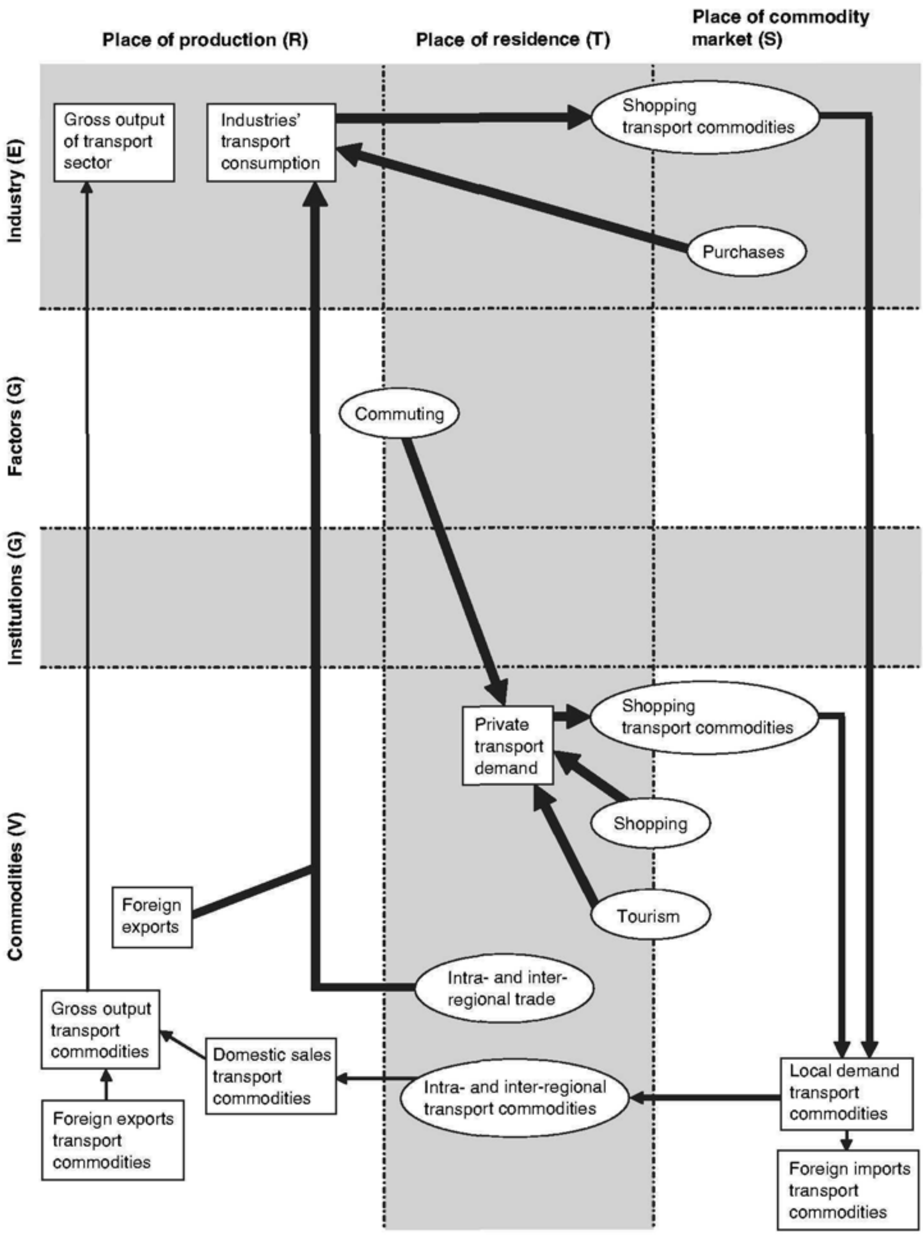


Figure 7. Demand for the transport commodity in LINE.

The equations of the model are represented by arrows showing the causal relationships between variables. The thickness of the arrows represents the different price concepts used in the model, which is of importance for modeling the role of transport in the regional economy. The thinnest arrows show basic prices, which are prices at place of production excluding value-added tax (VAT), commodity taxes, and transport costs. With increasing thickness of the arrows: basic prices plus are basic prices including transport costs between place of production and place of demand; market prices are basic prices plus adding VAT, commodity taxes, and trade margins; and market prices plus are market prices plus transport costs from place of demand prices to place of residence of the institution purchasing the commodity.

The model replicates a base year, and then is solved iteratively after a disturbance in the economic or transport system, converging on a new equilibrium.

Regional and sub-regional economic activity and interaction in space

The real circuit corresponds to a straightforward Keynesian model, and moves clockwise in Figure 6. Starting in cell RE in the upper left corner, production generates factor incomes in basic prices, including the part of income used to pay commuting costs. This factor income is transformed from sectors (RE) to sex, age, and educational groups (RG). Factor income is then transformed from the place of production (RG) to the place of residence (TG) through a commuting model after deduction of transport costs. Employment follows the same path from sectors (RE) to sex, age, and educational groups (RG), and further from the place of production (RG) to the place of residence (TG). Employment and unemployment are determined at the place of residence (TG). Employment refers both to the place of production and to the place of residence, whereas unemployment and the labor force, by definition, can only be related to the place of residence. This also explains the relationship of earned income to both the place of production and the place of residence, whereas other income, transfer incomes, and personal taxes, by definition, can only be related to the place of residence.

Disposable income is calculated in a sub-model, where, from earned income net of transport costs, taxes are deducted and transfer and other incomes are added. Disposable income is distributed from factors (TG) to households (TH). This is the basis for the determination of private consumption by type of household in market prices, by place of residence. This consumption is also transformed into commodities through a use matrix (TV). Private consumption is divided into tourism (domestic and international) and local private consumption. Private consumption is assigned to the place of demand (SV) using a shopping model for local private consumption and a travel model for domestic tourism. It can be noted that the value of local private consumption and tourists' consumption at the place of residence includes the cost of transport in connection with shopping and tourism, defined here as "market prices plus." In the same way, intermediate consumption (in market prices plus) is derived from the gross output at the place of production (RE), which in turn is transferred to the place of demand and is transformed into demand for commodities (TV). This also involves transport of commodities for intermediate consumption from the place of demand to the place of production.

Private consumption together with intermediate consumption, public consumption, and investments constitute the total local demand for commodities (SV) in market prices. In this transformation, transport costs have been deducted from market prices plus in order to determine local private consumption in market prices. Local demand is met by imports from other regions and abroad in addition to local production. Foreign imports (cost, insurance, freight) and inter-regional imports (basic prices plus) include transport costs. Through a trade model, exports to other regions and production for the region itself (in basic prices) are determined (from SV to RV). This involves transport of commodities to other regions and abroad. Adding exports abroad to domestic demand, gross output by commodity is determined (RV). Through a reverse make matrix the cycle returns to production by sector (from RV to RE). On the basis of gross output, sectoral intermediate consumption, gross value added and employment are determined.

Determination of commodity prices and transport costs. The interaction components of the real circle (commuting, shopping/tourism, trade) determine demand for the transport commodity, as illustrated in Figure 7. Demand for transport commodities (SD) is met in the same way as with other commodities, in the first step through shopping (the place where the vehicle is entered, the place of demand), and in the second step through intra-regional supply and inter-regional and international trade (the place of origin of the transport product, RE). Gross output of transport commodities at the place of production (RE) is in turn produced by the transport sector (mainly external transport) and other sectors (mainly sector-internal transport). On the basis of gross output in the transport sector, intermediate consumption, gross value added, and employment in the transport sector are determined. From here, the derived economic effects can be calculated in the same way as for other sectors and commodities.

The cost–price circuit: determination of commodity prices

Using the stylized version of the model shown in Figure 6, the counter-clockwise cost–price circle (not shown explicitly) corresponds to the dual problem. In the cost–price circle, production and demand are calculated in current prices, which in turn are transformed into relevant price indices. In cell RE, sector basic prices (current prices) are determined by costs (intermediate consumption, value-added and indirect taxes, net in relation to production). Through a make matrix, gross output by sector is transformed into gross output by commodity, both in basic prices (from RE to RV). These are then transformed from the place of production in basic prices to the place of demand in basic prices plus including transport costs (RV to SV) and further into market prices through inclusion of retailing and wholesaling costs and indirect taxes. Finally, private consumption is transformed from the place of demand in market prices to the place of residence in market prices plus (from SV to TV).

The cost–price circle and transport. The costs of producing transport commodities and the prices paid by consumers of transport commodities enter into the model. Starting in cell RE, the basic prices for production in the transport sector determine the basic prices for production of transport commodities (cell RV). Further, basic prices plus for transport commodities at the place of demand are determined. This is followed by the formation of market prices by the inclusion of trade margins and commodity taxes. These commodity taxes can include fuel taxes, road tolls, etc. (in addition to infrastructure improvements), which are determined by the spatial structure of the transport activities. Finally, the transport commodities are transformed from the place of demand in market prices to the place of residence/production in market prices plus, relating to interaction in the regional economy, such as inter-regional trade and shopping for intermediate goods and shopping by institutions (households, etc.).

The transport commodity in the economy

In economic models with a spatial dimension the transport commodities enter into the real circle and into the cost and price circle. In the real circle the demand for transport commodities is normally divided into intermediate consumption and final demand (private and governmental consumption and investments), as a share of intermediate consumption and final demand are used on transport commodities. Adding these together, the total demand for transport commodities is calculated. In demand-driven models, demand creates its own production through the determination of production from demand. In the real circle the multiplier process from demand to production and income and back to demand creates the direct, indirect, and induced effects of changes in the transport demand. In the cost–price circle the primary impact of the transport commodities is the margin function of transport commodities. As for the retailing and wholesaling commodities, the transport commodities are a necessary cost mark-up, which adds to the value of the commodities. Moving a commodity from its place of production to the market place of the commodity and further to the place of residence for the consumer and to the place of production of the producers adds to the value of the commodities transported.

Following on from the above, when analyzing the interaction between the regional economy and the transport system, the key economic object is the transport commodity, often largely ignored in CGE models. The transport commodity has two components. The first is the vehicle itself (walking, cycle, car, lorry, train, etc.) while the second is the infrastructure used by the vehicle. The transport commodities can be divided into transport services from individual modes of transport such as car, lorry, ship, etc., and collective modes of transport, such as buses, rail, and aeroplanes. Transport commodities can also be divided into mobile and immobile commodities, depending on the nature and accessibility of the supporting infrastructure. Cars can be regarded as mobile commodities, while ferries can be regarded as immobile commodities, as they are linked to locationally specific infrastructure (harbors), as is the case with trains. However, cars that use ferries and fixed links become immobile. Finally, the transport commodities can be divided into commodities that are bought externally from the point of view of the producer/consumer or transport commodities, which are bought by the producer/consumer (such as gasoline) in order to produce a transport service (such as lorry services in the production process or own transport for households as home production), which are internally produced transport services.

The demand for transport commodities – the real circle. The demand for transport commodity is a function of intra- and inter-regional interaction. Intra- and inter-regional interaction can be trade in commodities, shopping by type of demand (intermediate consumption, private consumption (including tourism), collective consumption, etc.), or commuting. The more widespread the interaction the bigger is the demand for the transport commodities. The theoretical model can thus be used as the point of departure for a more precise definition of the transport commodity.

First, immobile transport commodities can be defined as transport commodities where the place of production and the market place of the transport commodity are identical, as for example with a lorry using a harbor or a fixed link. The distinction between mobile and immobile transport commodities is of importance in the evaluation of regional economic consequences of changes in transport technology (e.g. the transition from ferries to fixed links), where immobile transport commodities will have a greater local economic effect than mobile commodities. Whether or not the distinction between mobile and immobile transport commodities is used in specific studies depends on the nature of the problem. For example, analysis of the regional economic consequences of a fixed link would require identification of the fixed link as an immobile transport commodity.

Second, in relation to internal transport commodities, where a firm or household owns its own vehicles, the transport commodities are produced internally and include both the investment expenditure related to purchase of the vehicle and the operating expenses. This division between external and internal transport is important in relation to the determination of the transport sector's importance, an issue that is dealt with in transport satellite accounts. The consequences of changes in demand for external transport commodities have a much greater effect on the economic activity of transport firms, while changes in internal transport costs will have direct effects on other types of firm, such as garages.

Third, the transport modes can be treated as different transport commodities, for example the division between goods and passenger transport modes. The first is usually referred to as transport related to production while the second is usually related to private consumption. This distinction is imprecise in relation to the division between intermediate and final goods, as passenger transport can also include productive activity, where passenger transport is an intermediate commodity and where goods transport can also include private consumption, where households use lorries supplied by a delivery firm. By using the commodity dimension a much more precise description of the economic function of the transport sector becomes possible, for example by subdivision of the car commodity into a household use and a firm use commodity.

The focus here is upon the economic function and the economic effects of transport activities, which means that it is relevant and necessary to use commodity definitions in relation to transport. The traditional subdivision of means of transport, undifferentiated by commodity, is related to traditional transport modeling and traffic planning, where traffic flows, capacity issues, and routing problems are the central issues of interest.

As described, the transport commodity has a function as a mark-up adding to the price of the commodity on its way from the place of production to the market place and further to the place of residence of the end user or the place of consumption of other producers. These transport mark-ups depend on: production costs in relation to the transport commodity; productivity in production of transport commodities; transport costs for the producers of the transport commodity in relation to delivery of the transport commodity to the market place; and transport costs in relation to accessing the transport commodity both by firms and households at the market place for the transport commodity.

The way in which the transport mark-up affects the commodity price of the commodities that are transported depends on the type of transport commodity involved. In the case of immobile transport commodities, the local economic effects of a change in transport mark-up are usually greater. The transport mark-up for other commodities depends on whether the immobile transport commodity is used in the transportation of the other commodities. For immobile transport commodities, the transport margin depends on local production costs and productivity in production of the immobile transport commodities. In the case of externally produced transport commodities compared with those that are internally produced, the transport mark-up depends on relative productivity levels and production costs.

Linking the real and cost–price circles in the spatial economy. The link from the economy to transport is through demand for the transport commodity in the real circle (Figures 7 and 9). The link from changes in the prices of the transport commodity occurs through the cost–price circle, where changes in transport costs affect price indexes, in turn affecting demand and real disposable income. These changes can include transport infrastructure changes, changes in prices and costs in the transport sector, and changes in commodity taxes for the transport commodity, including road pricing and fuel taxes.

4. The key dimensions

The issues illustrated by LINE indicate how to analysis of the complex relationships between transport and economic growth can be defined using a number of model dimensions, will be used to examine specific studies.

Macroeconomic and microeconomic approaches

Macroeconomic approaches deal with the way in which transport contributes to overall economic growth. As Vickerman (2002) points out, there are three basic ways in which transport can fit into a typical growth model:

- through the enhancement of investment and productivity;
- as a contributor to market integration, working through expansion of demand and dynamic scale economies;
- increasing efficiency as an endogenous contribution to total factor productivity, including greater openness to trade and ease of technology transfer.

This third mechanism deals with impacts on the rate of growth rather than the level of economic activity. Microeconomic approaches on the other hand are based upon modeling the behavior of firms and households, given changes in the transport system. Cost–benefit analysis of transport investment projects is based upon microeconomic theory. Here a key issue is the assumption that firms in sectors using the transport commodity are perfectly competitive, in which case a reduction in transport costs will appear directly in price changes, implying that economic benefits can be measured by willingness to pay for use of the transport system, an assumption upon which cost–benefit analysis is based.

Perfect or imperfect competition?

The widespread use of cost–benefit analysis in the evaluation of the benefits of transport system investments is based upon the assumption that transport users operate under conditions of perfect competition. If this is not the case, then as Venables and Gasiorek (1999) have pointed out, cost–benefit analysis will not, in general, correctly evaluate the benefits. Even when perfect competition can be assumed, cost–benefit analyses do not always incorporate the full regional economic effects, as multiplier effects, and effects on the labor, land, and property markets are often not included. Furthermore, if there is market failure in the transport sector, it is assumed that these consequences can be properly valued. If transport-using firms operate under conditions of perfect competition, then the benefits of a cost reduction in transport will immediately transplant themselves into the economic system and can be measured as benefits (which should include changes in regional economic activity).

If the transport-using firms operate in imperfect markets, however, where price (p) does not equal marginal cost (mc) (where either p > mc, which implies that private marginal benefits (pmb) are less than social marginal benefits (smb) or, less commonly, p < mc, implying that pmb > smb), then assessment of benefits is more problematic, for a number of reasons. First, firms with market power may not pass on transport cost reductions. Second there may be situations where poor accessibility protects a geographical monopoly. Third, the effect of transport cost reductions on the imperfectly competitive sectors is important for the assessment of benefits. If these are exposed to competition and they contract, then benefits will be greater, and the converse is also true. Fourth, improved transport can also reinforce the effects of scale economies in a sector, reducing the number of firms. Figure 8 illustrates the different possible combinations of imperfect markets in transport-using sectors (columns) and the transport sector itself (rows). Column 2 shows the case of perfect competition in transport-using sectors. Cell E is where there is perfect competition in both and where cost–benefit analysis will include the full benefits, if correctly specified. Cost–benefit analysis will evaluate benefits correctly in both cells B and H if the transport-related externalities are correctly valued and again if all benefits are included. In column 3, cost–benefit analysis will not, in general, provide a correct assessment of benefits, as the consequences of the transport improvements for firms that are imperfectly competitive are varied. There can also be cases where firms are charging a price that is below marginal cost and are receiving a subsidy, as indicated in column 1. In this case, transport improvements could, for example, replace partly or completely a subsidy designed

	Transport-using sectors (subsidies): p < mc (pmb > smb)	Transport-using sectors (perfect competition): p = mc (pmb = smb)	Transport-using sectors (imperfect competition): p > mc (pmb, smb)
Transport prices < msc Negative externalities Congestion User prices too low	(A) General subsidies and unvalued external costs. Too much traffic	(B) Economic benefits overestimated with congestion (external effects)	(C) Transport and transport user benefits are of opposite sign
Transport prices = msc No externalities Optimal capacity User charges correct	(D) Benefits overestimated as subsidies are reduced	(E) Economic benefits = transport benefits	(F) Total economic benefits > transport benefits
Transport prices > msc Positive externalities Spare capacity User charges too high	(G) Transport benefits and transport user benefits are of opposite sign	(H) Expand transport usage by reducing user charges	(I) Spare capacity in transport sector understates total economic benefits. Reduce user charges to give welfare gains

Figure 8. Possible combinations of imperfect markets in transport-using sectors (columns) and the transport sector itself (rows). (Based on Standing Advisory Committee on Trunk Road Assessment, 1999.)

to compensate for poor accessibility. In the first and third columns, benefits are not correctly reflected in the market demand curve.

Reduced or structural form?

An important distinction can be made between single-equation and structural-equation models. Single-equation, reduced-form, models are models where the endogenous variable is on the left-hand side of an equation, being a simple function of the exogenous variables. While they are typically easy to formulate and have more limited data requirements, there are a number of reasons for preferring structural-equation models. They provide an improved capability for formulation of an overall and consistent theoretical framework for explanation of the behavior of regional actors. They also reflect more closely the fact that regional economic growth is a complex process involving a large number of sub-processes for different actors in the regional economy with interaction between these actors. On the other hand, structural-equation models are more difficult to operationalize because of data problems.

Types of interaction

An important dimension is whether the modeling approach deals with all four principal types of traffic flow (trade, commuting, shopping, and tourism) or with only some of them. Changes in one type of flow will affect other types.

Treatment of space

An issue closely related to which types of interaction are treated is whether or not the approach makes a clear distinction between the location of production, the residence, and the market (demand), both for commodities and for factors.

The transport commodity

The way in which the transport commodity is treated in the modeling approach is important. The transport commodity may appear independently in the model, or sometimes it is simply assumed that a certain percentage of any transported commodity disappears under transport, representing the cost of transport (the "iceberg principle" – Samuelson, 1954). Further, if the transport commodity is modeled explicitly, is a distinction made between the users and producers of the transport commodity?

Partial or general models?

Here the prime distinction is between whether the market for the transport commodity is modeled in isolation or whether the market for transport is modeled as one of the set of all markets. A further distinction can be made between models that distinguish between commodity markets and factor markets.

Demand or supply side modeling

Whether the demand side of the economy or the supply side, or both, are modeled explicitly is an important feature distinguishing different models. This is in turn related to the treatment of the transport commodity. Models can also be demand driven, supply driven, or both.

5. Different modeling approaches: concrete studies

5.1. Transport modeling

The theoretical foundation of transport modeling is the four-step sequential transport model (Wilson, 1974). The basic model structure is shown in Figure 9. Given estimates of future demography and economic activity by zone in step 1, the following three steps derive the consequences for traffic flows and the transport system.

In this approach, a complex transport commodity is used, the users of the transport commodity are included, it is a partial model, dealing only with the

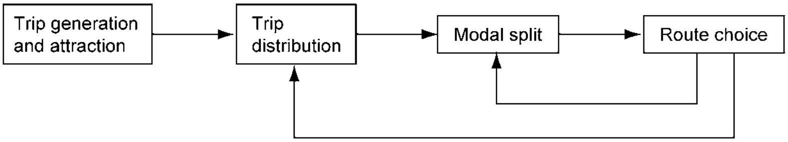


Figure 9. The sequential transport model.

transport sector, and it is in structural form. Different types of interaction can be modeled, though there is not explicit differentiation between the location of production. The model is not designed to estimate the regional economic effects of transport system improvements; rather, the causal relation is the reverse – given exogenous changes in economic activity by zone, what are the consequences for traffic flows and the transport system? It should, however, be noted that the model is able to incorporate other effects, such as road pricing and fuel taxation.

5.2. *Production function models*

There are a number of aggregate models that seek to measure the contribution of transport infrastructure to economic growth in a region or nation. A typical model, in the time series version, has the following general form:

$$Y_{tr} = Af(L_{tr}, K_{tr}, TI_{tr}, OI_{tr}),$$

where Y is value added, A is technological change, L is employment, K is capital stock, TI is transport infrastructure and OL is other infrastructure, for region r at time t. In relation to the dimensions identified above, the approach is macroeconomic, has the single-equation reduced form, and operates with a simple transport commodity; the model is supply side driven, and is partial as factor markets are at the core of the explanation, and there is no explicit modeling of the demand and supply sides of the economic system. Aschauer's (1989) seminal paper uses a Cobb–Douglas formulation that indicates that in the USA the elasticity for non-military infrastructure investment in relation to output per unit capital was as high as 0.39 and for core infrastructure in relation to productivity was 0.24. This suggests that the contribution of infrastructure to output growth is considerable. Since the publication of this paper, estimates of output elasticity have been reduced markedly, and Aschauer's work has been criticized on econometric grounds, including the existence of common trends, the direction of causality, and possible missing variables (Gramlich, 1994). More recently, Lau and Sin (1997) suggest that output elasticities are of the order of 0.1. A large number of studies using this basic approach have been reviewed by a number of authors (Banister and Berechman, 2000; Rietveld and Nijkamp,

2000) An alternative to production– function-based models is cost–function-based models with the following type of specification:

$$C = f(\bar{w}, K_{\mathrm{p}}, K_{\mathrm{g}}, t, Y),$$

where C is the variable cost function for the private sector, $\bar{w}$ is a vector of input prices, K_{p} is private capital, and K_{g} is government capital, t is technological progress, and Y is output. Then, the partial derivatives represent the shadow prices of each right-hand side variable. There are many problems involved in this approach, including measuring the price of public capital and choice of functional form. Seitz (1993) has used this type of model to examine marginal benefits from the expansion of transport infrastructure, concluding that investment in capital infrastructure enhances profitability of private capital.

5.3. *Accessibility models*

A number of studies use changes in accessibility after investment in transport infrastructure to estimate changes in regional economic activity (e.g. Rietveld and Bruinsma, 1998). This approach is based upon establishing an aggregate measure of accessibility for each zone before a change in the transport system occurs. This measure is then recalculated after the change, and a measure of change in aggregate accessibility for each zone derived. A typical measure of aggregate accessibility is economic potential:

$$P_i = \sum_{j=1}^{n} E_j c_{ij}^{-\beta},$$

where P_i is economic potential in zone i, E_j is a measure of economic activity (such as employment) in zone j, c_{ij} is a measure of interzonal transport cost, and β is a parameter. P_i is thus a measure of accessibility to economic activity in the entire zonal system. The problem remains of how to relate changes in accessibility to changes in economic activity. Proportionality is sometimes assumed, or an estimate of the proportional change is derived from comparative studies, where before and after data already exist. In terms of the dimensions identified, the approach is in a single-equation form, based upon a simple transport commodity, and is partial, addressing the commodity market, with no explicit modeling of the supply and demand sides of the economic system. Evers et al. (1987) used this approach to assess the regional economic effects of the proposed high-speed rail link from Brussels to Amsterdam and Hamburg. Jensen-Butler and Madsen (1999) have examined the regional economic effects of the Femern Belt link using this approach. In general, the effects predicted from this type of approach are more modest than those arising from the production function approach.

5.4. Inter-regional SAM models

Input–output models are a subset of SAM models, involving production sectors, final demand, and, in the inter-regional version, trade flows between regions to satisfy intermediate and final demand. Jensen-Butler and Madsen (1996) have used inter-regional input–output modeling to evaluate the regional economic effects of the Danish Great Belt fixed link. Essentially, regional economies are modeled using input–output techniques. From the individual regional model, inter-regional exports and imports by sector can be derived, using the method of computing residuals. Inter-regional exports in any one sector are found by subtracting sales to intermediate consumption and sales to final consumption from sectoral gross output. Inter-regional imports to a sector are found by subtracting primary inputs, international imports, and intermediate inputs from gross inputs. Adjustment is made for cross-hauling. These imports and exports are converted into monetary flows between regions and sectors in space, using double-constrained entropy maximizing methods, based upon a generalized transport cost function. Changes in employment, unemployment, GDP, and disposable income are among the measures of changes in economic activity that emerge from this type of study, though there is no direct measure of welfare changes.

When generalized transport costs between some regions change, there are two effects. First, there is a distributional effect, where the total inter-regional exports and imports are held constant and new trade patterns emerge, favoring some regions more than others. Second, there is a growth effect, arising from growth in inter-regional exports because of declining transport costs working through price elasticities. In the Danish case, the distributional effects appear to be more important than the growth effects. It is estimated that the Great Belt Link would create 2200 new jobs (about 0.1% of national employment), of which 1200 would be in the Greater Copenhagen area. In terms of the modeling dimension identified, the transport sector is identified, and the users and producers of transport are represented (though own production of transport in other sectors is not explicitly modeled). Commodity and factor markets enter into the overall model, though only the commodity markets are modeled explicitly, so the model is perhaps general rather than partial in its approach. The model is structural rather than reduced in form, though it only contains a real circle: the model is demand-driven and there is no explicit supply side modeling. As is well known from input–output analysis, there is no supply side reaction and there are no supply side constraints on expansion of demand for commodities. Technological coefficients are constant, and in addition to fixed proportions technology, there is fixed proportions consumption. While this approach models explicitly regional economies, albeit on the basis of flows of sectoral production in monetary terms, rather using commodities, benefits (or costs) arising from (supply side) changes in the transport system have to be modeled in an *ad hoc* manner, such as through the introduction of price elasticities to adjust final demand.

5.5. *Land use/transport interaction models*

Land use/transport interaction (LUTI) models link changes in location, land use, and economic activity to changes in traffic flows brought about by changes in the transport system. While Wilson (1974) adopted a broad definition of this class of model, these models are now, in a narrower definition, principally associated with models developed by two UK consultancy firms, which are similar to the SAM-based models. The basic aim of these models is to forecast physical flows of freight and passengers and associated changes in land use. The MEPLAN model for the Channel Tunnel (ACT et al., 1992) is of this type, and has three sub-models. The first is a regional economic model, which is input–output based with three elements: determination of regional final demand; determination of the level of regional production and regional imports; and changes in regional final demand that create inter-regional trade flows. The second is a conversion model, which converts annual values into tonnes and passenger trips. The third is the transport model that predicts modal split and route assignment for inter-regional flows. As accessibility changes because of infrastructure investment, these models concentrate on locational choice and distributional effects of infrastructure investment, tending to ignore net additions to economic activity because of the transport system change. These models share the same problems as input–output approaches, noted above, and the measures of benefits employed in the approach are similar.

5.6. *LINE*

This inter-regional model developed for the Danish economy is operational in a structural form. It operates at a high level of spatial, sectoral, and commodity disaggregation, which is regarded as necessary because of the complex nature of the transport–economy interaction. At present, the model is basically linear in its formulation and assumes perfect competition in both transport and transport-using markets, which are modeled explicitly. The model uses explicitly the concepts of places of production, places of residence, and places of markets, and all four types of interaction (trade, commuting, shopping, and tourism) are incorporated into the same model. These simplifying assumptions mean that LINE is a model developed in the CGE tradition, with interaction between supply and demand sides of the economy, while at the same time it is strongly anchored in empirical data and operates at a high level of disaggregation and detail. A set of benchmark values for a given year is established for key variables, at the regional or subregional levels, and these can be at high levels of disaggregation in terms of sectors, institutions, and factors of production, depending upon the task in hand. A disturbance is introduced into the system, for example a region-specific change in transport costs. The model is solved iteratively, and converges on a single equilibrium, this being the counterfactual equilibrium, to be compared with the benchmark. A number of developments are being introduced, including the effects of congestion externalities, the introduction of imperfect competition, and inclusion of substitution effects, between inputs to production and between commodities consumed.

The standard version of LINE has 12 industries, aggregated from 133; for factors, there are seven age, two gender, and five educational groups; for households, there are four types, based upon composition; for needs, there are 13 components of private consumption, eight components of governmental consumption, and ten components of gross fixed capital formation; there are 20 commodities, aggregated from 131; and, finally, there are 277 regions (municipalities), usually frequently aggregated to 17 counties. LINE has been designed with a high degree of flexibility, so that different elements of the model can be used in different types of analysis, or they can be removed by aggregation.

The model has been used in a number of Danish studies. One example is the case of modeling the removal of substantial tolls on the key east–west fixed link in Denmark, over the Great Belt. These tolls are around US $30 for cars and about five times this figure for lorries.[a] The model provides estimates for 17 Danish regions of price changes, for demand met by domestic production, for intermediate and final demand, for gross output, for exports and imports, and for private consumption both at the place of demand and at the place of residence, when tolls are set to zero. This is followed by calculation of the consequences for demand, production, and income, with estimates of changes in disposable income, real disposable income, private consumption, both by place of residence and by place of demand, imports, exports, gross output, and GDP at factor prices. Finally, changes in employment at the place of production and the place of residence, and changes in earned income, income transfers, taxes, and in nominal disposable income, are calculated. At the national level, employment is estimated to rise by 0.21% and GDP by 0.19%. There are important differences in the geographical incidence of benefits, which are largely are positive as perfect competition is assumed, though in places there are negative effects as route choice changes, because of a new configuration of routes.

[a]The assumptions used and the full specification of the model equations is to be found in Madsen and Jensen-Butler (2004).

5.7. Spatial CGE models

The Scottish CGE model is a two-region model with Scotland and the rest of the UK (Harrigan et al., 1991) that has been used to assess inter-regional effects of economic change. As a two-region model, transport costs are not modeled explicitly. Ferguson et al. (2003) examines the effect of an increase in exports from Scotland on GDP in both regions under different wage setting conditions in the regional labor markets.

CGE models are being used increasingly to evaluate the consequences of changes in the transport system. Changes examined include new transport infrastructure (Oosterhaven et al., 2001), removal of institutional barriers to movement (Bröcker, 1998), and implementation of road pricing (Munck, 2003). Pilegaard (2002) has examined the welfare economic consequences of subsidizing commuting using a CGE model.

Bröcker (1999) analyses changes in welfare arising from construction of the "TEN-T" European transport corridors. These corridors include the Danish–German Femern Belt link and the Danish–Swedish Øresund link. The welfare gains are measured in relation to regional GDP (using the method of equivalent variations). More than 800 regions are used, but only two types of goods, tradables and non-tradables. There is one factor of production, and households consume both types of good. Firms use both types of good plus the single production factor to produce a non-tradable good and a regional subset of tradables. Households receive factor income, plus a net income flow from other regions, and income is spent on local non-tradable goods and a composite of tradables, stemming from all regions, and preferences are represented in Cobb–Douglas form. Firms combine tradables, local non-tradables, and factors to produce an intermediate good that is either transformed into a local good or into different brands of tradables, using Cobb–Douglas technology, and costs are determined by the intermediate good. Firms are price takers on the input and local good markets, and there is monopolistic competition in the market for tradables. This allows calculation of the price of tradables, derived from the sum of the cost per unit of the intermediate good and the number of brands. These assumptions permit derivation of demand and supply of tradables as functions of the regional prices of tradable output and the price per unit of a composite tradable, which includes transport costs. Transport costs are modeled using the iceberg principle, and trade flows are determined by a production-constrained gravity model. Factor income is derived from the two sets of prices. Thus, a set of equations provides estimates of regional demand and supply, factor income, trade flows, and prices, as a function of a limited set of exogenous variables and parameters. Calibration assigns values to parameters and exogenous variables so that the equilibrium solution reproduces the observed data. These are obtained from a variety of sources. Some are borrowed from the national accounts for each country, others are taken from international trade data and used at a regional scale, and others are taken from various sources of empirical evidence.

The results show relative welfare increases for all zones. In the case of the Øresund link, Copenhagen experiences a 0.08% increase, while for Malmö it is 0.5%. For Schleswig Holstein, and Hamburg, increases are only of the order of 0.01%. For the Femern Belt link, Copenhagen experiences a 0.24% increase, Storstrøm in Denmark (the region closest to the fixed link) 0.42%, Malmö 0.31%, and Schlwesig Holstein 0.03%. Some zones experience negative growth. The overall pattern is not surprising, and corresponds to that found by Jensen-Butler and Madsen (1999). The main benefits of these southward-oriented links fall in Scandinavia, as they achieve better access to the large German and European markets to the south. The potential gains for Germany are much more modest.

This study illustrates the strengths and weaknesses of the spatial CGE approach. Its principal strength is that it builds directly upon microeconomic theory; it is a model in structural rather than reduced form, incorporating imperfect competition and product diversity, and both supply and demand sides are represented, though at unrealistically high levels of abstraction. Apart from the regional dimension, it operates at a very high level of sectoral and institutional aggregation, with only one homogenous production factor; there is no distinction between the place of residence and the place of the market, and it deals only with one type of interaction – trade; the transport commodity is not treated explicitly, and transport costs are not treated satisfactorily through the use of the iceberg principle. The link to the empirical base is very tenuous.

References

ACT, Institut für Raumplanung, Universität Dortmund and Marcial Echenique (1992) *The regional impact of the Channel Tunnel throughout the Community. Final report for the directorate General XVI of the Commission of the European Communities*. Paris: ACT Consultants. Dortmund: IRFUD. Cambridge: Marcial Echenique.

Arrow, K.J. and F.H. Hahn (1971) *General competitive equilibrium*. San Francisco: Holden-Day.

Aschauer, D.A. (1989) "Is public expenditure productive?" *Journal of Monetary Economics*, 23:177–200.

Baum, H. and J. Korte (2002) "Introductory report," in: *Transport and economic development. Report of the 119th Round Table on Transport Economics*. Paris: ECMT/OECD.

Banister, D. and J. Berechman (2000) *Transport investment and economic development*. London: University College of London Press.

Brakman, S., H. Garretsen and C. van Marrewijk (2001) *An introduction to geographical economics*. Cambridge: Cambridge University Press.

Bröcker, J. (1998) "How would an EU-membership of the Visegrád countries affect Europe's economic geography?" *Annals of Regional Science*, 32:91–114.

Bröcker, J. (1999) "The economic impacts of the TEN projects on the Øresund region," in: A.K. Andersen and A. Karlström, eds, *The regional development impacts of the Øresund bridge*. Stockholm: Department of Infrastructure and Planning, Royal Institute of Technology.

Dixit, A.K. and J.E. Stiglitz (1977) "Monopolistic competition and product diversity," *American Economic Review*, 67:297–308.

European Conference of Ministers of Transport (2001) *Assessing the benefits of transport*. Paris: ECMT/OECD.

Evers, G.H.M., P.H. Van der Meer, J. Oosterhaven and J.B. Polak (1987) "Regional impacts of new transport infrastructure: a multisectoral potentials approach," *Transportation*, 14:113–126.

Ferguson, L., D. Learmonth, P.G. McGregor, D. McLellan, J.K. Swales and K. Turner (2003) "The importance of macroeconomic constraints in the operation of a two-region computable general equilibrium model," in: *North American Regional Science Association Annual Conference*. Philadelphia.

Fujita M., P. Krugman and A.J. Venables (1999) *The spatial economy. Cities, regions and international trade*. Cambridge: MIT Press.

Gramlich E.M. (1994) "Infrastructure investment: a review essay," *Journal of Economic Literature*, 32:1176–1196.

Haddad, E.A., G.J.D. Hewings and M. Peter (2002) "Input–output systems in regional and inter-regional CGE modeling," in: G.J.D. Hewings, M. Sonis and D. Boyce, eds, *Trade, networks and hierarchies*. Heidelberg: Springer-Verlag.

Harrigan, F., P. McGregor, R. Perman, K. Swales and Y.P. Yin (1991) "AMOS: A macro–micro model of Scotland," *Economic Modeling*, 8:424–479.

Hewings, G.J.D. and M. Madden, eds (1995) *Social and demographic accounting*. New York: Cambridge University Press.

Isard, W. and I.J. Azis (1998) "Applied general inter-regional equilibrium," in: W. Isard, I. Azis, M.P. Drennan, R.E. Miller, S. Saltzman and E. Thorbecke, eds, *Methods of inter-regional and regional analysis*. Aldershot: Ashgate.

Jensen-Butler, C.N. and B. Madsen (1996) "Modeling the regional economic effects of the Danish Great Belt link," *Papers in Regional Science*, 75:1–21.

Jensen-Butler, C.N. and B. Madsen (1999) "An eclectic methodology for assessment of the regional economic effects of the Femern Belt link between Scandinavia and Germany," *Regional Studies*, 3:751–768.

Kehoe, P.J. and T.J. Kehoe (1994) "A primer on static applied general equilibrium models," *Federal Reserve Bank of Minneapolis Quarterly Review*, 18, No 2.

Krugman, P. (1998) "What's new about the new economic geography?" *Oxford Review of Economic Policy*, 14:2.

Lau, S.-H.P. and C.Y. Sin (1997) "Public infrastructure and economic growth: time series properties and evidence," *Economic Record*, 73:125–135.

Madsen, B. and C.N. Jensen-Butler (2002) "Regional economic modeling in Denmark: construction of an inter-regional SAM with data at high levels of disaggregation, subject to national constraints," in: G.J.D. Hewings, M. Sonis and D. Boyce, eds, *Trade, networks and hierarchies*. Heidelberg: Springer-Verlag.

Madsen, B. and C.N. Jensen-Butler (2004) "Theoretical and operational issues in sub-regional economic modeling, illustrated through the development and application of the LINE model," *Economic Modeling*, 21:471–508.

Madsen, B., C.N. Jensen-Butler and P.U. Dam (2001a) *The LINE-model*. Copenhagen: AKF-Forlaget.

Madsen, B., C.N. Jensen-Butler and P.U. Dam (2001b) *A social accounting matrix for Danish Municipalities, SAM-K*. Copenhagen: AKF-Forlaget.

McGregor, P.G., J.K. Swales and Y.P. Yin (1995) "Migration equilibria in regional economies: a multi-period CGE analysis of an improvement in local amenities," in: J.C.M.J. Van den Bergh, P. Nijkamp and P. Rietveld, eds, *Recent advances in spatial equilibrium modeling*. Heidelberg: Springer-Verlag.

McGregor, P.G., J.K. Swales and Y.P. Yin (1998) "Spillover and feedback effects in general equilibrium inter-regional models of the national economy: a requiem for inter-regional input–output?" in: G.J.D. Hewings, M. Sonis, M. Madden and Y. Koimura, eds, *Understanding and interpreting economic structure*. Heidelberg: Springer-Verlag.

Munck, K.J. (2003) "Assessment of the introduction of road pricing using a computable general equilibrium model," in: *TRIP Seminar on the Economic and Environmental Consequences of Regulating Traffic*. Copenhagen: AKF.

Nijkamp, P. and E. Blass (1994) *Impact assessment and evaluation in transportation planning*. Dordrecht: Kluwer.

Oosterhaven, J., T. Knaap, C. Ruijgrok and L. Tavasszy (2001) "On the development of RAEM; the Dutch spatial equilibrium model and its first application to a new railway link," in: *41th Congress of the European Regional Science Association*. Zagreb.

Partridge, M.D. and D.S. Rickman (1998) "Regional computable general equilibrium modeling: a survey and critical appraisal," *International Regional Science Review*, 21:205–248.

Petersen, T.W. (1997) "Introduktion til CGE modeler," *Nationaløkonomisk Tidsskrift*, 135:113–134.

Pilegaard, N. (2002) "Commuting behaviour, regional differences and labor market imperfections – in a CGE framework," in: *TRIP Seminar on the Economic and Environmental Consequences of Regulating Traffic*. Copenhagen: AKF.

Prud'homme, R. (2002) "Introductory report," in: *Transport and economic development. Report of the 119th Round Table on Transport Economics*. Paris: ECMT/OECD.

Rietveld, P. (1996) "Transport infrastructure, productivity and employment," in: B. Madsen, C.N. Jensen-Butler, J.B. Mortensen and A.M.B. Christensen, eds, *Modeling the economy and the environment*. Heidelberg: Springer-Verlag.

Rietveld, P. and F. Bruinmsa (1998) *Is transport infrastructure effective?* Heidelberg: Springer-Verlag.

Rietveld, P. and P. Nijkamp (2000) "Transport infrastructure and regional development," in: J.B Polak and A. Hertjee, eds, *Analytical transport economics*. Cheltenham: Elgar.

Round, J. (1988) "Incorporating the international, regional and spatial dimension into a SAM: some methods and applications," in: *Recent advances in regional economic modeling*. London: Pion.

Samuleson, P.A. (1954) "The transfer problem and transport cost. II: analysis of effects of trade impediments," *Economic Journal*, 64:264–289.

Seitz, H. (1993) "A dual economic analysis of the benefits of the public road network," *Annals of Regional Science*, 27:223–239.

Shoven, J.B. and J. Whalley (1992) *Applying general equilibrium. Cambridge surveys of economic lliterature*. Cambridge: Cambridge University Press.

Standing Advisory Committee on Trunk Road Assessment (1999) *Transport and the economy*. London: SACTRA, UK Department of the Environment, Transport and the Regions.

Van den Bergh, J.C.M.J., P. Nijkamp and P. Rietveld (1996) "Spatial equilibrium models: a survey with special emphasis on transportation," in: J.C.M.J. Van den Bergh, P. Nijkamp and P. Rietveld, eds, *Recent advances in spatial equilibrium modeling*. Heidelberg: Springer-Verlag.

Venables, A.J. and M. Gasiorek (1999) *The welfare implications of transport improvements in the presence of market failure. Part 1*. London: SACTRA, UK Department of the Environment, Transport and the Regions.

Vickerman, R. (2002) "Introductory report," in: *Transport and economic development. Report of the 119th Round Table on Transport Economics*. Paris: ECMT/OECD.

Wilson, A. (1974) *Urban and regional models in geography and planning*. London: Wiley.

World Bank (1994) *World development report 1994: infrastructure and development*. Oxford: Oxford University Press.

This Page Intentionally Left Blank

Chapter 13

INFRASTRUCTURE POLICY

ROGER VICKERMAN
University of Kent, Canterbury

1. Introduction

Policy toward transport infrastructure can be divided into two main parts: investment in new infrastructure; and the management, regulation, and allocation between users of existing infrastructure. Although early transport infrastructure was often provided by the private sector through turnpike roads or railways, the state has always had an interest, not least because of the role of transport infrastructure in national defense. However, state involvement in both provision and management of infrastructure is for reasons more than this: the economic characteristics of infrastructure involving elements of public good, externality, and natural monopoly imply a state interest. In this chapter we consider how policy toward infrastructure has developed in four main parts. First, we consider the basic economic characteristics of infrastructure in order to define what objectives the state might have for both provision and management. Secondly, we examine the possibility of a self-financing transport infrastructure system; the public utility question. Thirdly, we examine the role of private finance in public infrastructure. Fourthly, we consider the problems associated with the maintenance and operation of infrastructure and how to set efficient regulatory incentives to infrastructure managers or private sector providers.

2. Infrastructure characteristics and policy objectives

Infrastructure consists essentially of large, lumpy investments characterized by a high degree of asset specificity. This makes decisions on the appropriate level of investment difficult since it is frequently necessary to choose between investments that are smaller or larger than the apparent optimal capacity to serve revealed demand. Once provided, capacity can only be changed in discrete steps. Most transport infrastructure is fixed in its location and use, and thus has no value if it becomes redundant. Since capacity is fixed, at levels of use below capacity (no

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

congestion) the marginal cost of using infrastructure is zero (Dupuit, 1844). It is these inherent characteristics of infrastructure that provide the basic argument for public sector provision: there will be a basic market failure in which a private market will find it difficult to finance and provide an optimal level of capacity.

Even if a private sector operator is able to provide infrastructure, the public sector has an interest because of the natural monopoly problem. Lumpiness and asset specificity imply that there will in many circumstances only be one provider of infrastructure on any one route, and the state will wish to avoid the provider exploiting this monopoly power. Natural monopoly implies high barriers to entry, not least because of the time taken to provide a competing infrastructure. The state will thus need to regulate the infrastructure provider, and in many cases problems with regulation will lead to direct provision or nationalization of the service.

In addition to the market failure rationale, there is an externality rationale for state interest in infrastructure provision. It is argued that infrastructure may lead to wider economic benefits than those captured by any charge that a provider can impose on direct users (Aschauer, 1989; Gramlich, 1994; Standing Advisory Committee on Trunk Road Assessment, 1999). This may justify the provision of infrastructure out of general taxation rather than through direct charging for use.

Thus we have a basic argument that appears to imply that infrastructure should be provided, or at least strictly regulated, by the state. But how does the state determine the level of capacity to be provided and how it should be allocated between competing users, whether these are individual final consumers such as on roads or transport service providers as with airports or rail infrastructure? There are two aspects to this question. First, we face the typical public goods problem of how to get users to reveal their demand for infrastructure at levels of usage below capacity. Second, we need to resolve the question of how to fix charges for use, especially where short-run marginal cost is zero, or close to zero.

For a long period, transport infrastructure provision, especially roads provision, was based in most countries on a "predict and provide" policy. The key to this was simply predicting the level of demand using standard transport models and allocating this to the network. Where network capacity failed to meet predicted demand there was a prima facie case for investment in additional capacity. Increasingly the demand predictions became more sophisticated, allowing for the evaluation of time savings, reductions in accident costs, etc., but the basic principle remained. Little consideration was given in policy terms to the allocation of capacity between users despite the increasing recognition of the congestion problem and its formal analysis by such contributors as Tinbergen (1957) and Walters (1961). The line that was followed was essentially the zero marginal cost rule, first expounded by Dupuit (1844) for the use of infrastructure with significant sunk costs and zero opportunity costs and later developed by writers such as

Pigou (1920) and Knight (1924), who treated the road allocation problem as a difficult case.

In many countries, therefore, the cost of road infrastructure was met by a fixed annual charge entitling the vehicle owner to access to the network with no further direct charges. The earlier attempts to create payments for infrastructure through, for example, turnpike roads had often resulted in failure, as high charges led users to choose alternative toll-free routes, even where these were of lower quality. The exceptions to this were typically where specific infrastructure such as bridges or tunnels granted monopoly power to owners. Only when limited-access highways were developed was there a return in many countries to the turnpike idea of toll charging. This was based on the provision of a higher-quality infrastructure, essentially one with less congestion, than the free to access public road infrastructure.

The development of policy toward other transport infrastructure reflected similar priorities. By the end of the first quarter of the twentieth century most railways, with the exception of those in North America, were either fully nationalized or largely under public sector control. Investment in infrastructure thus became more subject to the financial constraints of government than to rigorous economic appraisal. Allocation of capacity was an internal matter for the railway, which determined priorities for different types of traffic largely on administrative grounds. The development of air travel reflected the same general approach, with publicly owned airports allocating landing slots in a largely administrative way. International flights in particular were heavily regulated and dominated by flag-carrying airlines, which, again with the exception of North America, were state-owned, heavily state-financed, national flag-carrying airlines.

In all of these cases the development of networks occurred largely haphazardly in response to perceived localized needs or pressures without any real thought given to the logical development of networks. Not least among these problems was the lack of consideration of how development in one part of a network may affect the rest of the network. Only in the second half of the twentieth century did policy start shifting toward consideration of network development, but frequently this was aimed at the wider economic impacts believed to be associated with infrastructure development than with consideration of the appropriate policy for the management of the infrastructure itself. Thus, in the USA the development of the Interstate Highway system in the 1950s was seen as a means of accelerating the development of the economy by providing a federal level infrastructure not subject to the priorities of individual states (Mohring and Harwitz, 1962). This model was used as a reference point by the EU in the 1990s, when it moved to establish a set of trans-European networks covering most modes of transport as a means of promoting both increased competitiveness and enhanced cohesion (European Commission, 1994; Vickerman, 1994; Turro, 1999). Similarly, individual countries have established national infrastructure plans such as the Schéma

Directeurs for roads and high-speed rail in France or the Bundesverkehrswegeplan in Germany.

In this development we see that infrastructure has largely been seen as an instrument toward achieving wider policy goals rather than the subject of policy in its own right. This has tended to reduce the effectiveness with which infrastructure has been integrated into overall transport planning.

3. Infrastructure as public utility

Unlike other network industries such as energy or telecommunications, there has been much less consideration of transport infrastructure as a public utility that can be provided in a self-contained manner, by one or more operators, subject to overall government regulation. The basic analytical framework of this was, however, laid out by Mohring and Harwitz (1962), who showed how the nature and measurement of highway benefits is affected by the way that highways are financed. This was a response to the provisions of the Act establishing the Federal Interstate System that established a highway trust fund designed to provide an explicit linkage between tax revenues associated with highways and federal highway expenditures.

The problem of such provisions, rather like the problem that had earlier been encountered by the Road Fund in the UK (Barker and Savage, 1974), is how to allocate the tax burden between users and non-users if in a period of heavy investment the receipts from the former are insufficient to cover expenditure. Secondly, if tax receipts from users are treated as part of general tax revenues, how can the linkage be established between the revealed demand for road space and expenditure on roads? The taxation of road users has, of course, increasingly been for reasons more than just the provision of road space. Concern for the environmental impact of traffic and as an imperfect means of controlling congestion have led fuel taxes in many countries, especially in Europe, to increase much faster than other costs. This has led to total tax receipts being much greater than the direct expenditure on the road network. This meets the usual prescription of public finance that taxes should not be hypothecated, but ignores the efficiency aspects of allocating infrastructure use.

It has, however, been shown by various authors that, allowing for the full costs of road transport including environmental, accident, and congestion costs, it is possible to envisage a self-financing road system that both generates sufficient revenues to cover its own full costs and provides a mechanism for the efficient allocation of capacity between competing users (e.g. Newbery, 1988a,b, 1994; Peirson et al., 1995; Sansom et al., 2001). In these circumstances it is only a short step to considering a road network as equivalent to any public utility. The usual objection to such an approach is the difficulty of charging the individual road user for an individual trip. Reliance on annual license fee charges and fuel taxes, supplemented by parking charges, tends to average out the costs between users, and results in cross-subsidy between different user groups, undercharging peak hour urban users, and overcharging off-peak rural users.

Direct road charging has been felt to be impractical except for tolls on limited-access highways, and even these typically charge a standard toll per vehicle type. Area-wide charging schemes, such as that introduced in London in 2003, or cordon charges such as those used in Oslo, or, prior to full electronic road pricing, in Singapore, have differentiated by time as well as vehicle type. However, these have been conceived as congestion charging schemes designed to allocate traffic to the existing road capacity rather than full road pricing schemes. Hence they are based (albeit somewhat arbitrarily in terms of the actual charge) on the difference in marginal cost between peak and off-peak, rather than being designed to recover the full cost of the network. Given the limited geographic scope of such schemes, retention of the existing taxes and charges on vehicle ownership and use has been necessary. In such circumstances it is difficult to convey to road users the full cost of using the network. This would be necessary before creating a genuine public utility approach, even though the continued use of multi-part tariff structures may be justified, as is common with energy or water utilities, which frequently have both a standing connection charge and a charge for use.

The other problem that arises with a road network as a public utility is the appropriate size of the network. There are two concerns here: the regulation of monopoly power and the fair charging of residents and non-residents. The regulation issue is not very different from that of any public utility, and we deal with this in the following section. A key to this has, however, been to achieve the proper balance between any scale economies, including network economies, and the promotion of competition. This implies there are either or both competing network and/or geographic limits to the network size. The latter allows for both potential competition and benchmark or reference competition. The incidence of charges on residents and non-residents is an issue specific to transport, however. The services of energy, water, and other utilities are delivered to a specific address, and hence both the connection charge and the usage charge can be billed to an identifiable user, and any attempt to avoid payment dealt with. Road users make use of road networks away from their home, and even if the solution was to create national road authorities there would be issues relating to international traffic as evidenced by the problems encountered in the introduction of charging systems for heavy goods vehicles in the EU. The problem that arises is, first, how to bill users and, second, how to avoid the incentive to shift tax burdens unfairly onto occasional users so as to subsidize resident users (De Borger and Proost, 2001).

4. Private finance and provision of infrastructure

The responsibility of the public sector for both planning and financing the provision of infrastructure has changed over time. In many countries, much of the initial development of transport infrastructure was undertaken by the private sector. This included railways in the UK and the USA, often leading to considerable competition; competition that often posed problems later. In other countries the state took a more direct interest from the start. In almost all of these cases, infrastructure owners were integrated operators of the infrastructure, and the services provided on the infrastructure. The increasing complexity of infrastructure and the financial problems being faced by many of these integrated operators led to an increasing movement toward state ownership from the 1920s and 1930s, reaching its high point in the period after the Second World War. This affected, for example, national railway operators in most European countries and urban public transit operations in many countries.

This problem in the provision of traditional public transport coincided with the need both to renew its infrastructure and to expand new infrastructure for the rapid rise in new means of transport, notably roads and airports. The burden of these massive investment programs was shouldered largely by the public sector in an era of mega-projects (Altshuler and Luberoff, 2003). This period was, however, fairly short lived, since by the 1970s two counter forces were beginning to impose restrictions on policies of the uncontrolled expansion of major infrastructures.

The first of these was a questioning of the policy in its own right. As we have noted above, the "predict and provide" philosophy that dominated this period was challenged. This was not least the result of increasing pressure to consider the environmental consequences of unlimited expansion, but as Altshuler and Luberoff point out, for many urban projects it was a growing recognition of the problems in using tax revenues for projects that had distinctly unequal impacts across communities, especially in urban areas.

The second problem was the more purely financial problem for the public sector in the face of the need to manage public sector deficits, which led the public sector in many countries to turn to the private sector as a means of financing infrastructure investment. Furthermore, concern at the inability of the public sector to manage infrastructure efficiently has also led to calls for increasing private sector involvement. The public sector's role thus changed from a primary concern with finance to one of prioritizing the planning of projects to be financed by alternative forms of private sector finance.

This private sector involvement in finance can take many different forms, from the complete handing over of responsibility to a private sector firm to develop and exploit an infrastructure, through the use of the concession or franchise to allow a private sector operator to exploit an infrastructure for a fixed period, at the end of which the infrastructure reverts to the public sector, to a partnership between public and private sectors whereby the private sector provides finance in return for accepting an agreed share of any risk. The broad principles are the same whatever the scale of the private sector involvement. The private sector provides investment finance, typically allowing an investment to take place earlier than might otherwise be possible, and accepts an agreed share of any risk. In return the private sector receives a cash flow that may arise either from any direct charge such as a toll, or from some imputed payment made by the public sector. The risk may arise from problems during construction, leading to higher costs, or from a failure to estimate accurately the level of usage and hence leading to lower revenues.

From the point of view of the public sector, this seems to be an effective solution to the financing problem. The private sector provides finance, which obviates the need for the public sector to borrow or to raise taxes and may also provide more effective management. The problem for the public sector is to determine the appropriate length of any concession agreement and/or the allocation of risk between the public and private sectors. The length of the concession will determine the total value of the concession to the private investor; it also determines the length of the period during which the public sector has no day to day control over the operation of the infrastructure. It is thus vital to ensure that the optimal conditions are placed on the concession when it is let. This is one element in the distribution of risk. The basic interest in involving the private sector is that the private finance takes a share of the risk; the private sector's interest will be to reduce as far as possible the degree of risk to which it is exposed. The problem is that there is imperfect knowledge, leading to problems of asymmetric information. If the private sector is involved in the construction, it may have much better knowledge of the potential construction risks than the public sector. On the other hand, the public sector may have better information about future levels of demand. Both of these factors could lead to higher costs in the development of the infrastructure, as the private sector seeks to insure itself against the downside risks. Such costs could outweigh any potential gain believed to derive from the greater efficiency of the private sector in the management of the development of the infrastructure. Flyvbjerg et al. (2003) have provided detailed evidence on the relative incidence of cost overruns on the construction of major infrastructure, showing that there is relatively little difference between public and private infrastructure in this respect.

There remains the question of how to generate revenues for the privately financed infrastructure project (or indeed for a publicly financed one within a public utility framework). Direct charging through tolls is the most obvious way of reflecting the full costs of the provision. An alternative method, used for example in the UK for a number of DBFO (design, build, finance, and operate) projects on the major highway network, is that of shadow tolls, where the state pays to the private operator a toll equivalent based on traffic flows out of general revenue.

This, of course, breaks the direct link between the eventual user and the perceived cost of using the infrastructure. The problem for the road user is that additional direct charges through tolls (or indeed congestion charges) are perceived as additional taxes on top of the existing range of taxes imposed on road users through annual vehicle licenses, fuel taxes, and parking charges. This suggests that it is problematic to introduce such charges piecemeal: they need to be part of an overall re-balancing of the charges for infrastructure use, with consistency between publicly and privately provided parts of the network, with any additional charges being clearly related to higher quality such as less congestion.

Turning to private finance does not remove all problems for the public sector since the main planning problems still remain with the public sector in most forms of private financing. Moreover, a number of new ones emerge. Given the differential willingness of the private sector to finance different types of project, the public sector may find that its priorities are difficult to sustain. This may lead to controversy, in which the private sector is seen to be more concerned with profit than the public interest. Thus, even in the case of private sector finance and management of infrastructure, the public sector will need to develop a regulatory structure.

5. Regulation of infrastructure

With infrastructure provided directly by the public sector, it is taken for granted that the infrastructure will be provided in the public interest. The introduction of a private sector interest raises the question of how the public sector's vested interest in public infrastructure should be exercised. There are two questions here: the regulation of monopoly control over the allocation of infrastructure use and the means of ensuring the infrastructure is maintained effectively over the lifetime of any concession.

Most privately provided infrastructure is provided under a concession or franchise in which the concessionaire is allowed to exploit the infrastructure for a fixed period. In most cases the charging policy of the infrastructure operator is regulated to avoid the exploitation of any monopoly advantage. This may be either via a direct control on prices, for example the regulation of tolls, which is common on bridge or motorway concessions, or via rate of return regulation, which is more common in conventional public utilities.[a]

Concession agreements require that the infrastructure is returned to the public sector at the end of the concession in good working order (a point originally noted by Pigou, 1920). It is therefore in the public interest to ensure that the infrastructure

[a]See Laffont and Tirole, 1993, for the basic analysis; Grout and Stevens, 2003, for a recent assessment, and for a discussion related specifically to transport, Helm and Thompson, 1991; Quinet and Vickerman, 2004).

is adequately maintained during the concession. Furthermore, this is not just a necessary condition at the end of the concession since failure to maintain the infrastructure during its lifetime will reduce the level of service provided and hence the value of the infrastructure. Small and Winston (1988) have provided a basic analysis of this problem, showing that traditional engineering approaches have overestimated road surface life and underestimated lifetime costs. If traffic is greater than that forecast, or consists of a different composition, for example relatively more heavy freight traffic than expected, then the lifetime of the asset may be shorter and incur higher maintenance costs. It is not just a case of the relationship between construction standards and lifetime maintenance. The management of maintenance has increasingly been seen to be one of the major concerns in both road and rail infrastructures, as shown by Vickerman (2003) in a comparison of road and rail maintenance regimes in the UK. The problem for any regulator is that the true quality of the infrastructure is only known to the infrastructure provider, and hence it is difficult to devise a regulatory system that forces the provider to reveal this true quality. This is compounded by a situation in which the infrastructure provider subcontracts management and/or maintenance of the infrastructure to others. Generally, neither a totally centralized regulatory system, in which the public authority sets precise targets for all levels of infrastructure provision, nor a totally decentralized system, in which the market is left relatively free to regulate itself, will produce an optimal output. As Caillaud et al. (1996) have shown, some coordination in incentive setting is desirable to ensure that information asymmetries are reduced.

6. Conclusions

Policy toward infrastructure involves a number of critical inter-relating decisions: investment, optimal allocation of capacity, method of financing construction, and regulation of operations. Much infrastructure is complex. Most infrastructure is part of a network, and, thus, even if parts can be separated off for private sector construction and operation, they cannot be managed completely independently of the rest of the network. The history of transport infrastructure provision has seen it move in most countries from an initial development in the private sector with minimal regulation to the public sector as it became more expensive for private sector firms to maintain it in the face of competition from alternative modes of transport, through a period of massive expansion in the public sector to coping with increasing pressures on public budgets which led to retrenchment and the re-introduction of private finance and management.

While we have outlined the principles that should underpin a coherent approach to infrastructure policy, there has been little evidence of such coherent policy in most countries. Infrastructure has always been subject to the pressures of vested interest groups and politically motivated decision, which often distort attempts to develop consistent policy. Frequently, policy toward transport infrastructure is motivated by non-transport ends, as seen in both the 1950s policy to create the Interstate Highway system in the USA and the later emergence of the trans-European networks in the EU. This is even more so at the local or regional level.

However, we have seen an increasing recognition of the need to introduce more consistent policy to the allocation of capacity on networks, particularly road networks, and to the evaluation of investments in alternative networks. Thus, the most recent policy document of the EU (European Commission, 2001) is based on the concept of the "fair and efficient pricing" of infrastructure, with consistent charging of all costs in all modes, as well as an intention to shift the balance between modes and eliminate bottlenecks. The first indication of the pricing policy is in the acceptance of a kilometer-based charge for goods vehicles on the major highways of several member states using satellite technology. However, it is still not clear that pricing proposals in this and other policies are fully integrated with investment appraisal and decisions. Rather as with the original use of the Road Fund in the UK, there is a fear that public authorities will be tempted to use infrastructure access charges as general tax revenues rather than contributions to external costs and the allocation of capacity. It is for this reason that there have been suggestions that roads, like other transport networks, should become part of a separate public authority and run like any public utility.

Infrastructure policy thus remains at the core of our understanding of the role of transport. There remains a need to identify more clearly the links between policy toward transport and its infrastructure and the use of transport as a means to achieving other wider policy goals in the national, regional, or local economy. The principles of this date back to some of the original scientific contributions on transport and have been added to by more recent work on the economics of regulation and information. However, vested interests and public concern have limited the full application of these principles; much work remains to be done.

References

Altshuler, A. and D. Luberoff (2003) *Mega-projects: the changing politics of urban investment*. Washington, DC: Brookings Institution.

Aschauer, D.A. (1989) "Is public expenditure productive?" *Journal of Monetary Economics*, 23:177–200.

Barker, T.C. and C.I. Savage (1974) *An economic history of transport in Britain*, 3rd edn. London: Hutchinson.

Caillaud, B., B. Jullien and P. Picard (1996) "Hierarchical organization and incentives," *European Economic Review*, 40:687–95.

De Borger, B. and S. Proost, eds (2001) *Reforming transport pricing in the European Union*. Cheltenham: Edward Elgar.

Dupuit, J.A. (1844) "De la mesure de l'utilité des travaux publics," *Annales des Ponts et Chaussées*, 8.

European Commission (1994) *Growth, competitiveness and employment*. Luxembourg: Office for Official Publications of the European Communities.

European Commission (2001) *European transport policy for 2010: time to decide*. Luxembourg: Office for Official Publications of the European Communities.

Flyvbjerg, B., N. Bruzelius and W. Rothengatter (2003) *Megaprojects and risk: an anatomy of ambition*. Cambridge: Cambridge University Press.

Gramlich, E. (1994) "Infrastructure investment: a review essay," *Journal of Economic Literature*, 32:1176–1196.

Grout, P. and M. Stevens (2003) "Financing and managing public services," *Oxford Review of Economic Policy*, 19:215–234.

Helm, D. and D. Thompson (1991) "Privatized transport infrastructure and incentive to invest," *Journal of Transport Economics and Policy*, 25:231–246.

Knight, F.H. (1924) "Some fallacies in the interpretation of social costs," *Quarterly Journal of Economics*, 38:582–606.

Laffont, J.J. and J. Tirole (1993) *A theory of incentives in procurement and regulation*, Cambridge: MIT Press.

Mohring, H. and M. Harwitz (1962) *Highway benefits: an analytical framework*. Evanston: Transportation Center, Northwestern University.

Newbery, D. (1994) "The case for a public road authority," *Journal of Transport Economics and Policy*, 28:235–254.

Newbery, D. (1998a) "Road user charges in Britain," *Economic Journal*, 98:161–176.

Newbery, D. (1988b) "Road damage externalities and road user charges," *Econometrica*, 56:295–319.

Peirson, J., I. Skinner and R. Vickerman (1995) "Estimating the external costs of UK passenger transport: the first step toward an efficient transport market," *Environment and Planning A*, 27:1977–1993.

Pigou, A.C. (1920) *The economics of welfare*. London: Macmillan.

Quinet, E. and R.W. Vickerman (2004) *Principles of transport economics*. Cheltenham: Edward Elgar.

Standing Advisory Committee on Trunk Road Assessment (1999) *Transport and the economy*. London: Stationery Office.

Sansom, T., C. Nash, P. Mackie, J. Shires and P. Watkiss (2001) *Surface transport costs and charges: Great Britain 1998*. Leeds: Institute for Transport Studies, University of Leeds.

Small, K.A. and C. Winston (1988) "Optimal highway durability," *American Economic Review*, 78:560–569.

Tinbergen, J. (1957) "The appraisal of road construction: two calculation schemes," *Review of Economics and Statistics*, 39:241–249.

Turro, M. (1999) *Going trans-European: planning and financing transport networks for Europe*. Oxford: Pergamon.

Vickerman, R.W. (1994) "Transport infrastructure and region building in the European Community," *Journal of Common Market Studies*, 32:1–24.

Vickerman, R.W. (2003) "Maintenance incentives under different infrastructure regimes," in: *Proceedings of the 2nd Workshop on Applied Infrastructure Research*. Berlin: Technische-Universität.

Walters, A.A. (1961) "The theory and measurement of marginal private and social costs of highway congestion," *Econometrica*, 29:676–699.

This Page Intentionally Left Blank

Chapter 14

INTEGRATED TRANSPORT STRATEGIES

ANTHONY D. MAY, CHARLOTTE KELLY and SIMON SHEPHERD
University of Leeds

1. Introduction

There has been growing interest in recent years in the development of integrated transport strategies. Their origins can be traced to a growing realization that a "predict and provide" approach was unlikely to provide a solution to growing transport problems (Goodwin et al., 1991), an acceptance that efforts to improve the supply of transport had to be matched by measures to control transport demand (Institution of Highways and Transportation, 1996), and heightened interest in the role of land use planning as a complement to transport policy (Greiving and Wegener, 2003). While several government agencies have recently advocated the use of integrated approaches (UK Department of Environment, Transport and the Regions, 1998; European Commission, 2001), the more visionary local authorities were appreciating the need for such approaches a decade earlier (May and Gardner, 1990; May, 1991; May and Roberts, 1995). However, there is still considerable confusion as to what is meant by integration, and how best it can be achieved.

In this chapter we provide a set of possible definitions, and then outline possible principles on the basis of which integration might be achieved. We then consider in greater detail two possible approaches: the pursuit of synergy and the removal of barriers. We illustrate these with a number of examples taken from predictive analyses. We then consider the role of sensitivity analysis as a means of identifying possible combinations of policy instruments. We conclude with broad guidance on the ways in which pairs of policy instruments might be combined to achieve integration. Our techniques for developing optimal transport strategies offer an analytical approach to the design of integrated strategies, and Chapter 39 on this topic therefore serves as a sequel to this chapter.

2. The meaning of integration

As noted above, several government publications have advocated an integrated approach, with the UK government going as far as to specify integration as an

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

objective of its transport policy (UK Department of Environment, Transport and the Regions, 1998, 2000). However, few policy documents define what they mean by the term. Much of the debate focuses on operational integration of fares and service levels specifically within public transport. There is, in practice, a more strategic form of integration, which is directly relevant to strategy formulation – the integration of policy instruments to achieve greater performance from the overall strategy. Such integration can occur in five broad ways:

(1) integration between policy instruments involving different modes;
(2) integration between policy instruments involving infrastructure provision, management, information, and pricing;
(3) integration between transport measures and land use planning measures;
(4) integration with other policy areas such as health and education;
(5) integration between authorities within a conurbation.

Integration of types 1 to 3 draws on the increasingly wide range of types of policy instrument available. One problem faced in developing such strategies is the lack of information on the performance of individual instruments. In our Web-based knowledge base, KonSULT (http://www.elseviersocialsciences.com/transport/konsult/index.html), we consider some 60 different types of policy instrument, grouped into six categories: land use; infrastructure provision; management and regulation; information provision; attitudinal change; and pricing. We assess the potential contribution of each to a range of policy objectives, in different contexts, both from first principles and on the basis of well-documented case studies. We also attempt, for each policy instrument, to identify those which might complement it by reinforcing its benefits (the pursuit of synergy), by reducing financial barriers, by reducing political and acceptability barriers, and by compensating losers. We return to this issue in the final section.

While a combination of policy instruments is likely to perform differently, against a given objective, from any one of the constituent instruments alone, the impact of the combination will depend on the types of instrument and the levels at which they are implemented. Much will depend on their ability to influence the underlying attributes of numbers of journeys, journey length, modal share, and, to a lesser extent, time of day and route taken. Some instruments will also change the transport supply, and hence the costs to users. Costs of implementation and operation, and revenues generated, will also be relevant to the impact of the instruments, alone and in combination. All of these will be affected by the scale and intensity with which a policy instrument is used; fare changes, for example, can vary in percentage change, by time of day, and, potentially, by route and area. The number of possible policy combinations is thus very extensive, and our parallel work on policy optimization (see Chapter 39) has been developed to assist in their analysis.

A carefully designed integrated strategy, particularly of types 1 to 3, should be better able to achieve the objectives set for it than any one or more policy instruments taken on their own. Some of the forms of integration outlined above may prompt a wider set of objectives; for example, integration of transport and land use (type 3) may well raise a wider set of development objectives, while integration with other policy areas (type 4) will require an understanding of their objectives (Jones et al., 2003). As with any strategy, it is important to be clear as to those objectives before the strategy is developed, since the combination of policy instruments suitable for, say, the pursuit of economic development will differ from those which best meet environmental or health targets. Where integration is to take place between authorities (type 5), it will be important for them to have a common understanding of their objectives, and of their relative importance. From this point of view, defining integration as an objective in its own right clouds rather than clarifies the issue. One needs to be clear what integration is designed to deliver, rather than seeing it as an end in itself.

3. Possible integration principles

Most approaches to strategic integration focus on one of two types of principle: the pursuit of synergy (May and Roberts, 1995) and the removal of barriers (May et al., 2003).

The pursuit of synergy involves finding pairs of policy instruments that reinforce one another. Obvious examples are the provision of park and ride to increase rail or bus patronage; the use of traffic calming to reinforce the benefits of building a bypass; the provision of public transport, or a fares reduction, to intensify the impact of traffic restraint; and the encouragement of new developments in conjunction with rail investment. However, true synergy is harder to achieve through these means, and it will help to define the term and its associated concepts more precisely.

The removal of barriers implies identifying factors that hinder the implementation of an otherwise desirable policy instrument, and using a second instrument to overcome them. Key barriers to any strategy will often be finance, public acceptability, and concerns that some members of society will be adversely affected. Thus, integration can contribute to the removal of barriers in three ways. Firstly, it can involve measures that make other elements of the strategy financially feasible. Parking charges, a fares increase, and road pricing revenue may all be seen as ways of providing finance for new infrastructure. Secondly, integration can package measures that are less palatable on their own with ones that demonstrate a clear benefit to those affected. Once again, an example is to be found in road pricing, which attitudinal research demonstrates is likely to be much more acceptable if the revenue is used to invest in public transport (Jones, 1998).

Thirdly, integration can involve measures that compensate losers. The selection of these depends on the side-effects that arise from other elements in the package. For example, road pricing could lead to extra traffic outside the charged area, which could be controlled by traffic management, and could adversely affect poorer residents, who could be helped by exemptions or concessionary fares. We consider the issue of barriers more fully below.

4. The concept of synergy

Policy instruments interact with each other in different ways. The term "synergy" is often used loosely to describe the effects of positive interactions between instruments. It is useful in practice to identify four terms that describe how the different instruments in policy packages combine with each other (Mayeres et al., 2003).

Complementarity

Complementarity exists when the use of two instruments gives greater total benefits than the use of either alone. This can be represented using the following notation:

$$\text{welfare gain } (A + B) > \text{welfare gain } A,$$

$$\text{welfare gain } (A + B) > \text{welfare gain } B.$$

Additivity

Additivity exists when the welfare gain from the use of two or more instruments in a policy package is equal to the sum of the welfare gain of using each in isolation. This can be represented as

$$\text{welfare gain } (A + B) = \text{welfare gain } A + \text{welfare gain } B.$$

Synergy

Synergy occurs when the simultaneous use of two or more instruments gives a greater benefit than the sum of the benefits of using either one of them alone:

$$\text{welfare gain } (A + B) > \text{welfare gain } A + \text{welfare gain } B.$$

Additivity and synergy can therefore be considered as two special cases of complementarity.

Perfect substitutability

Perfect substitutability exists when the use of one instrument eliminates entirely the welfare gain from using another instrument. This can be represented using the following notation:

welfare gain (A + B) = welfare gain A = welfare gain B

The term "welfare gain" is used here generically to reflect the full range of net benefits relevant to a given set of objectives. It can be derived from a conventional cost–benefit analysis, or from a more extensive multicriteria analysis, but should relate to the underlying objectives of the authority concerned. It is perhaps worth asking why, in these terms, anything other than complementarity should be achievable. In a simple system, it seems unlikely that the application of two changes that are mutually reinforcing should achieve more than the sum of the parts. For example, an increase in frequency and a reduction in fares on a single bus route are both likely to increase patronage, but both will to some extent attract the same users, and the increase from the two combined is likely to be less than that from the sum of their individual impacts. However, the transport–land use system in a city is not simple; interactions between modes and routes, lags in response, and feedback between transport and land use could all potentially result in discontinuities in the impact of policy instruments that could give rise to synergy.

5. The treatment of barriers

A barrier is an obstacle that prevents a given policy instrument being implemented, or limits the way in which it can be implemented. In the extreme, such barriers may lead to certain policy instruments being overlooked, and the resulting strategies being much less effective. Barriers can be grouped into four main categories (May et al., 2003):

- *Legal and institutional barriers*. These include lack of legal powers to implement a particular instrument, and legal responsibilities that are split between agencies, limiting the ability of the city authority to implement the affected instrument. A survey of European cities (May et al., 2001) indicates that road building, pricing, and land use are the policy areas most commonly subject to legal and institutional constraints. Information measures are substantially less constrained than other measures.
- *Financial barriers*. These include budget restrictions limiting the overall expenditure on the strategy, financial restrictions on specific instruments,

and limitations on the flexibility with which revenues can be used to finance the full range of instruments. Road building and public transport infrastructure are the two policy areas that are most commonly subject to financial constraints, with 80% of European cities stating that finance was a major barrier (May et al., 2001). Information provision is the least affected.

- *Political and cultural barriers*. These involve lack of political or public acceptance of an instrument, restrictions imposed by pressure groups, and cultural attributes, such as attitudes to enforcement, which influence the effectiveness of instruments. The European survey showed that road building and pricing are the two policy areas that are most commonly subject to constraints on political acceptability. Public transport operations and information provision are generally the least affected by acceptability constraints (May et al., 2001).
- *Practical and technological barriers*. While cities view legal, financial, and political barriers as the most serious that they face in implementing land use and transport policy instruments, there may also be practical limitations. For land use and infrastructure measures these may well include land acquisition. For management and pricing, enforcement and administration are key issues. For infrastructure, management and information systems, engineering design, and availability of technology may limit progress. Generally, lack of key skills and expertise can be a significant barrier to progress, and is aggravated by the rapid changes in the types of policy being considered.

Integrated strategies are particularly effective in overcoming the second and third of these types of barrier, and integration between authorities may help reduce institutional barriers as well. This apart, it is usually harder to overcome legal, institutional and technological barriers in the short term. It is often difficult to overcome a barrier without to some extent reducing the performance of the overall strategy. One which is feasible within a given financial constraint, or is modified to satisfy public opinion, is almost certainly less effective against the underlying objectives than one which is unconstrained in these ways. The pursuit of synergy and the resolution of barriers are thus to some extent in conflict in the design of integrated strategies.

In the short term, a strategy designed to overcome barriers may, however, ensure that something is implemented, rather than risking outright rejection of a better performing, but less acceptable, strategy. However, strategies should ideally be developed for implementation over a 15–20 year time-scale. Many of these barriers will not still apply 20 years hence, and action can be taken to remove others. For example, if new legislation would enable more effective instruments such as pricing to be implemented, it can be provided. If split responsibilities make achieving consensus impossible, new structures can be put in place. If finance for investment in new infrastructure is justified, the financial rules can be adjusted. Barriers should thus be treated as challenges to be overcome, not simply impediments to progress.

6. Some examples

There is as yet little empirical evidence of the benefits of integrated strategies, partly because the concept is sufficiently novel for there to be few implemented strategies, and partly because it would in any case be difficult to design a *post hoc* evaluation that successfully isolated the impacts of a combination of policy instruments. We therefore need to draw on analytical studies, most of which have focused on the pursuit of synergy. We review these first, and then look at an example of an attempt at the same time to overcome financial and acceptability barriers.

6.1. The London congestion charging study

The investigation of beneficial interactions between instruments was only a side issue in the research into congestion charging in London (May et al., 1996), as the emphasis was on the impacts of road user charging itself. Five additional complementary strategies (CS) were combined with alternative road user charging schemes to assess the potential combined impacts:

- CS 1. Bus priorities and traffic calming.
- CS 2. Improved rail frequencies.
- CS 3. Improved bus and rail frequencies together with bus priorities.
- CS 4. New rail infrastructure.
- CS 5. A combination of CS 3 and CS 4.

The study found that when adding congestion charges to each of the above strategies the highest additional economic benefit differed according to the congestion charge level. For example (Figure 1), at a congestion charge level of UK £2 for crossing the cordon the combination of congestion charging and bus priorities and traffic calming generated the most economic benefit. With the charge at UK £8, improved rail frequencies generated the largest economic benefit. Across all charges CS 5 generated the least extra economic benefit. In terms of environmental benefits the study found that at lower charge levels the reduction in emissions was three times greater with the combined strategy than without it. It also discovered, as shown in Figure 2, a largely linear relationship between economic benefits and emission reductions except at the highest traffic reduction levels, which implies that the two objectives can be met simultaneously.

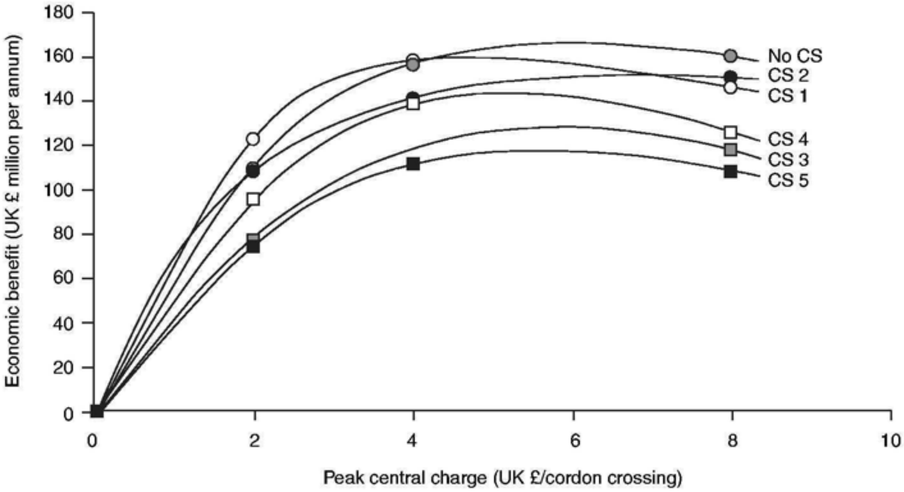


Figure 1. London cordon charging with complementary strategies. (Source: May et al., 1996.)

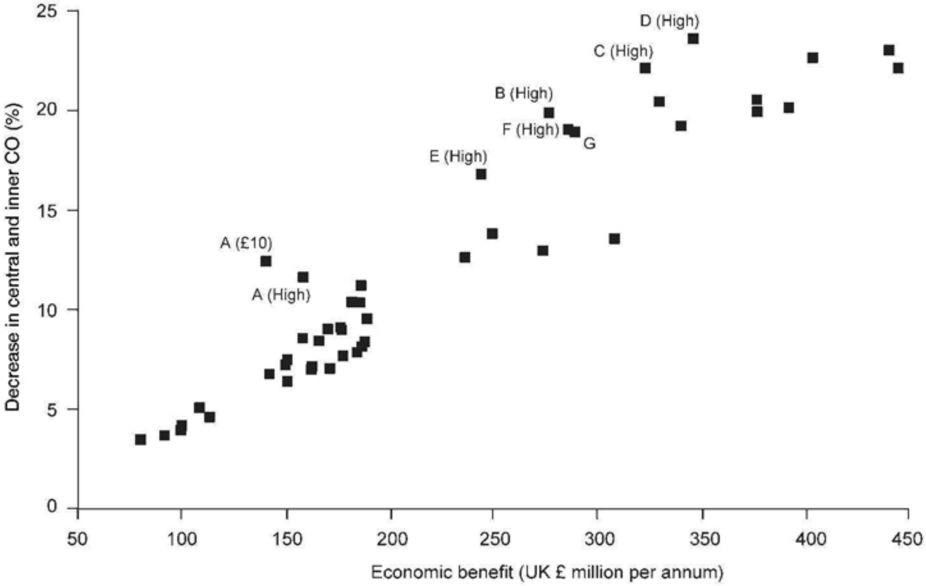


Figure 2. Relationship between economic benefit and pollution reduction for congestion charging options in London. (Source: May et al., 1996.)

Table 1
Evidence of synergy between increased car operating costs, faster public transport services and lower fares for Dortmund (Wegener, 2004)

		Difference from reference scenario in 2021 (%)						
		Trips	Mean trip length	Public transport trips	Car trips	Car-km per capita	Car owner-ship	CO_2 emission per capita
214	Car operating costs, +75%	–2.78	–14.77	+6.49	–3.61	–20.98	–6.24	–18.89
412	Public transport times, –5%	0.00	+0.02	+1.15	–0.06	–0.12	–0.05	–0.04
421	Public transport fares, –50%	+0.75	+2.49	+11.84	–0.42	–0.68	+1.95	+1.62
Total		–2.03	–12.26	+19.48	–4.09	–21.78	–4.34	–17.31
712	(214 + 412 + 421)	–2.00	–11.35	+26.68	–4.93	–23.03	–3.88	–17.43
Synergy				+7.20	–0.84	–1.25		–0.12

In general, the study showed evidence of complementarity between the instruments, but not of additivity or synergy.

6.2. *The PROPOLIS study of Dortmund*

The PROPOLIS study (Lautso, 2004) conducted a comparative study of the performance of a range of policy instruments, and selected combinations, in seven European cities, using three different land use transport interaction models. One of the case studies was Dortmund, for which the policy options were tested using the IRPUD model (Wegener, 1998). For this particular case study, a specific attempt was made to identify synergy (Wegener, 2004). PROPOLIS developed a comprehensive set of outcome indicators for the environmental, social, and economic dimensions of sustainability, and a weighted aggregation of these for each of the three dimensions; it also used a conventional set of transport statistics to monitor performance (Lautso, 2004). The Dortmund analysis of synergy, however, was limited to a set of transport statistics and one outcome indicator: carbon dioxide emissions, as shown in Table 1.

The combination illustrated in Table 1 involves increases in car operating costs by 75%, potentially through fuel taxes or distance-based charges; reducing public transport journey times by 5% through bus priorities and similar measures; and halving public transport fares. It shows significant synergy in its ability to attract trips to public transport, with an increase 35% higher than that from the sum of the constituent elements. There is also clear evidence of synergy for car trips and car-km. As a result there is a modest indication of synergy for carbon dioxide emissions, with the combined reduction just under 1% lower than that for the sum of the constituent elements. The question arises as to why synergy should appear in elements of the overall travel pattern, but not in performance against aggregate policy indicators. One possible answer is that these synergistic changes in, for example, public transport use are balanced by changes in other modes and elements of travel, within more stable aggregate values, such as travel time budgets. In turn, these aggregate constraints can limit the achievement of broader synergy.

6.3. The Edinburgh integrated strategy study

The Joint Authorities' Transport and Environment Study (JATES) was commissioned by the Scottish Office, Lothian Regional Council, and Edinburgh District Council in 1990 (May et al., 1992). The clients sought a range of possible strategies to meet the identified broad objectives of efficiency in the use of resources; accessibility within the city; environmental enhancement; safety; economic development; equity; and financial feasibility.

Four possible land use and development scenarios were specified, and for each the future problems were identified, assuming a "do-minimum" strategy. These results were then used to identify a number of policy instruments that could be used to overcome these problems. These in turn were packaged into three hypothetical strategies, focusing on highway improvements; rail and public transport improvements; and better management of the existing infrastructure.

Extensive sensitivity testing of individual instruments was then conducted to sort the possible policy instruments into three categories:

(1) those which were clearly beneficial;
(2) those which were not and could be rejected; and
(3) those whose range of impacts were both positive and negative, and merited further investigation.

Those in category 1 were principally low-cost means of managing the transport system more effectively; they included traffic management, signal coordination, bus priorities, and traffic calming. Those in category 2 were primarily infrastructure projects, and the majority of inherited infrastructure proposals were rejected as having costs that far exceeded their benefits. Those in category 3 were in many ways the most interesting. They included the remaining infrastructure projects, fares reductions, a major pedestrianization project, and a proposal for road pricing in the city center. Preferred strategies were then defined

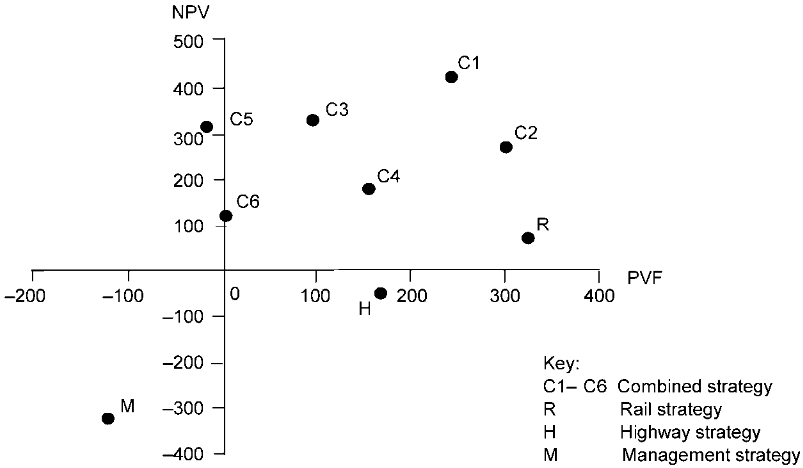


Figure 3. The economic and financial performance of the combined strategies for Edinburgh compared with initial strategies. Values are in UK £ million (1990 prices).

by combining the full set of instruments in category 1 with different combinations of those in category 3.

Since the finance available was uncertain, and road pricing was seen to be a particularly sensitive issue, six combined strategies were devised; two at each of three levels of financial outlay, one of each including road pricing. Other instruments included were those considered most likely to increase economic efficiency within the financial constraint. This approach offers an interesting application of both principles of integration: synergy between policy instruments is being sought, while at the same time an attempt is being made to overcome financial and acceptability barriers.

Figure 3 presents the results of an initial assessment of the six strategies in terms of the net present value of economic benefits (NPV) and the financial outlay (present value of finance (PVF), determined by discounting all cost and revenue streams to the present day). Higher values on the *y* axis are preferable in having higher economic benefit, but higher values on the *x* axis require greater financial outlay. The ideal position is thus toward the top left of the figure. The figure also shows the three initial hypothetical strategies: highway (H), rail (R), and management (M).

It can be seen that the strategies that included road pricing achieved higher levels of economic benefit at any level of finance than those that excluded it and that the benefits obtained were much less sensitive to the availability of finance with road pricing than without. Indeed, it appears possible to design a very effective strategy (C5) with no need for net financial support, provided that the revenues from road pricing can be hypothecated. It is clear that significant improvements in overall performance can be achieved by careful integration of policy instruments. Perhaps the most dramatic indication of this is the substantial improvement achieved in the recommended strategies (C1–C6) when compared with the initial hypothetical strategies (H, R, M). No attempt was made in this study to investigate the existence of synergy, but the results suggest that, while complementarity was found, true synergy was not present.

7. The application of sensitivity analysis

The approach adopted in the above examples of testing a range of combinations of policy instrument can be formalized through the application of sensitivity analysis. The strategic model MARS is used to represent the Edinburgh region in the UK. The case study is described in more detail in Chapter 39. The sensitivity approach adopted was as follows:

- A number of policy instruments were run in turn, at different levels over a feasible range, and the results were recorded in terms of welfare gains based on a standard cost–benefit analysis.
- The optimal level for each instrument was identified as that which maximized the cost–benefit analysis or objective function (termed OF) – this was either an internal optimum or a boundary value (the upper or lower limit of the practical range).
- Various combinations of policy instruments were then run with these optimal values to investigate the possible synergy effects.

The sensitivity tests were performed for each of the following instruments:

- peak fares (–50% to +100%);
- off-peak fares (–50% to +100%);
- peak frequencies (–50% to +100%);
- off-peak frequencies (–50% to +100%);
- peak road pricing charge to enter the city center (€2–6);
- off-peak road pricing charge (€2–6));
- parking charges in the city center long stay (€2–6));
- parking charges in the city center short stay (€2–6));
- road capacity changes peak (–10% to +5%);
- road capacity changes off-peak (–10% to 5%);
- fuel tax increases (–50% to +300%);
- fuel efficiency improvements (1 and 2% per annum).

The optimal, or practical, level for each instrument is shown in the first row of Table 2. The fares are set at their lower bound of –50%. The frequencies and peak road pricing variables are set at optimal values. The off-peak road charges are set at €2 – the lowest value tested, as any increase gives a negative objective function (OF) (the true optimum would be no off-peak charge). The parking charges are set to the same values as the road pricing variables to provide a better comparison even though these are not optimal for parking charges alone. Finally, the road capacity changes, fuel tax increase, and fuel efficiency are limited to what is thought to be practical. The other rows in Table 2 show the spatial coverage, OF value or welfare gain, and the PVF for each instrument.

From these results we can note that, in general, fares and fuel tax are the most effective instruments in increasing the value of the objective function. Also, peak instruments and area-wide applications are, in general, more effective than off-peak and spatially limited instruments. Table 2 also indicates the implications for financial constraints; public transport improvements require additional finance, while demand management measures can provide finance.

Table 3 shows the results for various combinations of the previous instrument levels given in Table 2. The first three rows show the sum of the OF values over the relevant single instrument results, the OF value when run in combination, and the implied synergy effect. Synergy is defined here as in Section 4, with a positive sign indicating synergy. The final row shows the PVF for the combined strategy.

The first strategy (Table 3, column 1) combines all instruments at their given levels. Note that this results in an OF value some €653 million lower than the sum of the individual effects. It shows that it is not a good idea to implement all instruments at their own optimal levels simply because they give improvements when applied individually (apart from the minor reduction due to off-peak road pricing). The second column in Table 3 shows the effect of removing the parking charge and road pricing instruments. Again, a negative synergy value is the result. However, the second strategy gives a better OF value than with the parking and road pricing charges included. This is because the charges overlap and are applied to the same element of generalized cost for some trips. Here we have identified that these instruments exhibit partial substitutability.

The third strategy removes fuel tax increases and looks at the effect of applying peak cordon and long-stay parking charges with the other instruments. Once again, the negative synergy effect implies that there is some obvious overlap between the cordon and parking charges, and suggests that if these were to be applied together, some other combination of charges would give better results. Here we can conclude that the cordon and parking charges also exhibit partial substitutability.

Columns 4 and 5 in Table 3 show the effects for fares, frequencies, and road capacity with fuel tax or peak road charging respectively; in each case there is only one charging instrument and no improvement in fuel efficiency. Again, there exists a small negative synergy, but as the overlap between instruments becomes less then the model exhibits additivity across instruments. Column 6 removes fare

Table 2
OF and PVF for optimal or practical levels for individual instruments

	Peak fare	Off-peak fare	Peak frequency	Off-peak frequency	Peak road pricing	Off-peak road pricing	Parking: long stay	Parking: short stay	Road capacity: peak	Road capacity: off-peak	Fuel tax	Fuel efficiency per annum
Change in instrument level	–50%	–50%	50%	25%	€5 m	€2 m	€5 m	€2 m	5%	5%	200%	1%
Spatial coverage	Area	Area	Area	Area	Central	Central	Central	Central	Area	Area	Area	Area
OF (€ millions)	1162.4	406.8	155.8	50.8	373.7	–67.3	172.3	9.3	548.3	912.1	1177.5	238.6
PVF (€ millions)	–1217.1	–1484.9	–367.0	–177.1	1151.3	698.8	169.3	55.1	73.4	155.0	10105.1	–553.2

Table 3
OF and PVF for combinations and possible synergy effects (€ millions)

	All instruments	Fares, frequencies, road capacity, fuel tax, fuel efficiency	Fares, frequencies, road capacity, road pricing peak, long-stay parking, fuel efficiency	Fares, frequencies, fuel tax, road capacity	Fares, frequencies, road capacity, road pricing peak	Frequencies, road capacity, road pricing peak	Frequencies, road capacity, road pricing peak, long-stay parking
Sum of individual OF	5140.3	4652.3	4020.8	4413.7	3609.9	2040.7	2213.0
OF of combination	4487.1	4561.5	3797.0	4372.2	3580.6	2050.2	2033.8
Possible synergy	–653.2	–90.8	–223.8	–41.5	–29.3	9.5	–179.2
PVF of combination	6450.0	4910.9	–2704.2	6263.4	–2233.8	828.4	874.9

changes from the combination in column 5. Here for the first time there is limited evidence of synergy (though perhaps it should be pure additivity, given possible convergence errors). Column 7 adds long-stay parking to this. Once again, there is negative synergy. It also confirms that combining the current road pricing cordon charges with the additional long-stay parking charges results in partial substitutability.

The synergy effects for the given combinations appear to imply that the model is largely additive across instruments, i.e. there is little interaction between them. Only when an instrument affects the same element of generalized cost, as is the case with the parking charges, cordon charges, and fuel tax for some car trips, is there any non-additive effect.

Regarding the PVF values, Table 3 shows that if fare reductions are to be implemented then these can only be afforded including the complementary instrument of fuel tax increases. Increased frequencies and improved capacities can be afforded by including the cordon charges and, as noted above, this combination exhibits a small amount of true synergy. In general, it is possible to use the OF and PVF values together to identify combinations that maximize performance within a given financial constraint. We return to more formalized ways of doing this in Chapter 39.

These model results have displayed the concepts of complementarity, additivity, and substitutability for a selection of instruments. Pure synergy effects were not found to be present with the given objective function and instruments tested, largely because many pairs of instruments appeared to be additive in their effects. However, it would be wrong to conclude that synergy could not be identified in such tests.

8. General design guidance

As noted earlier, one problem with the design of integrated strategies is the sheer number of different types of policy instrument that can be used. As a result, it is difficult to be certain how each pair of instruments will interact in general, let alone being able to predict their potential in any one context. In this chapter we have introduced the concepts of different types of integration, and of the potential for integration to achieve complementarity, additivity, and synergy, or to help overcome barriers of finance and acceptability. In our KonSULT website we attempt, for each policy instrument, to identify those other instruments that might complement it by reinforcing its benefits (the pursuit of complementarity or synergy), by reducing financial barriers, by reducing political and acceptability barriers, and by compensating losers.

The matrix in Figure 4, taken from our decision-makers' guidebook (May et al., 2003), provides a high level summary of these assessments. It suggests ways in

These instruments	contribute to these instruments in the ways shown					
	Land use	Infastructure	Management	Information	Attitudes	Pricing
Land use		✱				✱
Infrastructure	✱+		O			
Management	✱+	✱O+			✱	✱O+
Information	✱	✱O	✱O+		✱	✱O+
Attitudes	✱O	✱O	✱O			O
Pricing	✱+	✱□+	✱□+	□O	✱	

Key:
✱ Benefits reinforced
□ Financial barriers reduced
O Political barriers reduced
+ Compensation for losers

Figure 4. An integration matrix.

which the different types of policy instrument can complement those in other categories, through one or more of these approaches to integration. The matrix can clearly only serve as a broad design guide, but it may help to stimulate policy makers to consider a wider range of solutions to their transport problems.

9. Conclusions

Integration as a principle in urban transport policy is frequently advocated, but rarely defined. A distinction can be drawn between operational integration, usually of public transport, and strategic integration, between transport policy instruments with land use, with policy instruments in other sectors, and spatially between parts of a conurbation. Given the range of policy instruments and the scales at which they can be implemented, design of effective integration strategies is complex. Integration should be designed to serve agreed objectives rather than as an objective in its own right.

Most integrated strategies are developed either in pursuit of "synergy" or as a means of overcoming barriers, or both. Synergy as defined is a special case of complementarity, in which the benefits from the sum of the elements is greater than the sum of their individual benefits. It is debatable whether it can be realized in practice.

Financial and acceptability barriers, in particular, can be overcome by careful integration of different policy instruments, thus increasing the chance of the strategy being implemented. However, the overall benefit is likely to be less; thus, overcoming barriers is likely to be in conflict with pursuit of synergy.

The case studies investigated show little evidence of synergy in performance against objectives, though there is some evidence of synergy in responses within the transport system, such as trips by a given mode. It is not clear why this is. It may be that synergy is harder to achieve with a single objective, since the instruments that contribute to it will to some extent duplicate one another in their impacts. It may be that synergy becomes more apparent when objectives are in conflict, though much will then depend on the balance between these objectives.

It should be stressed that few of the objectives were developed specifically in pursuit of synergy. More research is needed to investigate the concept, and more examples of the analysis of integrated strategies would be welcome.

References

European Commission (2001) *European transport policy for 2010: time to decide*. Brussels, EC.

Goodwin, P.B., S. Hallett, F. Kenny and G. Stokes (1991) *Transport: the new realism. Report to Rees Jeffreys Road Fund*. Oxford: Transport Studies Unit.

Greiving, S. and M. Wegener (2003) "Integration of transport and land use planning: state of the art," in: *Proceedings of the 9th World Conference on Transport Research*. Amsterdam: Elsevier.

Institution of Highways and Transportation (1996) *Guidelines for developing urban transport strategies*. London: IHT.

Jones, P.M. (1998) "Urban road pricing: public acceptability and barriers to implementation," in: K.J. Button and E. Verhoef, eds, *Road pricing, traffic congestion and the environment*. Cheltenham: Edward Elgar.

Jones, P.M., K. Lucas and M. Whittles (2003) "Evaluating and implementing transport measures in a wider policy context: the Civilising Cities initiative," *Transport Policy*, 10:209–21.

Lautso, K. (2004) *PROPOLIS: planning and research of policies for land use and transport for increasing urban sustainability final report*. Brussels: European Commission.

May, A.D. (1991) "Integrated transport strategies: a new approach to urban transport policy formulation in the UK," *Transport Reviews*, 11:223–47.

May, A.D. and K.E. Gardner (1990) "Transport policy for London in 2001: the case for an integrated approach," *Transportation*, 16:257–77.

May, A.D. and M. Roberts (1995) "The design of integrated transport strategies," *Transport Policy*, 2:97–105.

May, A.D., M. Roberts and P. Mason (1992) "The development of transport strategies for Edinburgh," *Proceedings of the Institution of Civil Engineers: Transport*, 95:51–9.

May, A.D., D. Coombe and T. Travers (1996) "The London congestion charging programme: assessment of the impacts," *Traffic Engineering and Control*, 37:66–71.

May, A.D., B. Matthews and T. Jarvi-Nykanen (2001) "Decision making requirements for the formulation of sustainable urban land use transport strategies," in: *9th World Conference on Transport Research*. Seoul.

May A.D., A. Karlstrom, N. Marler, B. Matthews, H. Minken, A. Monzon, M. Page, P. Pfaffenbichler and S. Shepherd (2003) *Developing sustainable urban land use and transport strategies. A decision makers' guidebook*. Leeds: Institute for Transport Studies.

Mayeres. I., S. Proost, G. Emberger, S. Grant–Muller, C. Kelly and A.D. May (2003) *Deliverable D4: synergies and conflicts of transport packages. SPECTRUM (Study of Policies Regarding Economic Instruments Complementing Transport Regulation and the Understanding of Physical Measures)*. Leeds: Institute for Transport Studies.

UK Department of Environment, Transport and the Regions (1998) *A new deal for transport: better for everyone*. London: The Stationery Office.

UK Department of Environment, Transport and the Regions (2000) *Transport 2010: the 10 year plan*. London: DETR.

Wegener, M. (1998) The IRPUD model (http://irpud.raumplanung.uni-dortmund.de/irpud/pro/mod/mod e.htm).

Wegener. M. (2004) *Synergies between policies in policy packages. PROPOLIS technical note S7W10.* Dortmund: Spiekermann and Wegener.

Chapter 15

LONG-TERM PLANNING

EDWARD WEINER and ELIZABETH S. RIKLIN
US Department of Transportation, Washington, DC

1. Introduction

The goal of this chapter is to provide an overview of the transportation planning process, highlighting the considerations inherent in effective long-term planning. Long-term transportation planning assesses existing conditions, identifies the future demand for travel, generates alternative strategies to support the projected demand, evaluates those alternatives and recommends to decision-makers a practical portfolio of policies, programs, and investments that would maintain and improve the mobility of people and the movement of goods, while supporting the area's goals and objectives.[a] To be effective, a long-term transportation plan must be grounded in the community's vision, reflect early and frequent public involvement, have a comprehensive systems perspective, be developed with strong technical methods and the cooperation of the host of transportation agencies, and rest on realistic financial projections.

2. Background

In some ways, the long-term transportation planning process and planning techniques have changed little since the 1960s. Yet, in other ways, transportation planning has evolved over these years in response to changing issues, conditions, and values, and a greater understanding of urban transportation phenomena. Current transportation planning practice is considerably more sophisticated, complex, and costly than its highway planning predecessor, and involves a wider range of participants in the process.

Modifications in the transportation planning process took many years to evolve. As new concerns and issues arose, changes in planning techniques and processes

[a]For a history of the development of transportation planning in the USA see Weiner (1999)

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

were introduced. These modifications sought to make the planning process more responsive and sensitive to those areas of concern. Urban areas that had the resources and technical ability were the first to develop and adopt new concepts and techniques. These new ideas were diffused by various means, usually with the assistance of the national governments and professional organizations. The rate at which the new concepts were accepted varied from area to area. Consequently, the quality and depth of planning is highly variable at any point in time.

3. Overview of the transportation planning process

The process of identifying transportation problems and identifying effective solutions to those problems is called transportation planning. However, the planning process is more than merely listing highway and transit capital investments; it requires developing strategies for operating, managing, maintaining, and financing an area's transportation system in such a way as to advance the area's long-term goals. While there are universal elements of the transportation planning process, as depicted in Figure 1, each transportation plan, and the process to develop it, is unique, and should be tailored to best meet local circumstances and needs.

4. Visioning and transportation goals

Transportation planning is most effective when it establishes policy goals for the transportation system that are based on the area's long-term vision for itself:

> Effective transport investment policy must be derived primarily from an explicit view of the desired future city form. A simplistic approach of improving the transport infrastructure in order to solve specific transport problems, is usually ineffective, and sometimes counterproductive. Transport policy alone cannot be effective. It must be coordinated with other related policy decisions to form a coherent policy.
>
> (Gur, 1999)

The development of transportation policy goals and objectives is generally seen to require multidimensional participation that embraces at least:

- involvement of local officials and provision for broad stakeholder and modal involvement;
- tying in of the goals and objectives to action, and make them meaningful to stakeholders/customers;
- promotion of safe and efficient transportation movements;
- support of multimodal transportation uses;
- support for the land use and economic development policies of the city and region;

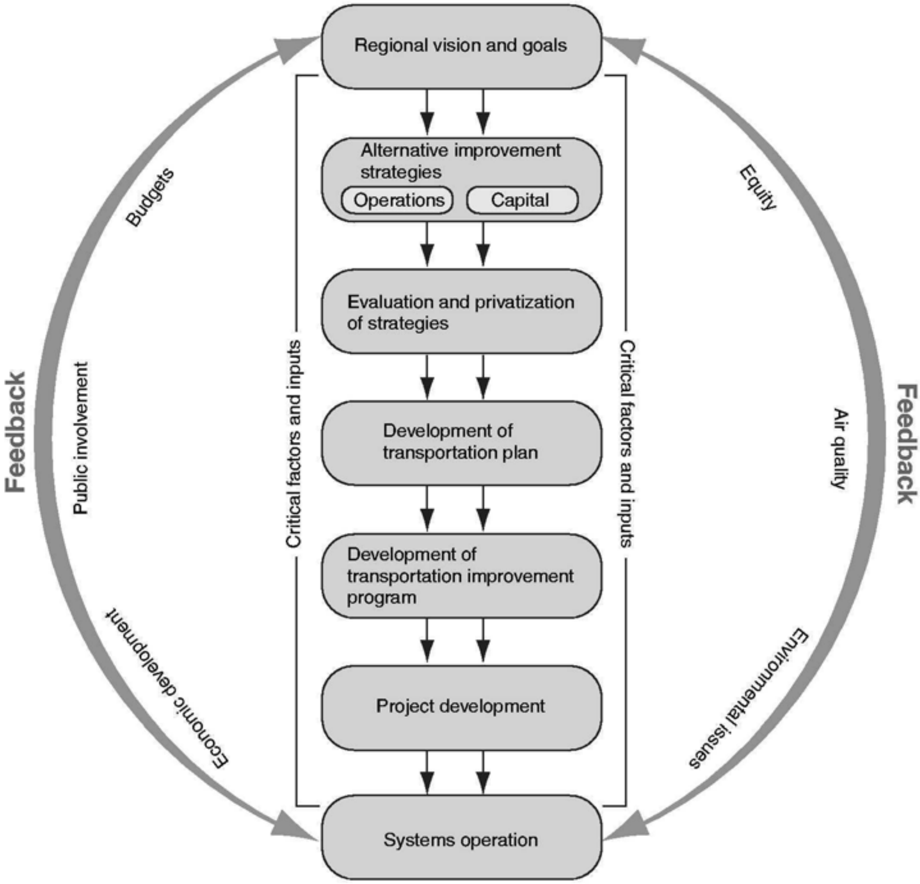


Figure 1. The long-term planning process (US Department of Transportation, 2001).

- sensitivity to environmental concerns;
- provision of sufficient specificity to guide plan development;
- coordination of local and regional policies with state or province plans;
- specifying roles in terms of who will be making the policy decisions and whether those roles are advisory or decision-making;
- inclusion of timelines and mechanisms for modifying and updating polices.

Thus, the transportation planner's first task is to work with the area's citizens and officials to understand their long-term vision for the area (TRANSPLUS, 2003). A vision plan provides broad policy goals for what the area will look like, and reflects what is important for the future. To develop a vision, planners, citizens, and officials need to consider several characteristics of their area and how they expect these characteristics to change over the next several years. Characteristics include projected population growth; projected economic changes; projected development patterns; preserving the human and natural environment; and enhancing the quality of life in the area. The vision's goals are the foundation for plans to improve an area's transportation system.

4.1. *Transportation goals*

Transportation goals and objectives that can be quantified and relate to the operational and physical condition of the transportation system are needed to evaluate the transportation system's current conditions as well as alternative management strategies and investment scenarios. With an understanding of the area's broad vision and an appreciation for how the transportation system can support that vision, transportation planners can begin to work with officials and citizens to develop policy goals for the transportation system. These goals provide the overall context within which the transportation system is operated, maintained, and developed. They need to be specific enough to guide the development of the plan but not too inflexible to respond to changing conditions and implementation priorities. Every area will have different priorities and, therefore, different goals and objectives. The important thing is that they are developed in a consultative manner, are measurable, and are used to guide plan development.

4.2. *Performance measures*

Performance measures can demonstrate how well the transportation system is doing its job of meeting public goals and/or expectations of the transportation network . Examples of performance measures are:

- *Accessibility.* Percentage of population within x minutes of y percent of employment sites; whether special populations such as the elderly are able to use transportation; whether transportation services provide access for underserved populations to employment sites; also whether services comply with legislation on accessibility by the disabled.
- *Mobility.* Average travel time from origin to destination; change in average travel time for specific origin–destination points; average trip length; percentage of trips per mode (known as mode split); time lost to congestion; transfer time between modes; percent on-time transit performance.

- *Economic development*. Jobs created and new housing starts in an area as a result of new transportation facilities; new businesses opening along major routes; percentage of region's unemployed who cite lack of transportation as the principal barrier; economic cost of lost time.
- *Environmental quality of life*. Environmental and resource consumption; tonnes of pollution generated; fuel consumption per vehicle-kilometer traveled.

Some methods used to measure performance include tracking average speeds and accident rates. Many areas monitor how close they are to achieving specific goals, such as the mobility of disadvantaged populations, levels of air quality, and the health of the economy. Performance measures provide feedback to the decision-making process – answering questions such as whether the performance of the transportation system (or economy, air quality, etc.) is getting better or worse over time and whether transportation investments are making a difference.

Through the development of the long-term transportation plan, the planning agency can take the lead in creating performance measures that provide information critical to regional and local decision-makers. Because performance measures strongly influence the goals and objectives of the planning process, their development and ongoing support can become part of the planning activities. If performance measures are to be developed, they should be subject to the public involvement program.

4.3. Land use and transportation

Land use and transportation are symbiotic: how development is designed and spaced can greatly influence an area's travel patterns, and, in turn, the degree of access provided by the transportation system can influence land use distribution. The role of the transportation planner in land use decisions varies (Hirschman and Henderson, 1990). In some areas, transportation planning agencies are responsible for reviewing local land use decisions considered regionally significant. In others, land use decisions are solely the prerogative of local officials. Regardless of the transportation agency's role in decision-making on land use, transportation planners must make every effort to consider the comprehensive land use plans of their area and maintain a constructive dialogue with land use officials. In that way, each group is informed of actions that might affect the other (US Department of Transportation, 2001). In addition, when major capital investments in transit facilities are being considered, transportation planners need to work with land use planners in the proposed transit corridors to encourage residential and commercial densities that would support the capital investment (American Association of State Highway and Transportation Officials, 1977; Cambridge Systematics et al., 1998).

4.4. Public participation

The fundamental objective of public participation programs is to assure that the concerns and issues of everyone with a stake in transportation decisions are identified and addressed in the development of the policies, programs, and projects being proposed in their areas. The elements of effective public involvement identified by the US Department of Transportation (2001) are:

- clearly defined purpose and objectives for initiating a public dialogue on transportation issues;
- specific identification of who are the affected public and other stakeholder groups with respect to the plans and programs under development;
- identification of techniques for engaging the public in the process;
- notification procedures that effectively target affected groups;
- education and assistance techniques, which result in an accurate and full public understanding of transportation issues;
- follow through demonstrating that decision-makers seriously considered public input.

The public includes anyone who resides, has an interest, or does business in a given area potentially affected by transportation decisions. This includes both individuals and organized groups. It is also important to provide opportunities for the participation of all private and public providers of transportation services, including, but not limited to, the trucking and rail freight industries, rail passenger industry, taxicab operators, and all transit and para-transit service operators. Finally, those persons traditionally underserved by existing transportation systems, such as low-income or minority households and the elderly, should be encouraged to participate in the transportation decision-making process. A well-informed public has the best chance to contribute meaningful input into transportation decisions, through a broad array of involvement opportunities at all stages of decision-making.

5. Information

Data on the current social and economic characteristics of the area, the area's expected development, the characteristics and performance of the transportation system, and the travel behavior of the households and business units is fundamental to the development of an effective long-term transportation plan (Fawcett, 1997; Douglas, 2000). Transportation planning agencies need to develop and maintain data to identify and understand transportation problems, generate and evaluate alternative solutions, and monitor the plan's performance. In addition to collecting their own data on land uses, development patterns, and local travel behavior, effective transportation planning agencies cooperate with land use planning entities, economic development, and transportation-operating and environmental agencies to maintain a robust database that serves them throughout the planning process.

Geographic information systems (GIS) for transportation have the capability to greatly improve the efficiency and quality of transportation planning and program development. GIS can be used in a wide range of activities, including collating data inputs to the travel forecasting model, performing spatial analyses such as measuring the jobs–housing balance, and displaying model outputs on base maps (Johnston and de la Barra, 2000).

In the USA, state departments of transportation (DOTs) are starting to use GIS to support a range of transportation decision-making processes. The North Carolina DOT has used GIS as a key part of a new approach, the "phased environmental approach," to improve its process for integrating environmental issues into the transportation systems planning process. In Southern California, planning for transportation facilities in a region that has nearly 200 threatened or endangered species is a major challenge for San Diego Association of Governments. Many conflicts arise from pressing demands to accommodate growth while preserving natural habitats. The region has launched a massive effort to complete multiple-jurisdiction, multiple-habitat, and multiple-species conservation programs. These programs are responsible for the development and management of extensive biological and land management databases. The GIS approach was chosen as the best method for maintaining and analyzing these data. By maintaining continuous and comprehensive habitat conservation programs, the region's transportation planners have access to timely and accurate environmental data (US Department of Transportation, 1996).

6. Identify transportation needs

6.1. Assess current conditions

A first step in needs analysis is to measure the gap between the transportation system goals and objectives and current conditions. Using transportation goals and objectives that are quantified and relate to the operational and physical condition of the transportation system, transportation agencies conduct a conditions analysis to measure the condition of the transportation system and the service provided for vehicles, transit, and non-motorized modes (bicycles and pedestrians). Service objectives for roadways may include consideration of roadway capacity, design, and safety. Examples of transit service standards are population coverage and frequency of service.

6.2. *Project future conditions*

Plans are future oriented. While they often respond to a backlog of needs identified through an assessment of current conditions, they also address future conditions and plan for them. There are a number of ways to estimate travel demand. These range from simple techniques such as historical trend analysis to variants of more complex computer models that require large databases of demographic and socioeconomic information to forecast travel demand.

Travel demand models

Travel demand models estimate future travel demand and indicate how the elements of the transportation system would function given future travel demand and alternative transportation investment scenarios. For the past 40 years, transportation professionals have used a four-step approach in modeling transportation demand. Most modeling approaches use some form of these steps today. Given an understanding of what the land use, population, and employment levels are in a study area, the model forecasts four basic travel elements: how many trips will be made; the origin and destination of those trips; the mode – single-occupancy vehicle, high-occupancy vehicle, transit, walk, bike etc. – that will be used; and the route that will be used. The basic components of travel demand models are calibrated to reflect current travel behavior, building upon observed travel behavior and existing development patterns. Once calibrated, the models project future travel demand on the basis of expected increases in population and economic activity and anticipated development patterns. The models provide information to identify the impacts of alternative investment scenarios on regional mobility and traffic congestion to inform the public and local decision makers (Gur, 1999). The output from travel demand models are also used to estimate mobile source emissions that would result from the transportation sector under the various investment strategies.

Activity-based travel models

A promising alternative to the four-step modeling process is the activity-based approach (Engelke, 1997; Cambridge Systematics, 2000). Entirely different from the approach taken for the development of the four-step procedure, the activity-based approach predicts travel demand based on an understanding of the decision process underlying travel behavior. Activity-based models are based on the fact that as the activities engaged in a day are linked to each other, trips made to pursue them are also linked to each other (Ettema and Timmermans, 1997).

As models are based on current relationships, they cannot anticipate the impacts on travel of long-term lifestyle changes such as the dramatic increase of women's participation in over the workforce or the current aging of the population. The models are also limited in their ability to indicate the impacts of technology on passenger and goods movement. Effective transportation planning agencies appreciate the limitations of travel demand models, and use them as one of an array of tools to identify and evaluate transportation solutions.

The movement of freight is often as important as person movements in an area's transportation system. The efficient movement of freight within and through a region is critically important to industry, retail, agriculture, international trade, and terminal operators. Freight movements especially affect metropolitan areas with ports, cargo airports, intermodal freight yards, large trucking terminals, and shipyards. Freight considerations can be incorporated systematically into the planning process by (US Department of Transportation, 2001):

- defining those elements of a metropolitan area's transportation system that are critical for efficient movement of freight;
- identifying ways to measure system performance in terms of freight movement;
- developing freight-oriented data collection and modeling to identify problems and potential solutions;
- creating a freight movement advisory committee to identify important bottlenecks in the freight network.

7. Develop and evaluate alternatives

7.1. Generate alternative strategies and actions to address the gaps

With an understanding of an area's long-term vision, travel patterns, and existing and projected gaps in the transportation system, transportation planners can propose a range of transportation improvements that would result in a balanced transportation network that would support the area's expected development patterns and allows the transportation system to achieve a level of service that promotes safe and efficient operation.

Management strategies

In addition to identifying conventional capital improvements to the highway and transit elements of the transportation system to meet personal mobility needs, planners, in consultation with the public, need to generate non-motorized alternatives and travel demand management (TDM)[a] and transportation system management and operation (M&O) strategies. M&O strategies recognize regional transportation as an interconnected set of services and systems, and work to improve transportation system performance through better management and use of the existing transportation network. Some examples of user-oriented performance measures are average trip travel time, length of delay, and reliability of trip making. These are important indicators of how well the transportation system is operating. Intelligent transportation systems (ITS) are computerized communication tools that can help to facilitate better system M&O. For example, roadway video surveillance allows better responses to changes in network conditions such as allowing the clearing of an accident faster to keep traffic moving. ITS technologies also can be used to collect real-time data, such as travel speeds, which can be used to monitor system performance over time. Successfully implementing M&O strategies requires close coordination among the many different agencies and groups with responsibility for transportation system performance.

7.2. Evaluate alternative strategies and actions to address the gaps

Transportation has a profound influence on the lives of people. Decision-makers must consider fully the social, economic, and environmental consequences of potential changes in the transportation system (Batey et al., 1992; Swenson and Dock, 2003). To inform decision-makers, planners assess the benefits (travel time savings and improved mobility), costs, and impacts of alternative strategies or improvements that address the condition of the transportation system. A coarse evaluation can be undertaken to consider the full range of alternative strategies and identify those meriting further consideration. These can be then subject to more detailed analysis. The evaluation of alternative strategies for addressing transportation system deficiencies is best considered as technical information that is an input into the policy and goal-setting process that selects transportation system strategies. In this way, the overall policy goals and strategies established in the plan drive the selection of appropriate strategies. In most areas, this decision-making takes place at a combination of policy and planning levels.

[a]TDM is any action or set of actions designed to influence the intensity, timing, and distribution of transportation demand, in order to reduce traffic congestion, enhance mobility, or improve the community environment. Such actions can include offering commuters alternative transportation modes and/or services, providing incentives to travel on these modes or at non-congested hours, providing opportunities to link or "chain" trips together, and/or incorporating growth management or traffic impact policies into local development decisions.

Cost estimates

Cost estimates are necessary to compare the proposed transportation investments with available revenues (ECONorthwest and Parsons Brinckerhoff Quade and Douglas, 1995; Nelson and Shakow, 1996). It is important to estimate maintenance and operation costs, in addition to capital costs, as these will likely use a majority of the existing revenue resources. Costs should, therefore, be estimated for:

- maintaining the existing and proposed transportation system;
- designing and building new, expanded, or replacement facilities (roads, terminals, etc.);
- acquiring new transit vehicles and related capital costs (maintenance facilities, etc.);
- operating transportation services such as transit or ride-sharing;
- administering and planning the transportation system.

Estimates of costs can usually be based on existing historic data that would be available from the finance officer of the city/county and the transit agency. Estimates of new costs for facilities and services will generally be based on a combination of rough estimates and specific cost estimates. Cost estimates based on preliminary engineering, right-of-way appraisals, or operating plans only need to be done for the most immediate recommended improvements.

Most of the recommended improvements in a long-term transportation plan will need an "order-of-magnitude" cost estimate. These estimates are based on factors such as typical "per kilometer" construction costs for different types of roadways or the operating costs for similar transit services in other counties.

Benefits

Travel time savings, safer transportation facilities, improved mobility and accessibility (Weibull, 1980; Handy and Niemeyer, 1997), and reduced congestion are just a few of the benefits that can result from a transportation investment or strategy. Using performance measures identified early in the planning process, planners can indicate and compare how well transportation investments or strategies achieve the area's transportation goals and objectives and support the area's vision.

Incorporating environmental review

It is productive to have a general understanding of the region's environment, as well as environmental regulations and requirements, when developing transportation plans (Swenson and Dock, 2003). It is also productive in the early stage of plan development to compile existing information on the study area environment as part of the base on which the proposed regional transportation improvements will be superimposed. This will, at a broad level, allow planners to understand the likelihood of any potential adverse impacts associated with construction. Areas of concern include the potential impact on air quality, land use, noise levels, water quality, wetlands, flood plains, threatened and endangered wildlife, historical and archaeological sites, and hazardous materials sites. Where possible, identification of sites where proposed transportation improvements may potentially impact the environment or are presumed to be environmentally sensitive should be highlighted for more detailed analysis. In the United States, only when a proposed transportation project can be shown to not adversely affect the environment, or have its impact avoided, minimized or mitigated, can a transportation project advance into the construction phase.

Air quality issues play a major role in transportation planning. Air pollution is caused by the interaction of topography, weather, and human influences on the environment, such as manufacturing, use of petroleum-based products such as gasoline, and even small business activities, such as dry cleaning. The key transportation-related pollutants are ozone precursors, carbon monoxide (CO), and fine particulates – particles smaller than 10 μm, which are more likely to lodge in human lungs than larger particles. The ozone precursors are pollutants that combine to form ground-level ozone, which in turn is part of smog. Ozone precursors are volatile organic compounds (VOCs) and nitrogen oxides (NO_x). These pollutants all emanate in part from on-road mobile sources.

In the USA, the Clean Air Act (CAA) of 1990 and the Transportation Equity Act for the 21st Century (TEA-21) both require that transportation and air quality planning be integrated in areas designated by the US Environmental Protection Agency as air quality non-attainment or maintenance areas. Non-attainment areas are geographic areas that do not meet the federal air quality standards, and maintenance areas are areas that formerly violated but currently meet the federal air quality standards. If no violations of air quality standards have been found, the area is considered to be in compliance or attainment with federal air quality standards. The CAA identifies the actions that areas must take to reduce emissions from on-road mobile sources in non-attainment and maintenance areas. In the USA, the challenge for transportation decision-makers in non-attainment and maintenance areas is to decide on a mix of transit and highway investments that, combined with measures such as inspection and maintenance programs or reformulated gasoline, will keep emissions within the allowable limits for emissions from motor vehicles.

7.3. Identify the distribution of costs, benefits, and impacts

In addition to assessing the benefits, costs, and impacts of alternative investments and strategies to improve personal mobility and goods movement, transportation planners have the responsibility to identify the distribution of those benefits, costs, and impacts on the various segments of the population, especially underserved and underrepresented population groups.

8. Prepare a long-term plan

8.1. Long-term plan document

Key success factors for developing plans systematically include the following:

- have clearly established roles and responsibilities for who will develop the plan, how and when it will be adopted, and how and when the plan can be amended;
- use the planning team and the public consultation process to help develop the outline for the plan;
- ensure that the plan is a strategic and visionary document and not a project list or "wish list";
- during the planning process, document all technical data and methodologies as well as cited references and reports.

8.2. Set priorities

Since transportation needs typically outweigh expected revenues, it is important to prioritize the needs identified during the transportation planning process. Typically, there are an overwhelming number of potential improvements, so it is important that the planning process has an agreed upon approach to project prioritization. To set priorities successfully, it is necessary to establish formal prioritization criteria and apply it consistently to all programs and projects.

Transportation asset management

One strategic framework for making cost-effective decisions about allocating resources and managing infrastructure is transportation asset management. It is based on a process of monitoring the physical condition of assets, predicting deterioration over time, and providing information on how to invest in order to maintain or enhance the performance of assets over their useful life. One of the main goals of transportation agencies is preservation, to keep the infrastructure in operating condition. If roads, bridges, airports, transit facilities, ports, bicycle and pedestrian paths, etc. are not maintained, people and goods will not move as easily, resulting in reduced quality of life and diminished economic activity. Through an effective transportation planning process, transportation planners can support asset management for establishing priorities for improving the area's transportation assets. Decision-makers can use the transportation asset management process to set long-term priorities by considering the following (McNeil, 2003):

- What is our inventory of assets?
- What is the value of our assets (monetary, importance to region, other)? What are their functions? What services do they provide?
- What are the past, current, and predicted future condition and performance of our assets?
- How can we preserve, maintain, or improve our assets to ensure maximum useful life and provide acceptable service to the public?
- What are our choices for investing our transportation budget? What are the costs and benefits of such choices?
- What are the consequences of not maintaining our assets? How can we communicate those consequences?

8.3. Establish financial plan

The transportation plan needs to be realistic, and usually that means fundable. Without tying transportation projects to reliable funding, the recommended solutions that are developed can easily become a "wish list." However, limiting solutions to projects that do not exceed available revenue could result in providing a lower level of service than the community desires. The funding plan should be a multi-year financing plan based on the needs identified in the plan, and it should include an analysis of the participating jurisdiction's financing capabilities. To develop the transportation finance analysis, the following steps can be used:

- development of cost estimates for solutions;
- assessment of the ability to pay for these projects and services;
- development of financing policies;
- forecasting revenue from existing and potential sources;
- development of a financing schedule by matching transportation projects and services to revenue projections;

- establishment of policies to govern the management of the transportation financing program.

If probable funding falls short of meeting identified needs, the funding plan should contain a discussion of how additional funding will be raised, or how assumptions will be reassessed to ensure that level of service standards will be met or adjusted.

8.4. Transportation improvement program

Programming refers to a series of activities carried out by planners, including data assessment, appraisal of identified needs, and consideration of available or anticipated fiscal resources to result in the drawing up, scheduling, and planning of a list of identified transportation improvements for a given period of time. The programming of projects for funding should consider:

- timing of the need for improvements (i.e. when the facility falls below the locally established level of service under assumed growth rates);
- timing for fund availability.

Often plans will require more funds than are available. This means that the agencies engaged in planning should identify funding mechanisms to support implementation of the transportation plan or reassess their desired levels of service.

9. Monitoring and evaluation

For a plan to be successful it must be implemented effectively, and progress against plan objectives monitored. This provides the "feedback loop." Transportation planning includes continually monitoring the performance of the transportation system and ensuring that plans are being implemented to meet the intended objectives. The success factors for implementation and monitoring of the transportation plan include:

- developing an ongoing process, known to participants, for monitoring progress toward plan objectives;
- establishing a process for how decisions regarding implementation are to be made;
- establishing a process for conditions tracking system;
- establishing a well-defined process for how priorities will be set.

Many transportation plans have failed because they lacked an effective implementation plan and monitoring mechanism. These are required to "keep the plan alive," and ensure that the plan guides and shapes transportation decisions in the future.

10. Conclusion

Long-term transportation planning continues to evolve to provide better information to decision-makers on how to increase mobility and accessibility while minimizing the negative environmental (Swenson and Dock, 2003), equity (Forkenbrock and Schweitzer, 1997), economic, and health impacts of travel. Transportation planners are recognizing the need to increase their cooperation and coordination with land use, health and environmental agencies, and decision-makers, recognizing the links between transportation decisions and land use and economic development. To maximize the use of existing transportation facilities, transportation planners are working with their communities to develop effective performance measures, sharpening their focus on intermodal connections and intelligent transportation systems, developing methods that can analyze and predict the impact of TDM and system management techniques on the functioning of transportation systems and preserving existing transportation assets. And to further the development of sustainable transportation systems, transportation planners are considering all modes of travel – transit and non-motorized modes, as well as the auto. The development of long-term transportation plans is an increasingly complex effort to identify effective solutions to transportation problems.

Appendix

A typical general transportation plan (US Department of Transportation, 2001):

Executive summary

This section provides an overall summary of the plan's objectives, methodology, findings, and recommendations.

Section I. Goals and policy statements

This section presents the overall vision, goals, and objectives developed during the planning process. These form the overall umbrella for the direction of the transportation plan in terms of plan priorities.

Section II. Transportation element

Chapter 1. Introduction. This chapter outlines the purpose of the plan, the plan participants, and the organization of the document.

Chapter 2. Existing conditions. This chapter presents the existing condition of the transportation system in terms of: roadways (road and bridge conditions, traffic volumes, safety, other criteria); public or quasi-public transportation (transit, school bus, emergency service routes and facilities, air, and water); non-motorized transportation (bicycle pathways, pedestrian pathways, equestrian routes); and land use and population considerations, the plans and programs of other agencies and jurisdictions, and county-wide policies.

Chapter 3. Travel forecast. This chapter presents historical traffic trends, population and land use trends, population and demographic projections, population distribution, a future land use map, and future travel projections and trends.

Chapter 4. Alternative strategies evaluation. This chapter presents the determination of needs based upon existing conditions and travel. It presents the evaluation of alternatives for travel safety, levels of service and congestion, environmental impacts, financing, community support, and consistency with the plans of other agencies and jurisdictions.

Chapter 5. Priorities and recommendations. This chapter presents prioritized recommendations for improvements to the area transportation system including: levels of service, new corridors, road widenings, spot/intersection widenings, realignments or channelization, traffic control or signalization, shoulder improvements, paving, bridge replacements, or other physical improvements, pedestrian, bicycle, or equestrian improvements, transit and transit facilities, and land use/transportation linkages.

Chapter 6. The financing element of the plan. This chapter presents cost estimates for identified improvements, potential financing options, re-assessment of identified improvements based upon financial constraints, and the 3 year transportation improvement program for the area.

Chapter 7. Implementation and monitoring. This chapter provides the plan for continually monitoring the performance of the transportation system to determine the progress being made in improving system performance and to identify additional areas of improvement.

Section III: Appendices

A. References

B. Technical data and methodologies

C. Excerpts from other reports

References

American Association of State Highway and Transportation Officials (1977) *A manual on user benefit analysis of highway and bus-transit improvements*. Washington, DC: American Association of State Highway and Transportation Officials.

Batey, P., M. Madden and G. Scholefield (1992) "Socioeconomic impact assessment of large-scale projects using input–output analysis: a case study of an airport," *Regional Studies*, 27:179–192.

Cambridge Systematics (2000) *Development and application of a microsimulation activity-based model for San Francisco*. Oakland: Cambridge Systematics.

Cambridge Systematics, R. Cervero and D. Aschauer (1998) *Economic impact analysis of transit investments: guidebook for practitioners. Transit cooperative research program report 35*. Washington, DC: National Academy Press.

Douglas, B. (2000) *Data collection and modeling requirements for assessing transportation impacts of micro-scale design*. Prepared for the Travel Model Improvement Program (http://www.gisdevelopment.net/application/Utility/transport/index.htm).

ECONorthwest and Parsons Brinckerhoff Quade and Douglas (1995) *Evaluation of transportation alternatives: least-cost planning: principles, applications, and issues.*. Publication No. FHWA-PD-95-038. Washington, DC: Federal Highway Administration.

Engelke, L.J., ed. (1997) *Activity-based travel forecasting conference proceedings: summary, recommendations and compendium of papers*. Washington, DC: US Department of Transportation.

Ettema, D. and H. Timmermans, eds (1997) *Activity based approaches to travel analysis*. Oxford: Elsevier.

Fawcett, J. (1997) *Multimodal transportation planning data: final report*. Washington, DC: Transportation Research Board National Cooperative Highway Research Program.

Forkenbrock, D.J. and L.A. Schweitzer (1997) *Environmental justice and transportation investment policy*. Iowa: Public Policy Center, University of Iowa.

Gur, Y.J. (1999) "Urban form and transportation: degrees of freedom and interaction," in: *International Union of Public Transport 53rd UITP International Congress Conference on An Urban and Congestion Free 21st Century*. Toronto.

Handy, S.L. and D.A. Niemeyer (1997) "Measuring accessibility: an exploration of issues and alternatives," *Environment and Planning A*, 29:1175–94.

Hirschman, I. and M. Henderson (1990) "Methodology for assessing local land use impacts of highways," *Transportation Research Record*, 1274:35–40.

Johnston, R.A. and T. de la Barra (2000) "Comprehensive regional modeling for long-range planning: linking integrated urban models and geographic information systems," *Transportation Research A*, 34:125–136.

McNeil, S. (2003) *Transportation asset management bibliography*. Washington, DC: American Association of State and Highway Transportation Officials (http://assetmanagement.transportation.org/tam/aashto.nsf/home).

Nelson, D. and D. Shakow (1996) "Least-cost planning: a tool for metropolitan transportation decision-making," *Transportation Research Record*, 1499.

Swenson, C.J. and F.C. Dock (2003) *Urban design, transportation, environment and urban growth: transit-supportive urban design impacts on suburban land use and transportation planning*. Minneapolis: University of Minnesota.

TRANSPLUS (2003) *Promoting the integration of citizens and stakeholders in urban decision making. Deliverable D5*. Rome: ISIS (http://www.transplus.net).

US Department of Transportation (1996) *Community impact assessment: a quick reference for transportation*, FHWA-PD-96-036. Washington, DC: Federal Highway Administration.

US Department of Transportation (2001) *The metropolitan transportation planning process: key issues, a briefing notebook for metropolitan planning organization board members*. Washington, DC: Federal Highway Administration (http://www.planning.dot.gov/documents/BriefingBook/BBook.htm).

Weibull, J.W. (1980) "On the numerical measurement of accessibility," *Environment and Planning*, 12:53–67.

Weiner, E. (1999) *Urban transportation planning in the United States. An historical overview*, revised edn. Westport: Praeger.

Part 3

ASSET MANAGEMENT AND FUNDING

This Page Intentionally Left Blank

Chapter 16

TRANSPORTATION ASSET MANAGEMENT

ODD J. STALEBRINK
West Virginia University, Morgantown, WV

JONATHAN L. GIFFORD
George Mason University, Arlington, VA

1. Introduction

Transportation asset management (TAM) is an umbrella term used to describe a range of efforts aimed at systematically and cost-effectively maintaining, upgrading, and operating surface transportation infrastructure assets. TAM goes beyond the traditional engineering-oriented approach to managing these assets by introducing or expanding the role of economic theory and business-oriented practices in the delivery of transportation services. As such, TAM introduces a new set of tools and practices that seek to build and improve upon existing practices, rather than to replace them. This chapter provides an overview of the state of the art of these practices and tools, and their development.

2. Key functional areas of transportation asset management systems

TAM systems generally support four key functional areas (Switzer and McNeil, 2004):

(1) policy goals and objectives and performance measures;
(2) planning and programming;
(3) program delivery;
(4) systems monitoring and performance results.

Policy goals and objectives are established through consideration of a number factors, including the economic and social impacts of transportation infrastructure-related decisions; other programs that are competing for scarce public resources; the life cycle of transportation assets (i.e. life-cycle costs); the preservation of accountability; and orientation to customer needs. Hence, a TAM system takes an

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

enterprise-wide approach when setting policy goals and objectives at the same time as it emphasizes the use of economic theory.

Planning and programming[a] is policy driven and performance based, with a long-term view of facility condition, performance and costs. A TAM system should be able to provide relevant and reliable outcome and output measurements that aid in the analysis and identification of "red flags," and provide input to the establishment of new or improved policy goals and objectives. One of the most heavily emphasized aspects of TAM in relation to programming and planning has been to capture synergies from managing transportation assets across functions. Traditionally, pavement management systems, bridge management systems, and maintenance management systems have often been managed in isolation.

Consistent with its enterprise-wide emphasis, TAM facilitates the integration of these systems by allowing managers to pull information from a single database. The desired contribution of this database is that it would allow for consideration of alternative investments and courses of action based on relative merit, such as condition-based budgeting. In addition to serving planning and programming activities, this central database would also be critical in fulfilling two remaining functions – program delivery and system performance monitoring. The central database could thus help remove inefficiencies in program delivery and provide new ways to hold management accountable for work accomplished. Given its uses, the database is one of the most critical elements of TAM.

While a central database may be a critical facilitator of TAM, a major challenge for agencies considering the implementation of TAM is determining what data to collect, and how. To date, many TAM efforts have therefore been geared toward helping transportation agencies identify and assess the new data that would facilitate the operation and management of its resources. "New data" in the context of TAM refers to data that allow transport agencies to receive objective feedback on facility condition, performance, and cost at an enterprise-wide level.

3. TAM benefits

In order to understand the role and benefits of TAM in decision-making about physical infrastructure it is important to underscore that the functional areas mentioned above produce benefits at different decision levels: the transportation agency, constituents and end-users, and legislative and oversight bodies. TAM is most often discussed as a tool for enhancing the operational effectiveness of transportation agencies. Not surprisingly, enhanced cost-effectiveness and efficiency in the delivery of transport services are driving forces underlying the implementation

[a]Programming refers to the selection, sequencing, funding, and scheduling of projects.

of TAM. However, agencies also have stakes in ensuring accountability. For example, capital markets may reward transparency with increased access to capital at lower rates. TAM agency activities that have been identified to serve efficiency, effectiveness and accountability goals include (Winsor et al., 2004):

- creating and consistently updating an inventory of assets operated by the agency, including some measures of condition, use, and performance;
- changing methods of asset valuation to recognize user costs and benefits;
- fully reporting agency inventory and expenditure decisions;
- implementing maintenance and capital improvement activities based on evaluation of user costs and benefits, optimization of investments, and/or maintaining a certain level of overall system performance;
- building an overarching system that allows the comparison of investments across different asset groups

For legislative and oversight bodies, TAM promises to illustrate more accurately the full ramifications of making or postponing decisions about the physical infrastructure stock, and, thus, would serve to inform oversight and the appropriation process. For example, the life-cycle cost analysis focus in TAM would make explicit the "full" financial obligations that would arise from decisions to add to, preserve or maintain the infrastructure stock. Previously, no comprehensive effort has been made to systematically provide information on the full financial implications of these decisions.

Finally, TAM promises to enhance the transparency of the activities carried out by transportation agencies to citizens and to legislative and oversight bodies. Under TAM, transparency arises from standardizing and simplifying the provision of information, as well as by expanding the focus of accountability from the traditional emphasis on fiscal matters to include operations and full costs.

4. Transportation asset management systems tools

A critical gap exists between the needs and the availability of TAM tools and applications. This gap significantly hampers transportation agencies' efforts to incorporate TAM into system operation and management (Switzer and McNeil, 2004). Many of the traditional tools are insufficient to support the capability of TAM for increased system integration, and for trade-off and life-cycle analysis capability. The development of good performance-based applications is perhaps the most critical challenge to TAM. There is a great need for identifying best practices and benchmarks on performance measurement regarding TAM. Research is ongoing to develop a number of tools and applications to facilitate the needs expressed for TAM systems. Areas where performance-based applications are needed to support TAM include (Neumann and Markow, 2004):

- inventory applications that allow for identification of type, number, and location of infrastructure assets;
- condition applications that provide up-to-date measures of condition, service level, and performance of infrastructure assets;
- forecasting applications that provide assessment of changes in condition, service level, and performance over time;
- applications that provide estimates of cost of treatment of a facility;
- impact models forecasting the implication of investments.

Overall, the advancement of TAM, in both functional areas and tools, varies quite substantially internationally. In the USA, much of the effort still focuses on building a support base and momentum around the major ideas of TAM (i.e. integration, customer orientation, life-cycle analysis). The development of tools in the above areas is the greatest need. The applications are necessary to illustrate the benefits of TAM. The support and willingness to implement TAM is partly contingent upon such illustrations. Internationally, especially in many Commonwealth countries, practice has already advanced.

5. TAM development in Commonwealth countries

Interest in TAM arises from a number of factors, including changes in the transportation environment; significant changes in public expectations; and extraordinary advances in technology. Internationally, TAM efforts have been underway for over a decade, and the concept is now receiving wide acceptance in the global community. In particular, significant progress has been made in many of the Commonwealth countries, including Australia, the UK, and Canada. These countries undertook systematic asset management efforts in the mid-1980s and the early 1990s. The efforts were largely driven by "new public management reforms" and changes in public sector financial accounting and reporting practices.

The effect of changes in public sector financial accounting and reporting practices on TAM is exemplified by the Australian experience. In Australia, TAM emerged out of a nationwide financial accounting reform that culminated in 1993 with the release of two Australian accounting standards: Australian Accounting Standard 27 – Financial Reporting for Local Governments (AAS27), and Australian Accounting Standard 29 – Financial Reporting for Government Departments (AAS29). These standards represented a significant move toward the adoption of enterprise-based financial accounting practices by Australian governments. As such, they extended the use of public financial accounting information in public management from a focus on fiscal accountability to include operational accountability by providing information on productivity, cost effectiveness, and quality of public services (International Federation of

Accountants, 1997). The influence of AAS27 and AAS29 were made explicit in the release of the first comprehensive Australian TAM manual, released by the Institute of Public Works Engineering Australia in 1995. The manual relies heavily on the information made available by enterprise-based financial accounting reports (Barret, 1996).

The influence of "new public management reforms" on TAM development is exemplified by the UK case. In the UK, asset management efforts began in the mid-1980s, triggered by aggressive cutbacks in the public sector by the government. These cutbacks provided a justification for drastic changes in the public sector, and thus an opportunity to implement asset management systems. Similar to Australia, the development of asset management efforts in the UK occurred in conjunction with a reform of governmental accounting practices.

6. TAM development in the USA

In the USA, TAM has emerged more recently. Its development coincides with the completion of the interstate highway system, which expanded the central focus of transportation service delivery beyond construction toward operation, management, preservation, and reconstruction of the system, as a means of getting the most capacity out of existing facilities (US Federal Highway Administration, 2002). This expansion of focus represents a fundamental change for many transportation agencies, and it requires a new way of thinking about transportation service delivery. Economic theory and business-oriented practices play an increasingly important role in adjusting to this new environment.

Practice among the leading states in the USA reflects a very narrow approach to TAM; most efforts have been geared toward creating better inventories and collecting data. A more comprehensive approach that encompasses processes that allow for the comparison of investments across different asset groups may follow as the adoption of asset management principles broadens.

Two things set US TAM efforts apart from the experience in the Commonwealth countries mentioned above. The first is the relative infancy of the concept in the context of US transportation service delivery. The term has been attracting broad interest since the mid-1990s. Previously, it was generally practiced at an individual agency level, with little sharing of successful asset management practices across agency boundaries. As a result, TAM remains in an embryonic state in most US transportation service delivery.

Second, and perhaps more importantly, is the highly decentralized nature of transportation system ownership and operation in the USA. A large number of stakeholders – each with different agendas, needs, perspectives, and expectations – makes the development of TAM in the USA particularly challenging (Switzer and McNeil, 2004). In essence, TAM practices are emerging from the interactions between a diverse set of interest groups, including the three levels of government, industry groups, and commercial interests (Stalebrink and Gifford, 2002).

Government ownership and operational authority over the highway system in the USA resides primarily with the states, which is a reflection of its federal nature. The federal government provides funding and sets some high-level policy by imposing conditions on funding eligibility. The states generally own and operate statewide highway systems, and local governments own and operate some local highways and most roads and streets.

The commitment to TAM at the federal level has arisen, to a large extent, from the national surface transportation legislation enacted in 1991 and 1998 (Intermodal Surface Transportation Efficiency Act of 1991 and Transportation Equity Act for the 21st Century 1998). These laws significantly changed the role of the Federal Highway Administration (FHWA) in the state transportation programs. In essence, they increased the states' flexibility in deciding where and how to invest federal transportation dollars. They also reduced the FHWA's role in individual project decisions. At the federal level, TAM may therefore be viewed as an important tool for holding states operationally accountable for the use of federal funds (Clash and Delaney, 2000), or alternatively as a means of improving accountability to compensate for a reduced federal role in individual transportation project decisions. The federal government influence on TAM is officially carried out by the Office of Asset Management of the FHWA, which is part of the US Department of Transportation.

The states' involvement in TAM has been supported by the American Association of State Highway and Transportation Officials (AASHTO). Among its important functions is representing its members in national policy-making. The AASHTO has taken a leading position as a coordinator for the advancement of TAM as it relates to state transportation agencies. It works closely with the FHWA's Office of Asset Management to advance the application of TAM. The forum for this partnership is the AASHTO Taskforce on Asset Management, which was established in 1997 (McElroy, 2000). The AASHTO has played a key role in efforts to develop a TAM strategic plan and a US TAM guide (Cambridge Systematics, 2002). The AASHTO guide is the first formal US guide on TAM. It allows individual transportation agencies to define and assess the benefits of transportation asset management in the context of their own unique circumstances. Rather than providing for an entirely new framework, the document seeks to maximize the benefits of the "old" at the same time as it tries to integrate the "new" (Kadlec and McNeil, 2001). Both the AASHTO and the FHWA have indicated their commitment to assist in efforts by states to adopt TAM practices based on the guide.

Local government TAM activities are coordinated through the American Public Works Association (APWA). To support TAM implementation at the local level, the APWA coordinates its efforts through a taskforce on asset management. The APWA task force supports the implementation of asset management not only within transportation agencies but also in other public works domains (e.g. piped water, electric power, gas, telecommunications, sewerage, access roads, industrial parks, shell buildings, and rail services).

Two important non-governmental organizations active in the US TAM arena are the Transportation Research Board (TRB) and the Government Accounting Standards Board (GASB). The TRB is part of the National Research Council, which in turn is the operating arm of the National Academy of Sciences, the National Academy of Engineering, and the Institutes of Medicine. The TRB has a broad scope of activity in all aspects of transportation, with some 200 technical committees that deal with a variety of issues ranging from hard engineering to policy and management. The TRB established a taskforce on asset management in 2000, which became a standing committee in 2004. Its purpose is to encourage research related to TAM and to coordinate findings generated by other committees (McElroy, 2000). The committee meets twice annually and serves to build stronger liaisons with a broad area of activities that are related to TAM; emphasize the institutional aspects of TAM; promote research and publications related to TAM in TRB publications; and explore the applicability of proven private sector asset management tools in a public setting.

The GASB is the organization that establishes accounting standards for US government units below the federal level. The GASB has played an important role in TAM because of its adoption of a policy, called Statement 34, that requires government agencies to show their infrastructure assets in their financial reports (US Government Accounting Standards Board, 1999). Traditionally, infrastructure capital assets have been reported at the discretion of the individual transportation or government agency. Statement 34 requires transportation agencies to: establish an inventory of their existing infrastructure facilities; assign values to the facilities that make up this inventory, based on historical or estimated historical cost; and assign yearly depreciation charges to these values. In lieu of assigning depreciation charges based on historical costs, the agency may elect to report their infrastructure assets using a "modified approach," which requires them to adopt an asset management system that is capable of documenting and providing information on the estimated annual amounts needed to maintain and preserve its infrastructure assets at or above a prescribed level (McNamee et al., 1999).

In many respects, GASB Statement 34 is similar to the financial accounting reforms that occurred in Australia and the UK in that it represents a shift toward enterprise-based financial accounting practices and an enhanced use of financial accounting information in public management. The AASHTO and the FHWA have shown great interest in GASB Statement 34 as a facilitator of TAM.

TAM has also attracted substantial commercial interest, particularly from consulting firms. In the USA, these firms often perform work for states under contract, and in that role act both as developers and purveyors of best practices.

7. Concluding remarks

TAM systems have received considerable interest in the transportation community internationally. The advancement and development of TAM systems vary quite extensively both across countries and across transportation agencies and functions. Challenges to the widespread adoption of TAM systems include the development of solid performance measurements, and the availability of tools and methods for achieving the objectives of TAM. In addition, there are organizational and institutional barriers to TAM adoption. Many of the tenets of TAM, such as system integration, customer orientation, and life-cycle analysis, fundamentally challenge conventional approaches to transportation service delivery. Given these challenges, widespread adoption of TAM in the USA is not expected to occur until the next decade (Switzer and McNeil, 2004).

References

Barret, P.J. (1996) *Asset management handbook*. Canberra: Australian National Audit Office.

Cambridge Systematics (2002) *Transportation asset management guide*. Prepared for the National Cooperative Highway Research Program, project 20-24(11). Washington, DC: NCHRP (http://www4.trb.org/trb/crp.nsf/All+Projects/NCHRP+20–24(11)).

Clash, T.W. and J.B. Delaney (2000) "New York State's approach to asset management: a case study," *Transportation Research Record*, 1729:35–41.

International Federation of Accountants (1997) "Perspectives on accrual accounting," *Standards and Guidance* (September 17) (http://www.ifac.org/StandardsAndGuidance/PublicSector/OccasionalPaper3.html). New York: IFAC.

Kadlec, A. and S. McNeil (2001) *Applying the Governmental Accounting Standards Board's statement 34: lessons from the field*, Paper No. 01-3076 (CD-Rom). Washington, DC: Transportation Research Board.

McElroy, R.S. (2000) *U.S. Federal Highway Administration initiatives – 1999–2000 – transportation asset management.* Washington, DC: Federal Highway Administration, US Department of Transportation.

McNamee, P., D. Dornan and E. Chait (1999) *Understanding GASB 34's infrastructure reporting requirements: a PriceWaterhouseCoopers White Paper.* New York: PriceWaterhouseCoopers.

Neumann, L. and M. Markow (2004) "Performance-based planning and asset management," *Public Works Management and Policy*, 8:156–161.

Stalebrink, O.J. and J.L. Gifford (2002) "Actors and directions in U.S. transportation asset management," in: *The Transportation Research Board 81st Annual Meeting*. Washington, DC: National Research Council.

Switzer, A. and S. McNeil (2004) "Developing a roadmap for transportation asset management research," *Public Works Management and Policy*, 8:162–75.

US Federal Highway Administration (2002) "Asset management: preserving a $1 trillion investment," *Focus*, May:1–2.

US Government Accounting Standards Board (1999) *Basic financial statements – -and management's discussion and analysis – for state and local governments*. Governmental Accounting Standards Series, No. 171-A, Statement No. 34. Norwalk: US Government Accounting Standards Board.

Windsor, J., R. Adams, S. McNeil and L. Ramasubranian (2004) "Transportation asset management today: communities of practice collaboration in the transportation industry," *Transportation Research Record*, 1885:88–95.

This Page Intentionally Left Blank

Chapter 17

FINANCING TRANSPORT INFRASTRUCTURE

RICO MAGGI
University of Lugano

1. Introduction

It is difficult to consider the financing of transport infrastructure without reference to the ongoing discussion on the role of the state (e.g. privatization/deregulation issues) on the one hand, and the size of public debt and hence the amount of taxation and the allocation of general tax money to specific uses on the other. The dominating view in the public sphere is that infrastructure has to be planned and provided by the state. according to public "needs" In contrast to economists' focus on efficiency, practical interest is on distribution, and the central question is: who should contribute, how much, and in what way to the financing of this infrastructure? Until very recently it seemed clear that the construction of transport infrastructure would mostly be financed by general tax money, with users contributing in one way or another (fuel taxes, tolls) to the costs of operation. However, for some time now, participation of the private sector in financing through public–private partnerships has raised expectations that public funds would no longer be a relevant constraint on the realization of single infrastructure projects. Moreover, discussions on road tolls and, rarely so far, on the introduction of road pricing have provoked curiosity about the options available for financing and regulating the use of infrastructure. Unfortunately, the complexity of the issue, the variety of dimensions involved, and the number of instruments proposed tend to create more confusion among interested practitioners than they contribute to the solution of the problems and the orientation of transport policy toward a strategy based on economic principles.

This chapter imposes some order on all this. However, several circumstances challenge this ambition. First, lumpy investments and long-term considerations often create some serious problems for the purely marginal optimization that usually guides economic reasoning, and, secondly, real-world policies seem to contain rather small amounts of the ingredients recommended in economists' recipes. The former necessitates consideration of the agents that might reap the future benefits of the current engagement in infrastructure building – and

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

therefore might be willing to participate in financing – while the latter requires discussion of the reasons behind this discrepancy.

Section 2 provides a short description of some historical cases of financing transport infrastructure that have led to current practices, and should help set the main topic of the chapter in context. A subsequent section summarizes the economists' arguments on pricing and financing infrastructure. Alhough the pricing of infrastructure use is not the primary subject of this chapter, its discussion is essential, and is made all the more relevant because, in the perception of politicians and voters/taxpayers who often consider prices, user charges, and taxes to be one and the same thing. Therefore, the section will compare normative concepts with some considerations drawn from the positive economics of infrastructure as a distributional issue with a strong spatial dimension. This is followed by a description of the actual situation where pure public financing of infrastructure is sometimes substituted by public–private partnership financing and where toll schemes are gradually introduced that contribute to the financing of infrastructure construction. The discussion distinguishes between the various spatial contexts and dimensions of transport infrastructure.

2. The evidence from past to present

The growing problems in very many developed economies of keeping the public budget in balance, or, more realistically, of limiting debt to some sustainable level, have provoked an intensified interest in alternatives to purely public financing of important transport infrastructure and other investments and gaining access to private capital sources. This seems only logical after almost 30 years of deregulation in the transport sector during which time the general public has seen formerly state controlled transport systems replaced by privately operated and profit-driven business. However, the recent involvement of the private sector in financing large infrastructure projects (e.g. the Channel Tunnel) has taught us that, given the large-scale dimensions of some infrastructure and the complexity of modern transport markets, involvement of the private sector in financing infrastructure might in some cases be more difficult than regulating the private operation of services on existing networks.

Leaving aside ancient history (for an overview see Levinson, 2002), we find that, from a historical perspective, raising private capital for transport infrastructure projects was not unusual. Medieval bridges in England owed their existence to religious charity, individual donations, tolls, and public funds, as Steane (1997) notes. In the sixteenth and seventeenth centuries, bridges in cities were often constructed with buildings on them (e.g. the bridges in Paris, Florence, and Venice), and placing a market place on a bridge created income opportunities that justified the investment – and not only for transport reasons. Obviously, modern travel at much higher speeds has invalidated this option. In the nineteenth century, sluices in Europe were financed by private capital because the surrounding, drained, land could be used to conduct business affairs. In Colonial Connecticut, money for the construction and repair of roads and bridges came from the town's annual highway tax, levied on households according to how much land they owned. Starting in the middle of the seventeenth century in Great Britain and in the late eighteenth century in the USA, important roads were realized as turnpikes. Trusts and turnpike companies were assigned the right to raise tolls on their roads in compensation for their construction and, above all, maintenance. Toll farming (franchised farmers collected the tolls), as a sort of local public good provision, substituted forced work, and was itself substituted by public financing when the increasing integration of marginal roads and competing transport systems (canals, railways) made the roads less profitable in the second half of the nineteenth century (Levinson, 2002).

The engagement of the private sector in roads seems to follow cycles; it was important at the time of the turnpikes, but vanished with the arrival of the railway. Roads continued to be publicly funded with the arrival of automobiles because the road system had to be completed in a capillary fashion, and toll revenues would not have justified investment. Only in recent times, when motorways no longer satisfy huge and channeled traffic flows, and where, especially in large agglomerations, congestion signals excess demand, private capital is again attracted to road infrastructure.

Historically, the Dutch canals have represented an important transport network. From the late seventeenth century, their construction was financed by interested cities that created special societies for raising tolls from commercial users (Rienstra and Nijkamp, 1997). In the nineteenth century the larger canals were first constructed, financed, and operated by private companies; later, when the number of canals built could not satisfy the needs of the Crown and government, public financing took over; and in the twentieth century, private financing vanished. The trend seems again to be that private capital is attracted as long as infra-marginal projects are involved and the respective transport system is competitive. The new technology of the industrial age held the promise of significant profits, and public funding was thus absent. The state often provided the land though, and regulated tariffs and competition. For example, railways throughout Europe all started as private operations. In the second half of the nineteenth century most of the profitable railways in many European countries were built, but the national networks were far from being complete and integrated. With the motivation of equity and public obligations, the state stepped in and funded the development of railway networks from then onwards. Again a cyclical phenomenon can be observed: in the early railway era, investments were mainly private, but when the network became increasingly extended, often in view of goals of national coherence, private involvement vanished. When in the middle of the twentieth century the car began to dominate mobility in developed economies, the public sector had to finance railway infrastructure entirely. Only in recent decades, when roads in high-density urban areas became increasingly congested, have public transport solutions become, once again, attractive to private capital.

In the recent past, several important transport projects across Europe have been financed entirely or partially with private funds: the Channel Tunnel between France and the UK, the Great Belt Link between Denmark and Sweden, and the Mont Blanc Tunnel between France and Italy, as well as smaller projects such as bridges in Portugal and the UK and tunnels in the Netherlands. These projects have several common features: they all represent critical links in a larger network; they all promise a return on investment through user charges; and the private investment mostly takes the form of loans (only the Channel Tunnel was financed solely from stock issues). But private engagement does not mean that the investors finance the whole project nor that they necessarily bear all or part of the risk (Rienstra and Nijkamp, 1997). This interest of private investors in missing links is in line with the long-term cycles described above: the combined effect of network extension, competing modes, and land use development will create specific bottlenecks that guarantee high demand. In a certain sense, the current business opportunities in transport infrastructure are analogous to the bridge tolls in earlier times.

Though not following exactly the same paths as in Europe, the general trends in Japan and South-East Asia have been similar, but with one important exception: over the last decade the construction of urban mass transit networks in capital cities, and some intercity railway lines, have been linked to real estate development around the stations. This has significantly increased the expected return on investments in transport infrastructure for the consortia involved. There is no doubt that this kind of solution, where the benefits of the investment in infrastructure are capitalized into surrounding land uses, is feasible only if the property is one and the same and the possibilities of reaping the positive externalities from increased accessibility without investing does not exist.

In the USA one can find various forms of private sector participation in highway projects, but private investment in public transit projects is rare. Table 1 provides an overview, and confirms the finding that private investment is most probably attracted to investment with a strong missing link character in a large agglomeration. It should be noted, however, that the table illustrates the situation in 1998, and several projects did not turn out as expected.

A special case is Norway, which is sometimes considered as having the best practices for road financing by means of tolls (Odeck and Brathen, 2002). Toll financing has been used for financing road construction in Norway for over 100 years, and toll revenues make up 25–30% of the budget for road construction. The specificity of the Norwegian model lies in the non-profit toll companies, composed of local authorities and private investors. Odeck and Brathen discuss

Table 1
Sponsors and features of highway financing in the USA

Sponsor	Major features of financing	Examples
Private equity investors	Finance and develop the project using private resources	Dulles Greenway (Virginia) 91 Express Lanes project (California)
Private, non-profit entity	Issues tax-exempt debt backed by tolls (and without recourse to taxes) and oversees the project under the terms of the agreement between the state and the private developer	TH 212 (Minnesota) Southern Connector (South Carolina) Interstate 895 (Virginia) Tacoma Narrows Bridge (Washington) Arizona toll projects
Special-purpose public agency	Issues tax-exempt debt backed by tolls (and without recourse to taxes) and oversees the project under the terms of the agreement with a private developer	E-470 (Colorado) Orange County, California, transportation corridor agencies
State agency	Issues tax-exempt debt backed by tolls (and without recourse to taxes)	Some turnpikes
State agency	Issues tax-exempt debt backed by taxes	Most highway projects that are financed by debt
State agency	Finances highway on a pay-as-you-go basis using state taxes and fees plus federal aid	Most highway

Source: US Congressional Budget Office (1998).

why private investors in these partnerships are so interested in anticipating and accelerating the construction of critical links in sparsely populated areas, concluding that private capital is attracted due to the external benefits created for transport-dependent future business.

The general picture emerging is that whenever infrastructure promises a financial return on investment, the engagement of private capital is feasible. But because transport infrastructure takes various forms (Table 2), the solutions for financing and, above all, attracting the participation of private investors differ according to the type of infrastructure. In the case of relatively small single projects at nodal points, pure private financing is possible, and might become the rule in the future. At the other extreme, large national or continental networks or significant parts of them are, and most probably will continue to be, maintained and extended with public finance. In the case of new investments in urban transit,

Table 2
Different kinds of transport infrastructure

	Single project	Network
Small	Car parks, stations, multimodal terminals	Urban transit
Large	Bridges, tunnels, airports	Road and intercity rail

public–private partnerships will increasingly arise because private investors will have a share in the profits from developments associated with, for example, stations due to the investment. Finally, the realization of single large projects frequently represents the closing of a missing link. In this case public–private partnerships become relevant and feasible in the form of project financing, if forecasts of use promise significant incomes from tolls.

But why is public financing of transport infrastructure and the subsidiary engagement of private capital the rule? If public funds are limited and allocated in an inefficient way, why not follow standard economic principles? This implies shifting attention to pricing, putting the focus on users of infrastructure in the center, and reflecting on externalities. Economic rationale implies that infrastructure investment should be guided by the willingness to pay of users and third parties benefiting from the investment. Public bodies would then not be involved in financing as such but instead would regulate externalities. However, because the issue is not only complex but also the subject of other chapters in this series, it will not be treated in depth here. Rather, the normative arguments will be combined with a discussion of the distributional implications (who pays, who benefits?), which is of political relevance. The latter issue concerns various groups of users as well as third parties benefiting indirectly from infrastructure.

3. Financing transport infrastructure via pricing

According to general economics principles resources should be allocated in an optimal way, so that an additional sum spent (a further infrastructure project realized) returns exactly the same benefit for society in every possible investment. There should be no better alternative way to invest the same resources. This economic norm, derived from welfare economics, implies that the price charged for a unit of a transport service should exactly equal the marginal cost (i.e. the cost of producing this additional unit). This rule strictly holds in a competitive market setting, where charging a higher price and making profits would attract competitors. The problem with this so-called "first-best" rule in the case of transport infrastructure is that setting the price equal to the marginal cost normally results in economic losses for the infrastructure supplier. This is due to the lumpy investment that is typically associated with any infrastructure project. The high fixed cost results in decreasing average cost for most relevant markets. However, as long as the average cost is decreasing, setting prices equal to the marginal cost will not cover this average cost (decreasing average cost implies that an additional unit costs less than the average so far).

Hence, it will generally not be feasible to recover infrastructure cost by means of a general market rule. Infrastructure will not be supplied by a private firm (a natural monopolist in this case) if the rule is first-best marginal cost pricing. If a community wants to retain first-best pricing, it has to cover the ensuing deficit from cross-subsidies or general tax. This can and very probably will result in distortions elsewhere. In order to solve this conflict between short-run allocation efficiency and long-run cost recovery, economists have introduced second-best solutions combining marginal cost pricing with elements permitting the fixed cost to be recovered also. Charging infrastructure users their associated marginal cost and distributing the fixed cost has resulted in three variants: demand-oriented (elasticity based) price discrimination; two-part tariffs; and congestion and environmentally based pricing. Elasticity based pricing consists of market segmentation according to price elasticity of demand and then charging a mark-up on the marginal cost that is inversely proportional to the elasticity. With Ramsey prices set equal to the average cost, the fixed cost is fully recovered.

But not everyone agrees that second-best pricing can be more than a short-run option. Hensher and Brewer (2001) present the options in the following simple form: user charges are a function of the marginal cost and the price elasticity of demand. The options are to allow for monopoly pricing, i.e. a monopoly mark-up on marginal cost in order to permit private operators to supply the service (eventually regulated by a price cap). At the other extreme, the regulator can decide to charge the marginal cost only (a mark-up of zero) in order to maximize welfare and finance – in theory without distortion via general taxation. Finally, Ramsey prices can be used as a device for constrained welfare optimization, permitting the operator to break even. This implies a strictly short-run consideration. Following Walters (1968), the marginal cost, unlike in standard economics discussions on natural monopoly, is assumed to be short-run, i.e. depending on individual users' decisions on undertaking specific trips. Given what we know about the dimensions, life cycle, and hence lumpiness of many transport infrastructure projects, this is understandable. It must be made clear, however, that the textbook economics on Ramsey pricing and two-part tariffs does not, except for practical reasons, need a distinction between the short and the long run. The issue becomes of relevance if general economic impacts on land use, economic externalities for real estate owners, future generation impacts, and the like come into picture. In this case only costs linked directly to individual use will be of relevance for financing via pricing. This will, however, provide financing basically for the variable operating and

Table 3
Types of good and examples of transport services

	Rivalry	
Excludability	Low Economies of scale No congestion	High Diseconomies of scale Congestion
Impossible Expensive Undesirable	**Public good** Example: rural roads Supply: public	**Common good** Example: congested urban road network Supply: public plus regulation of externalities (e.g. road pricing)
Feasible Cheap Possible	**Club good** Examples: national motorway and railway networks Supply: public or private plus, for example, natural monopoly regulation	**Private good** Examples: urban transit, single congested roads or lanes, parking Supply: private plus user charges

maintenance costs and congestion and environmental externalities, but not for the construction and replacement costs. As a consequence, subsidies are necessary, and will have to be decided upon in a cost–benefit framework.

To be of relevance, these economic principles need applying to specific cases. As has been illustrated, transport infrastructure varies in size but also with respect to the kind of service provided. This requires a more differentiated economic argument with respect to the relevance of normative economics concepts and, more specifically, private sector involvement. Transport infrastructure can be characterized with respect to relevant economic notions such as contestability, potential for cost recovery from user charges (returns to scale) or externalities, and, more generally, with respect to the type of good (private, club, public), as illustrated in Table 3.

According to these criteria, significant differences exist, e.g. between extreme cases such as urban transit and rural roads. Urban bus and rail transport services exhibit a high potential for competition, provide a private good, have a cost structure permitting for recovery through user charges, and cause limited environmental externalities. As a consequence the "marketability" (World Bank, 1994) would be high, and private funding of infrastructure feasible. The trend, at least in Europe, has only recently been toward tendering contracts containing numerous public service obligations to private companies. As a consequence, the license holder will strive for efficiency, but prices will be politically set lower than marginal cost, and the system will be congested. Rural roads – in contrast to the club good offered by primary roads – represent a case at the other extreme of marketability. They have a low potential for competition, offer a public good, and the potential for cost recovery is low. In this case, private investors are absent, and a policy providing infrastructure use at the average cost of national road networks provokes excess supply.

The implications from a purely normative point of view seem straightforward. Urban road and mass transit networks could be optimized according to economic principles. Second-best pricing would include charges for congestion and environmental externalities, and multi-part tariffs would allow for cost recovery. Interchanges such as port and airport facilities, railway stations, and inter-modal terminals could be organized as clubs where the organized interests would negotiate pricing schemes that would satisfy conditions of cooperative game theory (Rothengatter, 2003). Finally, large national or continental road and rail networks would be publicly financed according to public service concepts, and congestion and externality charges would complement general tax and fuel tax revenues.

This sketch, however, is unrealistic, because it does not consider the distributive implications that drive policy-making in the transport sector. Communities normally finance infrastructure via general taxation, and taxes on fuel that do not represent marginal cost prices. As Winston (1991) noted for the USA, fuel taxes were a reasonable way to finance national road infrastructure as long as roads were uncongested and in good condition. However, in spite of obvious inefficiencies and funding problems, the above rationale is rarely relevant for infrastructure policy. From a public policy perspective, the focus is on distribution, and the central question is: who should contribute how much and in which way to the financing of transport infrastructure? The decision on how much of which infrastructure to build and where represents the flip side of the coin.

One relevant dimension of distribution is space. This dimension becomes especially relevant in federal systems, but is also of significance in more centralist states as long as there is a focus by politicians on local policy markets. If voters suffer from "fiscal illusion" (Mueller 1989) and do not perceive the true marginal effect on tax rates from an investment in infrastructure, then the bureaucrats and elected politicians will tend toward excess supply. In a spatial context this means that federal grants to local infrastructure create the illusion of a free ride. But even in the absence of fiscal illusions, financing local infrastructure through national general taxation implies to some extent a free ride created by transfers from national to local tax payers. This problem is of specific relevance in a transport infrastructure context due to the piecemeal and necessarily localized character of investments in transport networks. Hence, not only do local transport systems suffer from this, but also national ones.

An illustration of a different variant of the spatial distribution game regarding transport infrastructure is presented by Levinson (2002) in the form of a decision on whether to levy a tax or toll, and its implications for residents and non-residents

Table 4
Revenue mechanisms and user groups

	Residents	Non-residents
Tax	Hard ride: payment exceeds fair share	Free ride: no payment for use of road
Toll	Easy ride: payment is less than fair share	Hard ride: payment exceeds fair share

Source: Levinson (2002).

(Table 4). For taxation, we find a free ride situation opposite to that cited above – residents finance the infrastructure via a local tax and non-residents can take a free ride. If tolls are paid by residents only if they leave the boundary of their community, then non-residents will have to subsidize the infrastructure. The implication is that financing and provision of infrastructure will very much depend on the spatial set-up. Urban core regions will be interested in setting up tolls and pricing regimes, but the urban fringe will oppose this. The larger an agglomeration, and therefore the smaller the share of the core in attracting traffic, the lower is the possibility of introducing user charges, and vice versa. An illustration of the latter case is the Norwegian city of Trondheim, where, according to Small and Gomez-Ibanez (1998), one-third of the population lives inside the toll ring but obviously benefits from some of the road improvements. According to this logic, agglomerations will prefer tax schemes, preferably on a national level, which will allow them a free ride to downtown and receive cross-subsidies from rural regions.

Distribution obviously, however, does not only have a spatial dimension (for an overview see Rietveld and Verhoef, 1998). The whole concept of public service obligation to providers of transport services is based on equity issues. Much of the opposition to pricing schemes is based on the same grounds. It is independent of the effectiveness of the scheme. Pricing as a substitute or complement for general tax and fuel tax funding is usually perceived as regressive, which seriously hampers its feasibility. However, environmental externalities that should be internalized via Pigouvian taxes, according to economists, represent a distributional issue. In this case the free ride is taken at the expense of future generations – the same group that will have to pay for the free ride realized by present generations when financing large project by debt.

4. Public–private partnership

The pricing discussion reflects the normative approach of welfare economics, where benevolent politicians and efficient civil servants implement the optimal policy. The public choice perspective on distribution confronts this with the reality of political markets. However, as has become clear from a look at the historical development and recent trends, financing of transport infrastructure has often involved the private sector. In this short section, two variants of involvement of private investors are mentioned briefly. The first is project finance, and the second is value capture.

Private contributions to financing any infrastructure are normally based on business opportunities. These can arise from payments by users of the infrastructure or from third parties benefiting form the investment and willing to pay for this. Taking users first, the toll and pricing schemes mentioned in the previous sections do not necessarily have to be implemented by the public sector. Private investors will be interested in building and operating infrastructure with or without public participation as a function of the expected revenues from users. This kind of engagement of the private sector is categorized as project finance (Levy, 1996).

Project finance is a financing method that focuses not on the credit status of a company but on cash flows that will be generated by a specific project. In the case of transport infrastructure this means the income created by the future users of the project. Project finance always implies a transfer of risk to the private investor, whose return on investment will depend on the use of the infrastructure. The specificity of transport infrastructure makes this risk, sometimes, particularly high. Forecasting the use of transport infrastructure is difficult because transport is heavily regulated and hence the future use for the infrastructure will depend on the demand for specific services subject to changes in regulation. Typical examples are public regulations on user charges themselves, changes in regulations on competing modes of transport and infrastructure, and changes in complementary network developments. However, long-run demand for the services offered by the investor depends also on residential location and land use development, changes in transport technology, and the social acceptability of user charges for the existing infrastructure. All this makes long-term traffic forecasts unreliable, and normally do not extend beyond a time horizon of 15 years. Transport infrastructure, however, has a planning and construction period of 2–10 years and a lifespan construction from 20 years to over a century (Gerardin, 1993). Therefore, the usefulness of traffic forecasts as a base for project finance is limited. The consequence in the past has often been that overly optimistic travel forecasts created problems for the investors, and did not improve the prospects of attracting private capital to large infrastructure projects.

The likelihood of project financing can be significantly increased in different ways. One model becoming increasingly popular is public–private partnerships, which imply an involvement of the private investor that goes beyond pure financing and includes mobility management and critical investment and land use planning decisions. Depending on the wider transport policy context, this partnership will focus on the shaping of future mobility in the framework of traffic demand management (interventionist variant) or take the form of the engagement of the investor in economic activities creating the traffic for the infrastructure (value capture). Technically speaking, project finance takes mostly the form of BOT (build, operate, and transfer) contracts, where the private investor builds and operates the infrastructure for profit and transfers it back to the public sector at the end of the contract period, i.e. when their investment is amortized.

Turning to third parties benefiting from transport infrastructure investment, the focus is on accessibility. Business locations becoming more accessible due to an infrastructure investment, households being able to save on a second car due to a new metro line, and real estate developments along a transit project or around highway junctions are all actors that have gained value from a project. If these third parties can be convinced to contribute to financing, part of the added value can be "captured." One variant of this takes the form of land value taxation, taxing a portion of the additional value of properties resulting from improved accessibility. However, more in line with the idea of public–private partnership is the participation of third parties in non-profit companies for infrastructure development, as in the case of Norwegian toll roads, or as integrated and profit-driven developers of transport infrastructure and real estate development, as is the case in Japan, for example. Just like project finance, value capture in this form allows for a partnership between private and public investors that is mutually beneficial for the transport system itself in terms of ridership and for the investors in terms of direct or indirect benefits. Capture of added value from infrastructure development has always existed, and not just in the form of a land tax; for example, in the sixteenth century the Senate of Venice organized a lottery, where the stake was a "lotto" (piece of real estate) around the Ponte Rialto.

5. Conclusions

The financing of transport infrastructure has been discussed here as an issue between efficient pricing and political distribution. Historically, toll pricing and private investment in transport infrastructure has been frequent and important (e.g. in the railway era), and market orientation relevant. More recently, e.g. in the economic growth after the Second World War and the construction of large motorway networks as well as urban transit systems, public funding and operation has largely dominated.

Economists' models for the return to some economic principles promoting an efficient allocation of funds, environmental resources, and users' time are of limited relevance as long as distributional issues prevail. The conclusion therefore is that the increasing emphasis in the literature on analyzing and understanding the economics of transport policy and the mechanisms of collective choices in the sector will improve the feasibility of financing infrastructure via pricing.

References

Gerardin, B. (1993) "Financing transport in Europe," in: D. Banister and J. Berechman, eds, *Transport in a unified Europe*. Amsterdam: Elsevier.

Hensher, D.A. and A.M. Brewer (2001) *Transport. An economics and management perspective.* Oxford: Oxford University Press.

Levinson, D.M. (2002) *Financing transportation networks*. Cheltenham: Edward Elgar.

Levy, S.M. (1996) *Build, operate, transfer: paving the way for tomorrow's infrastructure*. New York: Wiley.

Mueller, D.C. (1989) *Public choice II*. Cambridge: Cambridge University Press.

Odeck, J. and S. Brathen (2002) "Toll financing in Norway: the success, the failures and perspectives for the future," *Transport Policy*, 9:253–60.

Rienstra, S.A. and P. Nijkamp (1997) "Lessons from private financing of transport infrastructure," *Revue Economique*, 48:231–246.

Rietveld, P. and E.T. Verhoef (1998) "Social feasibility of policies to reduce externalities of transport," in: K.J. Button and E.T. Verhoef, eds, *Road pricing, traffic congestion and the environment*. Cheltenham: Edward Elgar.

Rothengatter, W. (2003) "How good is first best? Marginal cost and other pricing principles for user charging in transport," *Transport Policy*, 10:121–130.

Small, K.A. and J.A. Gomez-Ibanez (1998) "Road pricing for congestion management: the transition from theory to policy," in: K.J. Button and E.T. Verhoef, eds, *Road pricing, traffic congestion and the environment*. Cheltenham: Edward Elgar.

Steane, J. (1997) *Medieval bridges in Oxfordshire.* Wantage: Oxfordshire County Council & Vale and Downland Museum (http://www.wantage.com/museum/Local_History/Medieval%20Bridges%20in%20Oxfordshire.pdf).

US Congressional Budget Office (1998) *Innovative Financing of Highways: An Analysis of Proposals,* Washington, DC: US Congress.

Walters, A.A. (1968) *The economics of road user charges. World Bank staff occasional paper No. 5.* Baltimore: Johns Hopkins University Press.

Winston, C. (1991) "Efficient transportation infrastructure policy," *Journal of Economic Perspectives*, 5:113–127.

World Bank (1994) *World development report 1994: infrastructure for development*. Oxford: Oxford University Press.

This Page Intentionally Left Blank

Chapter 18

A BANKING PERSPECTIVE ON TRANSPORT

NICHOLAS HANN and TIM MACK
Macquarie Bank Limited, Sydney

1. Introduction

This chapter provides a banker's view of some issues that have already been discussed in these handbooks. It will:

- introduce financial concepts relevant to the transport industry;
- review the funding of transportation infrastructure;
- discuss models for public transport finance;
- discuss new financing structures;
- summarize the financial products used in recent Australian and international project transportation;
- examine the future directions for financing transport opportunities.

The transport sector incorporates a wide range of industry subgroups that, by their definition, are all broadly involved by some means in the movement of people or goods from point A to point B. Therefore, a one-man trucking operation and Qantas both fall under the transport industry banner. Clearly there are significant differences between the many types of transport businesses. This chapter focuses on the financing of large-scale transport infrastructure such as railways, toll roads, and airports. Finance raised for these types of project is generally characterized as "project finance," and is explained below. This is in contrast to the small business finance that would be sought by a truck owner and the corporate finance that would be raised by Qantas.

2. A banker's view of the transport sector

In raising project finance for transport infrastructure it is important to consider the question "What do lenders look for and how can the transport sector meet their requirements?"(Macquarie Corporate Finance, 2000). Lenders look at a wide range of factors, both qualitative (e.g. management skill and experience) and

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

quantitative. Many factors impact both the expected performance of a project and the risks attached to achieving that performance. However, three key issues are often a focus in financing are assets, profitability, and cash flow:

- Lenders value physical assets highly. Traditional bank debt relies heavily on taking security over assets in the expectation that this will provide something to sell in the event of a default on the loan. Moveable and transferable assets are particularly highly valued for this reason. The transport sector is very asset rich, which opens up great potential for asset-based financing structures, but there is a huge difference between, say, the civil works and track work for a railway and the trains that run on it.

  The transport sector is a curious combination of the capital-intensive and the service-intensive. It is no accident that the majority of private finance raised has been in those transport sectors in which capital plays a significantly larger role than the provision of services, such as toll roads, airports, and shipping ports. For example, the new Western Sydney Orbital toll road will require approximately Aus. $1.5 billion in private finance just to cover the cost of building the road.
- A track record of profitability in the industry is very important to bankers, and this is where the transport sector as a whole is a weak prospect for finance. Standard & Poor's, a credit rating agency, award a low overall industry credit rating to the transport sector, with road transport getting a "C," and rail a "C/D." This is due to the fact that while most transportation companies have positive operating cash flow, they often struggle to earn an adequate return on their capital investments – this problem is reflected in high depreciation charges and low profitability. This issue is reflected in the ratings of individual companies. Of 27 major transport companies in Australia and New Zealand, 42% of the debt issued was unrated, 43% rated BBB/BB, and only 14% rated A (Standard & Poor's, 2001). Many transport infrastructure projects will never be very profitable – this is not because they are bad projects or even unfinanceable projects. Transport infrastructure can throw off large amounts of cash to pay financiers but be made unprofitable on an accounting basis due to high depreciation and interest charges. The ability to repay debt (with interest) and to provide cash returns to equity investors over the life of the asset becomes more important than the profitability. Many infrastructure concessions have been established as trusts, not companies, because trusts allow distribution of cash flow to investors ahead of creation of a profit.
- Cash flow is the most important feature of financeability, since it is cash that is required to service debt and pay a return to equity. Predictability of cash flow is particularly important. The track record is sometimes used as a substitute for predictability – although this will not be true in many cases.

> Cash flow is a very different method of valuation, and even today cash-flow-based lending, where repayment of debt is based wholly on the cash flows generated by a single asset or project (known as limited recourse financing or project financing) is a fairly specialized area of banking, albeit the area that has the most application to the transport sector. Where cash flows can be predicted with a high degree of certainty, project financing and securitization techniques allow a higher level of debt to be raised proportionate to equity than typical corporate financing, with gearing (debt to total funding ratios) of up to 90%.

Typical private sector corporations will have a mix of all three: assets, profitability, and cash flows. Transportation projects are often weighted more heavily toward one or two of the characteristics, rather than satisfying all three equally. This has an effect on the financing strategies employed. Table 1 illustrates the different types of financing instruments that are applicable to different strengths.

The transport sector therefore has a range of features, some of which make it attractive to bankers, and others that cause challenges. Equity investors have a similar but slightly more complex set of requirements. As a result, conflicting agendas are common between the different types of investor with interests in the transport industry.

This is complicated by a very high level of involvement by government in the transport sector around the world; as a provider of "free goods" within the sector (road networks and some public transport), an owner of commercial operations (rail, ports, airports); a prior owner, leaving behind in many cases distorted industry structures (airlines); and an extensive and often unnecessarily intrusive regulator.

Table 1
Financing instruments and their applicability to different strengths

Characteristic	Financing method
Assets	Traditional bank debt Asset backed securitizations
Cash flows	Securitization Project financing Private equity Public equity held through trust structures
Profitability	Public equity Corporate debt finance

The transport industry as a whole has an enormous financial critical mass. It is an asset class that financial markets simply cannot afford to ignore. Total funding for new fixed transportation assets in Australia in 2000–2001 was approximately Aus. $8.3 billion, $2.8 billion of which was provided by the private sector (Australian Bureau of Transport and Regional Economics, 2003).

Transportation claims the largest share of gross domestic product of the developing economies of any infrastructure sector, with an average forecast investment of 2.7% of GDP for East Asia, or some US $607 billion (World Bank, 1995). Transportation is the second largest recipient of development finance after energy, with an average of over US $5 billion per year through the 1990s. However, according to the World Bank, an increasing proportion of the investment in transportation is expected to come from the private sector.[a]

By playing effectively to banker's requirements, the most efficient financing models can be created. The three issues of assets, profitability, and cash flow will be recurring themes of this chapter as we go on to examine the types of institutional model that can maximize access to finance, and the types of financial instrument that have been developed to support these models.

3. How bankers look at projects differently from economists

Finance is taking on an increasingly important role in the planning and identification of transport projects rather than just being left in a "black box" to be opened only when the planning decisions have already been made.

Governments, investors, and banks apply the same basic economic tools to transport investment decisions, but they apply them in very different ways. Widely used evaluation tools are:

- NPV (net present value): the discounted value of the future cash flows associated with a project.
- Project IRR (internal rate of return): the implied discount rate associated with a project's cash flows given the up front price of a project.
- Equity IRR: the implied discount rate associated with the cash returns to equity given the up front equity amount paid by investors.

Time is usually the biggest differentiating factor – the time-scale over which a project must be financially viable is usually significantly shorter than the time-scale over which it can be shown to be economically viable. The difficulty for governments is often in assessing the viability and priority of projects. There are

[a]The performance of this private sector investment is summarized by the Australian Department of Foreign Affairs and Trade (1998).

three measures of viability – which unfortunately are not as distinct as they ideally should be. There are substantial grey areas between them.

- Economic: an economically viable project has a positive NPV, taking into consideration any "externalities." "Externalities" are the economic benefits associated with a project, but do not directly accrue to the project. For example, on a new toll road project, the time and fuel savings achieved by road users have economic benefits that do not flow to the owner of the road. The government may deem a project economically viable even if it is not commercially viable, if there are benefits to third parties.
- Commercial: a commercially viable project is one with an expectation of making a profit.
- Financial: a financially viable project differs from a commercially viable one in that the cash flows must have an acceptable degree of certainty to meet debt payments and provide sufficient cash return on equity. Only a financially viable project can be completely funded by the private sector without government assistance.

For example, say a new city passenger rail link is expected to cost $500 million to develop. Also, say that the net present value of the cash flows that the new rail link is expected to earn (after deducting the cost of running the rail operation) is only $400 million. This project will thus not make money, and is not commercially viable. However, if the government estimates that the project will also save it $200 million in road maintenance costs due to more people catching the train, the project is economically viable.

Similarly, say a new city passenger rail link is expected to cost $500 million to develop. However, say that the net present value of the cash flows that the new rail link is expected to earn is $700 million. This project is thus expected to make money, and is commercially viable. However, if the uncertainty associated with the project is too great (e.g. if the cost of building the rail link could blow out, or if the estimate of the number of passengers expected to use the train is very uncertain) financiers will be unwilling to invest their money in the project, and therefore it will not be financially viable.

The traditional project development approach has been to undertake an economic analysis first, and from that determine the level of government support. More recently, particularly in the case of the proposed Melbourne to Brisbane Inland Freight Rail Corridor, the focus of early feasibility studies has been much more on financeability (also known as "bankability") from the outset. We would suggest this has advantages in ensuring that the focus of the private sector and the government is aligned. The government can then look at filling any financing gaps identified on a co-investment basis with the private sector, and then conduct economic analysis to justify whether such support is required. If the government subsequently decides that private sector involvement is not appropriate, it has a comprehensive commercial and financial understanding of its own investment decision.

The private sector needs to have an active role in ensuring the right project is selected, as well as its role in providing the right delivery.

The financial structure of a project will reflect its risk profile. Financial markets, while imperfect, are still probably the best predictors and valuers of risk that we have. All to often by financing transportation by the balance sheet, governments ignore or fail to manage the risks associated with the project. Governments are largely unable to project finance – in other words there is usually recourse to the government as a whole (and accordingly the taxpayer) for any borrowings. In effect, therefore, the government finances with 100% of the equity risk in a project, making a mockery of its lower cost of debt capital. An exception to this, widely used in the USA but still embryonic in most other countries, is the use of revenue bonds, which (at least in theory) have recourse only to the revenues produced by the asset that they finance. These are described in more detail below.

4. Institutional models

4.1. BOOT/BTO/BLT

Throughout the 1990s build, own, operate, and transfer (BOOT) was the major model for the grant of concessions to the private sector over major transport infrastructure. The basic principles of BOOT schemes are also present – repackaged – in the current wave of public–private partnerships.

Several major variants of BOOT also developed in response to differing risk allocation or taxation requirements. For example:

- build, transfer, and operate (BTO), where the political and/or taxation requirement was for the immediate transfer of the asset back to the public sector after completion;
- build, lease, and transfer (BLT), where the government pays for the infrastructure over the concession life through lease payments.

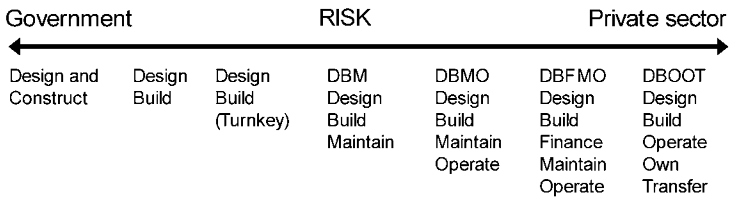


Figure 1. Degrees of risk transfer.

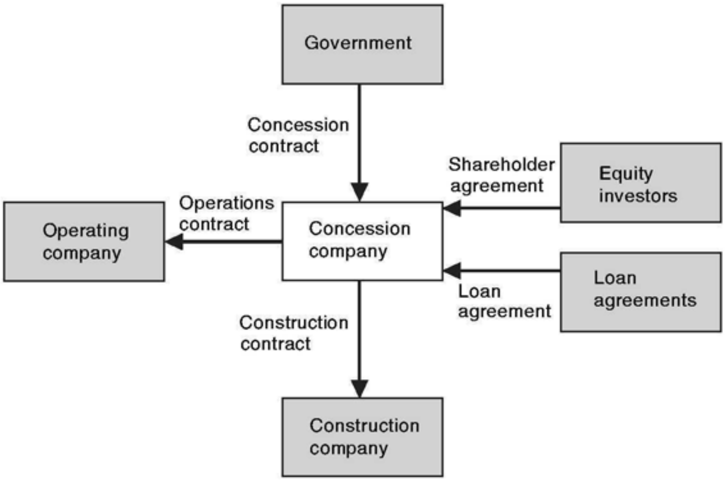


Figure 2. A typical franchising arrangement.

With a passion for greater detail, the BOOT label is now increasingly being replaced with DBFMO (design, build, finance, maintain, and operate) and its variants.

The different degrees of risk transfer are illustrated in Figure 1.

In Australia, the tendency in relation to most large-scale transportation projects (such as toll roads) has been to favor the delivery models that maximize the transfer of risk to the private sector.

4.2. *Franchising*

Franchising is the term used where the private sector is granted a right to provide a service rather than to own infrastructure, although in practice franchises in transport are often a combination of the two. Rebadging as a public–private partnership (PPP) or private finance initiative (PFI) is a catch all term for the introduction of private capital into services with strong governmental influence through the types of mechanisms described above. A typical franchising arrangement is shown in Figure 2.

The franchising model for privatization has introduced private sector management into public transport operations through fixed-term contracts. There has been significant transfer of commercial risk to the private sector in the transportation industry. A fairly recent example, from Australia, of the franchising model in action is the passenger train and tram franchise agreements entered into by the

Victorian government with private sector operators. These franchises essentially grant the franchisees vertically integrated control over their infrastructure and provide strong financial incentives for the franchisees to effectively maintain and upgrade the rail track and train and tram equipment. The mandated capital expenditure built into the subsidy profile and the effectiveness of maintaining government control at the end of the franchise earned plaudits. Although the length of the franchises and the continued need for operating subsidies fall short of full capital investment decision-making in the private sector, this deficiency could be regarded as a necessary interim step toward full private sector decision-making. Unfortunately, due to lower then expected passenger numbers and fare evasion, the franchises have struggled to survive.

One potential problem with the franchise model in some transportation projects is the actual or perceived lack of control over its assets. For example, in the rail freight industry, many rail freight companies put a very high strategic value on owning the track they operate. Ownership of the track gives them greater control over how the track is maintained and operated, which is obviously critical to their business.

So from a financial markets perspective we have the following dichotomy:

- Some disaggregation of a business is necessary to better understand and separate the risks of infrastructure and to ensure that each component of the business is funded at the cheapest possible cost of debt and equity. The modern freight-forwarding company is close to the virtual enterprise, carrying few assets on the balance sheet and earning a margin getting paid to provide a service to its customers, and paying someone else to provide the assets (e.g. trucks) and to do the work.
- However, where the full capital costs of investment are recognized – most clearly seen in major greenfield projects – disaggregating into separate business entities (not connected by strong contractual relationships) can pose serious problems for financing due to a lack of control.

Put simply, investors like to stand as close to the source of cash flows as possible. Owners of infrastructure assets need to be tightly regulated, with their cost of capital and management of assets their key strategic advantage. However, some disaggregating of the industry is almost inevitable, either driven by the government or put in place by the private sector as subcontracting arrangements.

4.3. The role of subsidy

In many areas of the transport sector the need to carve out a financially viable role for the private sector requires a level of government subsidy. This is particularly true in public transport, where there is very often a political unwillingness to set fares at fully commercial levels.

The level of government subsidy required by the transportation sector varies significantly between modes of transport and the type of project. For example, the aviation industry as a whole, and metropolitan toll road projects, do not generally require any government subsidies, while passenger rail transport requires heavy subsidies from the government in order to survive. The trucking industry does not require any direct subsidies from the government; however, it is indirectly subsidized by the government in the form of expensive highway networks that the public pays for.

One of the key issues in the reform of public transport financing around the world has been the strict separation of capital cost subsidies from operating cost subsidies, with up front capital cost subsidies seen as a necessary evil, in contrast to operating subsidies, which have been widely perceived as leading to a lack of financial discipline.

The UK Transport Act (1986), for example, established a requirement that PFI transport projects would only be eligible for capital grants provided that they could demonstrate that they would require no operating subsidies.

Capital cost subsidies are neither necessarily the most efficient form of financing nor do they allow governments the right incentive mechanisms to drive improved performance. New models that bridge the gap between capital and operating costs are consequently some of the most important developments.

Clarity of what structures are acceptable to government in various situations is of immense benefit in targeting private sector opportunities and speeding up the development and negotiation process. Shadow tolling and targeted operating subsidies of the sort we see in the Victorian train and tram franchises can be a useful first step.

Public–private funding mechanisms include:

- capital cost subsidies, which are tolerated when necessary, whereby the government pays for and owns the "low-value" assets, including perhaps tunnels or earthworks, and the private sector pays for and owns the higher-value-added component, such as transport, signaling and communications equipment;
- operating subsidies, which are increasingly only acceptable where the subsidy payment is conditional upon achieving stringent performance standards;
- shadow tolling – the payment by government of a usage based toll indirectly on behalf of the users – which fills the gap between a commercially viable charge and a politically acceptable one;
- financing support, whereby government-owned corporations invest alongside the private sector to provide debt financing to supplement commercial bank senior debt and higher-return private sector equity;

- residual risk support, whereby the government agrees implicitly or explicitly to underwrite the residual value of assets at the end of a franchise or concession term.

5. Key issues for governments and bankers

A broad range of the issues that are considered by bankers and governments in the development and financing of transportation projects are described. In respect of most of these issues, there is no “right answer” – they must be considered on a case-by-case basis for each project, depending on what the government and the private sector are trying to achieve.

5.1. A service or an asset?

The issue of whether providing transportation is a service or an asset is critical to a range of issues:

- PPP and, in particular, accounting treatment;
- the type of financing;
- the structure of concessions.

Walker (2003) has argued that PPP arrangements are actually two different contractual relationships: the initial contract is a design–build contract to provide an infrastructure asset; post-construction, the contract evolves into a service provision that Walker argues often confers a right on the private sector concessionaire to earn a guaranteed rate of return. In practice we do not see too many concessions structured in that way. The argument that a PPP contract can be split into its separate components is to ignore a fundamental component of PPPs, and a fundamental reason why they are preferable to conventional delivery – the integration and optimization of the complex choices between up front capital and long-term operating and maintenance costs. Governments everywhere have a tendency to under-invest in capital (or over-invest if they happen to have surplus cash in that budgetary period) and then to incur a maintenance deficit, resulting in more rapid deterioration of the asset than would be typical in the private sector. This is the fundamental problem of government ownership that PPPs are intended to address. Too often accounting pressures result in perverse pressures and incentives that distort the investment decision.

5.2. Risk transfer and off balance sheet

One of the major benefits of private sector financing of transportation projects is that they remove the public sector responsibility and liability in relation to providing the transportation. This is then reflected on the public sectors' accounts in the form of lower debt.

However, some PPPs act to replace one form of governmental long-term liability (to repay government debt) with another: to make lease or rental payments on the infrastructure or services.

The issue really depends on the degree of risk transfer – the extent to which the private sector is really taking a risk on performance of the asset in order to make its return. Generally PPPs in transport have been among the most far-reaching in terms of transfer of risk from the public to the private sectors.

5.3. Certainty of cost

Flyvbjerg et al. (2003) have recently identified a persistent track record of cost overruns in large-scale transport infrastructure projects globally in the vicinity of 10–40%, which supports the anecdotal evidence from projects such as the Boston Big Dig or the Channel Tunnel. They attribute this to mendacity on the part of transport planners with a bias toward getting projects approved.

This contrasts with the experience in UK PFI projects, which have a far stronger track record. A recent report by the UK National Audit Office (2003) found that of 37 PFI projects, including seven roads, 78% came in on budget compared with only 27% under conventional delivery. One of the greatest benefits of well-structured PPPs, particularly in the transportation sector, which has a very poor track record of construction cost management, is the ability to obtain genuine fixed price, lump sum, date certain, design and construction contracts, which also result in a product that meets performance durability and quality requirements for its concession life.

5.4. Cost of capital

The argument is often heard that because the government's cost of capital is its cost of borrowing money, which is lower than that of the private sectors' cost of capital, the government should provide the funding to all transportation (and other "social service") projects. Although the government's cost of borrowing is lower than the private sectors' cost of capital, this can be offset by risk transfer and other benefits achieved by having the private sector undertake the development and financing of transportation projects. Private sector financing of projects enables risk to be transferred to the private sector, which generally has an efficiency advantage over the government. In addition, even if governments use an estimate of their cost of capital that is higher than their cost of borrowing, the public sector's cost of capital is often a politically driven decision and/or bears little relationship to the risks associated with each project.

This means that it is wrong to fund a project using public sector funding rather than private sector finance simply because the government has a lower cost of funding. To attempt to address this issue, most governments do not use the government's cost of borrowing as its cost of capital when evaluating the merit of private sector financing of transportation projects. For example, when evaluating proposals for the private financing of a transportation project, the New South Wales government in Australia assumes that it would require a commercial return on its funding if the project was to be taxpayer funded (NSW Government, 2001).

5.5. *Termination and step in rights*

One of the reasons why private sector investments can achieve efficiencies that the public sector cannot is the discipline of insolvency in the private sector. Inevitably, private sector projects can and do go wrong. From a public policy perspective, this should not be seen as a bad thing. If private sector projects never failed, then almost by definition government is not transferring sufficient risk or not obtaining sufficiently competitive deals. However, all too often private sector failure is not allowed to occur because governments came under political pressure to prevent failure. There have been numerous examples of bailouts of transport concessions around the world by governments. The frequency of bailout has lead to a reliance on the "implicit" guarantee of government by financial markets even where no explicit guarantee exists. This in turn has distorted the market appetite for risk, and financial markets pricing and willingness to provide finance, which in turn increases the likelihood of failure.

5.6. *Ramp up*

With transportation projects, there is often a ramp up period before revenue levels reach capacity. The term "ramp up" refers to the early stages in the project where demand has not grown to long-run equilibrium levels. This has particular relevance to governments and bankers, as it is often difficult to meet debt service and other liabilities during the ramp up period. Standard & Poor's (2002) note for example that in Australian toll roads "Most of the facilities have under performed the traffic forecasts in the initial years, in some cases by as much as 20%, but the catch-up has been faster." Hensher and Goodwin (2004) explore the reasons for the underestimation of toll-avoiding behavior by travelers that drives this initial underperformance. The key from the financing viewpoint is to ensure that underperformance in the early years does not result in financial failure, as was the case in the Sydney Airport Rail Link. It is becoming increasingly possible to build in an allowance for the slow ramp-up of demand as financing structures become longer term and more flexible.

5.7. Shadow tolling

Private financing depends on a commercial revenue stream. Frequently, however, public policy and politics dictate that the direct user pay principle for public transport or transportation infrastructure cannot be implemented. A large number of transportation projects are therefore developed on a shadow tolling basis. Shadow tolling is a payment by the government for a service based on the number of passengers carried or availability measures. Ideally, it should fill the gap between a commercially viable charge and a publicly acceptable one. It should also be set at a level that does not have a large impact on demand, unless this is a clearly understood policy objective. It should also pass operating risk and decision-making to the private sector (compared with a subsidy based on key performance indicators).

5.8. Revenue and value capture

The other common problem of transportation projects is the high proportion of commercial benefits arising from the project, which are difficult to capture within the project itself. The most obvious of these is the effect that transportation projects have on increasing land and property values. Estimates from Hong Kong during the 1980s to 1990s (albeit during one of the world's strongest property booms) indicated that land within the vicinity of a mass transit station increased ten-fold in value after completion of the mass transit line. Governments tend to leave this increase in value on adjacent property on the table, and have struggled around the world to capture it effectively. However, increases in value can be potentially captured by, for example, selling rights to increased development. Experience indicates the property value capture can only fund the current project if it is fully integrated into private sector delivery.

6. Financing structures and options

6.1. Equity

Equity investors have an ownership interest in the project. As debt holders have greater seniority of claim on the firm's cash flows, equity is the riskiest form of investment. Financing through equity is the most expensive, as return on equity must compensate for the higher level of risk. Equity for private transportation projects was traditionally provided by the major construction companies and equipment suppliers, who begrudgingly contributed equity as a means to secure the contract. Such equity had a number of limitations. It was often simply a reinvestment of profit margins from the construction job, and frequently resulted in an increase in costs; as a result it did not add a great deal to long-term risk transfer; it was generally short term; and was limited in amount and very expensive. Equity for major transportation projects is now increasingly being sourced through specialized infrastructure investors and infrastructure funds, independent of construction contractors and able to exert greater pressure on costs and risk transfer. This equity tends to be much longer term in outlook and significantly cheaper than contractor equity.

The largest specialist infrastructure fund in Australia is the Macquarie Infrastructure Group, which was publicly listed on the Australian Stock Exchange in December 1996, raising Aus. $300 million of capital. It has now grown to be a world leader in private toll road ownership, with assets totaling over Aus. $9.5 billion, spread over 26 individual projects around the world, and with a market capitalization of Aus. $6 billion (at 31 December 2002). The Macquarie Infrastructure Group (2003) annual reports provide one of the most detailed descriptions of the financial features of toll roads.

6.2. *Leasing – finance leasing and operating leasing*

Leasing structures are one of the most commonly used forms of finance in transportation. Leasing falls into two broad categories: finance leasing and operating leasing.

Finance leasing is leasing that extends for the significant majority of the economic life of the asset such that the lessee (the user of the asset) is taking the majority of the risks associated with the asset. Under a finance lease, the lessee may also have the option to purchase the asset from the lessor at the end of the lease period for a bargain price. The lease is called a finance lease because the lessee is using a lease as a way of acquiring the asset without having to pay for it up front. This way of financing an acquisition is an alternative to simply borrowing the money from a bank to buy an asset. For this reason, accounting rules require assets acquired under a finance lease be accounted for as if the lessee "owns" the asset.

Operating leasing is leasing of an asset for only a short period of time. Under an operating lease, the lessee is only taking the asset to use temporarily, and either does not need the asset for its entire life, or does not want the risks and costs associated with using/owning the asset for its entire life.

Recently, we have seen a trend away from finance leasing and toward operating leasing. This has arisen for two reasons. First, one of the benefits of finance leasing has traditionally been that it had some taxation advantages to it. These advantages have, over time, been reduced, reducing the demand for financing leasing. Secondly, there is growing demand for operating leases where the lessor takes the risks associated with owning the assets, and for a growing number of lessors that understand and are willing to take these risks. This shift will result in more leases being treated as operating expenses through the profit and loss statement of lessees, rather than as a purchase and loan on their balance sheets.

6.3. Mezzanine debt

Mezzanine debt, or sub-debt, has lower priority in payment compared with senior and secured debt, but higher priority than equity capital. Mezzanine debt fills the gap in the risk profile between senior debt and equity, and therefore allows a better match between investors' risk appetites and that of the project.

Governments have at times provided mezzanine debt to assist in the financing of infrastructure assets such as toll roads (e.g. the Roads and Traffic Authority in New South Wales in Australia holds a mezzanine debt investment in the M2 Tollroad). Government agency mezzanine debt is a mechanism to reduce the weighted average cost of capital or debt service requirement to within financeable levels, and is an option that should be considered more often.

6.4. Bank debt

The banks are a significant source of debt funding for large transportation infrastructure projects. Bank debt for large transportation infrastructure projects is regularly sourced from large local banks, as well as large international banks (from Europe and the USA).

Bank debt is the most flexible type of debt available, with no penalties for early repayment. In particular, a major advantage of bank debt is that it can be drawn down over time, which is very useful in construction projects where the money is not needed all at once, but over time as money is spent on construction costs.

Bank debt is generally used as short- or medium-term financing (i.e. 2–10 years), rather than long term (i.e. 20–30 years). However, bank debt has been raised on some projects with a term of around 20 years.

6.5. Capital markets debt

Debt provided by the capital markets comes in many different forms. Examples include bullet bonds, amortizing bonds, and inflation-indexed bonds.

Although available in longer maturities than bank debt, one of the curses of transport infrastructure finance is that debt has not been available in maturities that match the cash flow payback profile of the project. In addition, a major constraint outside the USA and other similarly developed capital markets is the absence of a benchmark government bond issuance for the type of maturities that would support transport projects.

However, the markets are always developing and new debt products becoming available. In this current low-interest rate environment, demand for longer-term bonds has grown because they offer a higher interest rate for investors, and we have seen the term of bonds available in some markets increase to 40 years or even longer. Early examples were the Docklands Light Railway and the Lewisham Extension in the UK, which was the first time a PFI project combined bond and commercial bank financing, achieving a single 'A' rating. The structural complexity of transportation projects in the capital markets has led to the development of a significant monoline insurance market, particularly in the UK, to "wrap" the project risks and enhance the credit rating achievable to ensure access to cost-effective financing.

6.6. CPI indexed bonds

CPI indexed bonds are a type of capital markets debt. They provide interest payments and principal repayments that increase with the rate of inflation. Therefore, payments on CPI-linked bonds start off lower than traditional "nominal" bonds that have flat payment profiles, but increase to be higher toward the end of the term of the bond. CPI indexed bonds are, at least in theory, ideally suited to the financing of long-life transport infrastructure projects (such as toll roads), which have revenue streams either linked to or influenced by the rate of inflation.

For example, CPI indexed bonds have been issued to finance the following Australian projects:

- Aus. $200 million for the M2 Tollroad in Sydney in 1994;
- Aus. $350 million for the Melbourne CityLink in 1996;
- Aus. $108 million for the Brisbane Airtrain in 1998.

6.7. Municipal/revenue bonds

Municipal/revenue bonds are a special type of bond financing particular to the USA. This type of bond is very commonly issued by US public agencies who own infrastructure assets such as airports, toll roads, and rail networks. They carry no government guarantee, and are backed only by the revenue generated by the asset.

Interest paid to investors is tax exempt, although the public agency is able to claim a deduction for interest paid. In this way they are essentially a form of indirect subsidy from the US state and federal treasuries. Currently, revenue bonds linked to specific infrastructure assets contribute almost twice the value of general obligation municipal bonds in the USA. This tax-exempt market has been an impediment to the involvement of private sector equity into the US transportation market and has reduced the ability for risk transfer. However, a recent example, described in *Project Finance Magazine* (2003), of the combination of tax exempt financing (under the new US federal Transport Infrastructure Finance and Innovation Act) with the risk transfer benefits of private sector financing was Macquarie Infrastructure Group's financing of the SR125 toll road in San Diego, California, in 2003, which benefited from a US $140 million 35 year tranche of debt at a fixed rate equivalent to 30 year US treasuries as part of a US $635 million total financing package.

6.8. Securitization

Securitization describes the packaging of specific cash flows into a single-purpose bankruptcy remote entity, which then issues bonds repayable from those packaged cash flows. It is very similar to revenue bonds in many ways, without the tax exemption. A good recent example of securitization is the Melbourne rolling stock financing described below.

Table 2 provides a summary of the features of the major types of debt financing available.

6.9. National Express Melbourne rolling stock financing

Notwithstanding the other difficulties experienced in the 1999 franchising of the Melbourne public transport network (commuter rail and light rail) in Australia, as noted by Stanley and Hensher (2003), the financing structures employed were a great success, and a good example of how significant risk transfer can be achieved without substantially increasing the cost of capital.

The franchising of Victorian Trains and Trains was a unique project in that the cost of the franchises was very small compared with the future capital expenditure on rolling stock that the franchisees were required to make. For example, the National Express Group (NEG), the successful bidder for three of the five franchises that were tendered, agreed to buy the franchises for a total of Aus. $45 million, and in the process committed to Aus. $900 million worth of new rolling stock.

Table 2
Characteristics of debt financing

Product	Margins (over prime)	Terms
Bank debt	70–150 basis points	Flexible form of financing Early repayment with no penalty
Capital market		
Bullet bonds	60–130 basis points	Interest is paid over the life of the loan, and all the principle is paid upon maturity
Amortizing	60–130 basis points	Equal payments are made over the life of the loan, each payment servicing both interest and principal
Inflation indexed	60–200 basis points	Usually linked to the CPI. Can be either a bullet bond or amortizing debt. Provides a real interest rate
Sub-debt	300–500 basis points	Has lower priority in payment to senior and secured debt, but higher priority to equity capital

Some of the key objectives for NEG for the project were:

- off balance sheet ownership and financing structure for rolling stock;
- operating lease treatment of the rolling stock for accounting purposes;
- very low cost financing to be cost competitive with other bidders.

In order to achieve these aims, and those of the state, the answer was to have a separate owner of the rolling stock to then lease the rolling stock to the franchisee. However, the new leasing entity (the lessor) did not have any credit worthiness to enable it to borrow the funds necessary to fund the purchase of the new rolling stock.

The solution was the state government agreeing to guarantee the debt service payments on the debt raised to purchase the rolling stock. This gave the debt the highest credit quality (AAA/A1+), enabling the debt to be raised very cheaply and the future concession payments from the state to be minimized. Despite the guarantee, the state still managed to achieve the risk transfer it was looking for by having NEG put up Aus. $45 million to pay for the franchise, and GATX (a US-based rolling stock investor) put in the equity required to complete the funding of the new rolling stock.

Figure 3 gives an overview of the financing arrangements, and Hunter (2000) provides a more detailed description of the financing and its benefits.

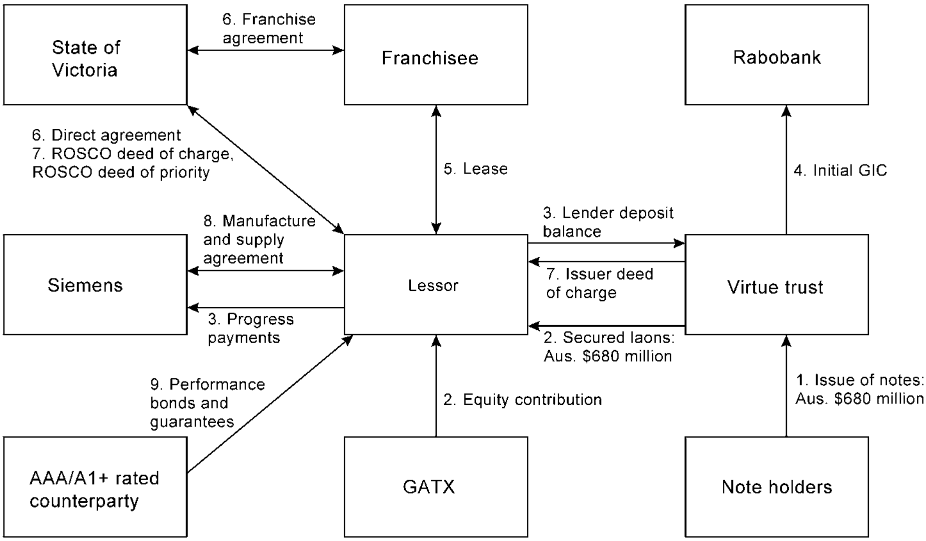


Figure 3. National Express Group Victorian rolling stock financing.

6.10. A tale of two airport rail links: Brisbane Airtrain and Sydney Airlink

The significance of differing financing structures can be seen very clearly in the fortunes of two airport rail projects developed over much the same time-frame in Australia, in Sydney and Brisbane, respectively.

In both cases the projects struggled with initial ridership considerably lower than forecast. In the case of the Sydney Airlink the financing structure relied heavily on commercial bank debt, with approximately 95% debt and only 5% equity. Although the concession was for 30 years the debt had a maturity of just 12 years, and required both payments of interest and repayments of principal (amortization) to commence shortly after completion of construction and commencement of commercial operations. Consequently, the private sector concession company was unable to meet its payments, and became insolvent within 6 months of opening, leaving the banks to step in. To the government's credit, it has resisted to date any temptation to buy the banks out.

The Brisbane Airtrain experienced similar difficulties but had a very different financing structure. In Brisbane, the 35 year concession was financed primarily with CPI indexed bonds with a maturity of 33 years. This structure also allowed for a back-ended repayment profile and substantial grace periods following completion of construction before both payment of interest and repayment of principal commenced. In addition, Brisbane Airtrain included substantially more equity capital than Sydney, with almost 30% equity allowing a greater buffer before problems were experienced in servicing debt.

Walker and Walker (2000) and Walker (2002) argue that the Sydney AirLink BOOT scheme demonstrates that risk transfer has been ineffective based on an initial feasibility study suggesting that the New South Wales government would receive a 2% rate of return (nominal breakeven of 23 years) on its investment while the private sector would receive a 21% rate of return (a 4 year nominal break-even). However, these commentators ignored the fact that the private sector was taking a higher proportion of the risk.

In fact, as a result of this risk transfer, the private sector equity incurred significant losses on the Sydney AirLink when passenger levels did not reach forecast levels. The result has been that this project has been a poor one from the private sector's perspective. However, from a public sector perspective the project has been a success in that had taxpayer dollars been used to fund the project, that money would have been lost. The use of the PPP model has saved the New South Wales government a significant amount of money

6.11. London Underground

In part as a reaction to a number of projects such as those in Melbourne, Sydney, and Brisbane where the private sector overestimated ridership, a major PPP transaction for the London Underground metro, closed in 2003, has retained ridership risk in the public sector with private sector concessionaires responsible for providing all of the infrastructure and services up to the point of delivery of the service to the end customer in return for performance-related payments from government. One such example is the Tubelines transaction, a landmark UK £6 billion program to upgrade and maintain the infrastructure of three lines of the London Underground network under a 30 year service contract. In order to ensure quality of service, the concession contract includes numerous performance criteria that influence the level of government payments.

Transactions such as that for London Underground are a growing trend, in which the government retains responsibility for establishing service levels and the risk of usage associated with those service levels. They are also a natural response to the tightening of private sector appetite for demand risk in transportation. However, the trend is worrying in concentrating on the contractual allocation of risk (through complex performance criteria and payment mechanisms and detailed legal documents) rather than transferring real ownership of the risk of providing a service, which users want at a price that they are prepared to pay.

Table 3
Advantages and disadvantages of private sector financing

Advantage	Disadvantage
Enables governments to undertake more projects without incurring more debt	Transaction costs (meaning only large-scale projects can be undertaken cost-effectively)
Achieves risk transfer away from taxpayers and cost certainty	Potential loss of control and/or flexibility for government
Delivers projects sooner	Complexity
Permits access to private sector innovation, expertise, and efficiency	Potential political costs (e.g. union backlash, perceived negative implications of "privatization")
Allows projects to be planned on a commercial whole-of-life basis (not just up front cost)	
Advances the 'user pays' principle to achieve balance between current and future generations	

7. Why use private sector finance?

There is a growing social need for improvements in transportation infrastructure. However, with government funds becoming increasingly more inaccessible, alternative methods for financing are necessary. Private sector finance is an effective way to deliver a project, as it has many benefits.

Can the benefits of PPPs in terms of lower-cost service provision and higher service quality be achieved without opening up monopolized transport markets to private sector finance and ownership? Stanley and Hensher (2003) contend that they can through negotiated performance-based approaches. From a bankers perspective we would argue that the transfer of financing risk, and even more the transfer of equity ownership risk, is essential to drive real risk transfer across other aspects of the deal and to create the incentives for private sector operators to deliver on lower costs and better service.

However, in determining whether the private sector financing of a transportation project is appropriate, the advantages and disadvantages listed in Table 3 should be weighed by governments and the public.

For the private sector, the advantage of private sector financing is clear: greater opportunities for development and investment in important transportation infrastructure. There are few disadvantages for the private sector in utilizing private sector finance. However, some design and construction companies can feel that the pursuit costs of bidding for privately financed projects are greater, and the business terms that they can achieve more onerous, than under a traditional government procurement contract.

8. Has private investment in transport infrastructure been successful?

The record of success of private sector involvement in the transport infrastructure sector is mixed. The transport sector has seen some of the biggest successes and some of the biggest failures in the PPP market place. In Australia, toll roads, in large part, have been successful investments for the private sector (e.g. the M4 in Sydney and the CityLink road in Melbourne), whereas passenger rail projects have not (e.g. the Brisbane Airtrain and Sydney Airport Rail link).

One of the strongest areas of success has been the Australian toll road sector. Standard & Poor's (2002) highlight this in a comparative study of 32 toll roads in the USA, UK, and Asia: " Relative to global trends, the privately funded toll-road projects in Australia have performed strongly and tend to lie at the lower end of the risk spectrum."

An example of a success story is the Hills Motorway Group that operates the M2 toll road in northern Sydney in Australia. Since opening in mid-1997, traffic levels on the M2 have increased by approximately 65%, and the road has become a key part of Sydney's road network. The road has benefited from the introduction of electronic tolling, which has increased the convenience for road users, increased traffic flow, and grown revenues. The road has significantly improved the flow of traffic in northern Sydney, and will become a key link in Sydney's ring network of roads. This has translated into daily revenues increasing by approximately 150% since opening and excellent growth in value for the company's public shareholders.

However, even the failures in transportation have a tendency to become successes over the longer term. For example, the Channel Tunnel, linking the UK and France, is currently doing well after a poor start. The Guangzhou–Shenzhen–Zhuhai Superhighway, a visionary project linking the three major cities of the Pearl River Delta, is another example of a project that experienced mixed financial success initially and grew substantially in value over time and through a series of refinancings. Long-term success is usually cold comfort to the initial investors, who will typically have lost their money long before the turnaround. The transport infrastructure sector may well be an industry in which it is better to be the second owner of many assets rather than the developer. Major financial restructurings are likely to be the norm. However, this is not an unusual feature of a number of industries that still attract large quantities of private financing. Development of the industry will inevitably progress in fits and starts as financiers become more conservative as a result of failures and less willing to take risk or provide substantial financing.

Table 4
Australian private road finance – evolving risk profile

Year	Project	Risks borne by the private sector	Funding mechanism
1986	Sydney Harbour Tunnel	Design Construction	Government guarantee
1988	Sydney M4 Motorway	Design Construction Traffic Tax	Bank debt Existing motorway
1990	Sydney M5 Motorway	Design Construction Traffic Tax	Bank debt Government contractor
1994	Sydney M2 Motorway	Design Construction Traffic Tax	Listed equity CPI bonds Bank debt
1995	Melbourne City Link	Design Construction Traffic Tax Competing rail system	Infrastructure bonds
1996	Eastern Distributor	Design Construction Traffic Tax Competing rail system	CPI bonds Bank debt Infrastructure bonds Subordinated debt

9. Trends in transportation finance

9.1. Evolving risk profile in privately funded road projects

The competitive pressure generated from embracing private sector delivery has incrementally improved deals for governments.

Since the initial private sector investment in the Sydney Harbour Tunnel in 1986, the risk profiles of subsequent road projects have evolved as governments, infrastructure providers and investors have become more familiar with the risks involved. Table 4 illustrates this.

Three additional toll roads are currently under development in Australia, the Western Sydney Orbital, the Cross City Tunnel, and the Lane Cove Tunnel, which are taking the financing model further. This is driven by the increasing transfer of risk into the private sector.

Blain and Wilkins (2003) describe almost the opposite process in effect in the indirect or shadow toll roads in the UK, known as DBFO roads, where the evolution of the payment mechanism by government has reduced the uncertainties associated with the payment stream, and therefore the risk taken by the private sector.

The lesson is that the private sector appetite for and understanding of risk can develop quickly, and financial markets can equally quickly evolve new financing instruments. This, however, depends on success. When projects fail from a financial perspective, markets can overreact. A current example is the difficulty that the private sector currently has in taking any material ridership risk on urban mass transit projects, having been over-aggressive on growth in the Melbourne public transport franchises, the Sydney airport rail link, and the Brisbane airport rail link.

9.2. Commercial models

There has been much discussion on whether a single commercial and institutional model is needed or appropriate. The consensus has been that it is not, and a case-by-case approach is the best. While we cannot argue with the logic of this, it does leave financial markets confused. We risk having the modern equivalent of the different gauges of the nineteenth century – a different commercial model in every state and in every segment of the industry. This will inevitably require a major rationalization at some stage in the future, either through government regulation or market consolidation.

9.3. Major barriers to private sector investment in public transport

There are a number of clearly identifiable barriers to private sector investment in public transport. The continued government dominance of ownership but lack of investment in public transport infrastructure simultaneously causes public transport owners to rely on government subsidies and a decline in their operating competitiveness, which is deterrent to major investment.

The range of different government interests in public transport, as well as the number of different regulatory agencies and procedures, creates a complex and confusing management web that deters private sector involvement. The very limited channels for private sector investment in infrastructure exacerbate this problem.

The government has a role to play in the provision and operation of public transport. It must, however, consider the imbalance of its rail public transport funding relative to government funding for roads. The government also needs to develop a firmer understanding of the commercial complexity of transport projects. It is vital that the government recognizes that the size and success of projects often depends on associated infrastructure developments.

10. Conclusions

A more effective understanding of the transport industry by both governments and financial markets is an essential base for financing support. Particularly, there needs to be an increased understanding of the link between capital investment decisions and the long-term operation of the project. This can best be achieved by ensuring that the private sector is engaged as early as possible in the planning of transportation projects, and is given maximum flexibility to propose ways to most efficiently design, build, operate, and finance large-scale transportation projects.

Financing needs to be targeted, and the benefits of asset-based financing need to be explored. Financing products will develop rapidly in response to new institutional models, and the current availability of financial products should not restrict PPP models. Failures should be an opportunity to learn and improve, although it needs to be recognized that private markets are likely to overreact to failures for a period of time. The long-term trends in transportation are, however, clear, with an increasing investment role for the private sector and a diminishing role for the government.

Despite political backlash at the time of writing, greater control of the sources of revenues (fares, tolls, freight charges, development revenues) in the private sector, subject to appropriate levels of regulation drawn from the experience of other industries, is the most likely route to generate innovative financing and successful outcomes. Creating the environment for and understanding of investment in public transport infrastructure will benefit both the government and the private sector. The end result will be an increasing availability of real equity and a public transport system that can be characterized not as "welfare on wheels" but as a commercially viable and sustainable system that transports contented commuters effectively and products efficiently.

References

Australian Bureau of Transport and Regional Economics (2003) *Australian Transport Statistics*, June. Canberra: Australian Bureau of Transport and Regional Economics.

Australian Department of Foreign Affairs and Trade (1998) *Asia's infrastructure in crisis – harnessing private enterprise*. Canberra: East Asia Analytical Unit.

Blain, R. and M. Wilkins (2003) "The evolution of DBFO payment mechanisms: one more for the road?" *Credit Survey of the UK private finance initiative and public private partnerships*, April. New York: Standard & Poor's.

Flyvberg, B., N. Bruzelius and W. Rothengatter (2003) *Megaprojects and risk: an anatomy of ambition,* Cambridge: Cambridge University Press.
Hensher, D. and P. Goodwin (2004) "Using values of travel time savings for toll roads: avoiding some common errors," *Transport Policy*, 11:171–182.
Hunter, A. (2000) "Financing passenger rolling stock in Victoria," in: *Euromoney yearbook*. London.
Macquarie Corporate Finance (2000) *Project finance: the guide to financing transport projects*, 2nd edn. Sydney: Macquarie Bank.
Macquarie Infrastructure Group (2003) *Annual report.* Sydney: Macquarie Bank
National Audit Office (2003) *PFI: construction performance – report by the Comptroller and Auditor General*. London: National Audit Office
NSW Government (2001) *Working with government, guidelines for privately financed projects*. Sydney: NSW Government.
Project Finance Magazine (2003) "SR125: the US PPP bridgehead?" *Project Finance Magazine*, May 30:241.
Standard & Poor's (2001) *Australian and New Zealand transportation: reliance on debt can derail credit stability*. New York: Standard & Poor's.
Standard & Poor's (2002) *Public finance criteria: tollroad and bridge revenue bonds and Australian toll road projects have investment grade potential*. New York: Standard & Poor's.
Stanley, J. and D. Hensher (2003) "Performance based contracts in public transportation: the Melbourne experience" in: *8th International Conference on Competition and Ownership in Land Passenger Transport (THREDBO 8)*. Rio de Janeiro.
Walker, R.G. (2002) "Public private partnerships: different perspectives," in: *The Australian Equities Investment Conference*. Sydney.
Walker, R.G. (2003) "Public private partnerships: form over substance?" *Australian Accounting Review*, 12.
Walker, R.G. and B.C. Walker (2000) *Privatisation: sell off or sell out? The Australian experience*. Sydney: Australian Broadcasting Corporation Books.
World Bank (1995) *World development report and infrastructure development in East Asia and Pacific: towards a new public private partnership*. Washington, DC: World Bank.

Chapter 19

FINANCIAL ANALYSIS: APPLICATIONS TO AUSTRALIAN TOLL ROAD ENTITIES

STEWART JONES
University of Sydney

1. Introduction

According to Hensher (see Chapter 6), "performance evaluation frameworks" transport entities in the public and private sectors are becoming increasingly involved in performance evaluation. Reasons include the aspiration to deliver services more efficiently and effectively as well as to better understand how the range of business activities contribute to overall organizational performance. He notes "The transport sector has come under particular scrutiny in recent years as governments, in their role as regulator and service supplier, seek to engender increased performance and productivity from all transport sectors – road, rail, shipping, ports, airports, airlines, trucking, bus, and coach." Hensher emphasizes the importance of performance measures in a strategic holistic sense. For performance measures to be useful in a strategic sense, they must encapsulate the integrated nature of business "to truly replicate their competitive strength." Hensher believes that it is not inputs, outputs, and singular success factors *per se* that are critical to performance evaluation, but rather how inputs fit together in the development of an overall strategic framework. While Hensher develops a qualitative framework for performance evaluation, this chapter introduces quantitative financial techniques for assessing the performance of toll road entities in Australia. The techniques explained in this chapter are critical for assessing different dimensions of the performance of the toll road industry from the perspective of investors, lenders, and other users. However, it is acknowledged that these measures by themselves only provide partial indications of performance as they do not describe the overall performance of an organization and its links to the organization's strategic, tactical, and operational (STO) setting.

Toll road entities are classified by the Australian Stock Exchange (ASX) as a subgroup of infrastructure and utility industry. In drawing attention to financial analysis techniques, this chapter compares the performance of two major Australian toll road entities, Transurban (TCL) and the Hills Motorway Group (HLY).

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

Financial analysts make extensive use of company financial statements in assessing the financial position and performance of listed companies. Information disclosed in these financial statements are also used by analysts to assess the relative value of a firm's shares on issue. The main interest here is to determine whether a firm's shares are mispriced – either overvalued or undervalued based on the underlying performance of the firm. Underpriced shares may represent buying opportunities for investors, whereas overpriced shares may signal "sell" decisions. This approach to stock price valuation is often referred to as fundamental or value analysis. While analysis of financial statements forms a key aspect of fundamental analysis, it does not constitute the sole source of information. Fundamental information can include a variety of firm-specific, industry-wide, and economy-wide information, such as general economic conditions, the impact of changes in government and regulations, and individual risk factors associated with specific industries (Foster, 1986). However, for the purposes of this chapter, this analysis is largely confined to financial analysis of company financial statements.

2. Regulations governing financial disclosure

Company financial statements in Australia (and throughout the world) are governed by a wide range of accounting regulations, forming the rules and principles that govern the structure, content, audit, and disclosure of financial information. It is important to understand these regulatory influences on financial disclosure to have a better appreciation of the "raw data" that forms the basis of financial analysis. In Australia, there are essentially three main sources of accounting regulation: the requirements of companies legislation; the Australian stock exchange listing rules; and accounting standards issued by the Australian Accounting Standards Board (AASB).

(1) *Requirements of the Corporations Act (2001).* The relevant legislation governing financial disclosure in Australia is known as the Corporations Act (2001). The Act is the supreme regulatory authority over companies because it is a statute passed by the federal parliament. The Corporations Act is a lengthy statute covering a range of issues relating to the formation, conduct, administration, and winding-up of companies. Among other things, the Corporations Act (Schedule 5) requires company directors to present to shareholders at the annual general meeting a statement of financial performance (or "profit and loss statement"), a statement of financial position (or "balance sheet"), and a statement of cash flows. Such financial statements are also required to be externally audited. These financial statements form a critical part of financial analysis, as numerous

performance measures used to assess companies are drawn directly from these published financial statements.[a]

(2) *Stock exchange listing rules*. Prior to 1997 the Australian Stock Exchange (ASX) was a non-profit, private sector organization comprising over 100 stockbroker corporations and partnerships. The ASX was formed in 1997 following the amalgamation of the six state-based exchanges and demutualization. Its main function is to provide a market for the trading, clearing, and settlement of equities and derivatives. It also provides market data relating to listed companies and indexes. The conduct of the ASX and its members is governed by the Corporations Act. Within this legal framework, the ASX operates as a self-regulatory body with its own comprehensive listing requirements. The ASX is primarily concerned with protection of investors as well as the promotion of an efficient and ethical share market. This orientation is reflected in the ASX's concern with timely financial disclosures by listed companies. The ASX regulates corporate financial disclosure through its Listing Rules.

Companies seeking to list on the stock exchange must comply with the financial disclosure requirements of the ASX and other listing requirements. These requirements, like those of the Corporations Act, have traditionally been concerned with disclosure rather than with technical accounting issues, such as the different ways to measure and classify transactions in the financial statements. Australian accounting standards provide the detailed rules for measuring and classifying accounting transactions. Under ASX Listing Rules, firms are required to disclose financial statements on an annual basis (Listing Rules 4.3–4.7A) and on a semi-annual basis (Listing Rules 4.1–4.2C). There are further rules for firms in the "new economy" (such as Internet and high-technology firms) and mining entities to disclose information on a quarterly basis (Listing Rule 4.7B). It is also noted that "Signal G" company announcements are required under Listing Rule 3.1 of the ASX, which sets out continuous disclosure requirements that Australian public companies must satisfy. ASX Listing Rules relating to continuous disclosure are enforced by Section 674 of the Corporations Act. Continuous disclosure is the timely release of company information to keep the market informed of events and developments as they occur.

(3) *Accounting standards*. Accounting standards provide detailed rules on how particular types of financial transactions and other events should be dealt with in an entity's financial statements. Accounting standards cover all

[a]Schedule 5 has now been subsumed under a new accounting standard, AASB 1034. The nature and role of accounting standards is discussed shortly.

manner of accounting transactions, such as accounting for inventories held by firms; accounting for research and development costs; foreign current transactions; leases; employee entitlements; valuation of non-current assets; accounting policy disclosures; and measurement basis for non-current assets, among many other transactions and disclosures. Accounting standards help to ensure that an entity's statement of financial performance, statement of financial position, and statement of cash flows present a "true and fair" representation of its performance and financial position. Accounting standards in Australia are issued by the Australian Accounting Standards Board (AASB), with oversight from the Financial Reporting Council (FRC), which is directly responsible to the federal government. Accounting standards that are issued by the AASB are known as Approved Australian Accounting Standards. The AASB is given authority under the Corporations Act to issue accounting standards that are legally binding on all publicly listed companies in Australia.

Having explained the major sources of regulation, it is clear that three primary financial statements feature prominently in the Corporations Act, the ASX Listing Rules and Approved Australian Accounting Standards. We now briefly explain each of these financial statements, as they form the basis of the analysis of TCL and HLY provided in this study (see the Appendix for financial statement summaries of these entities).

3. The primary financial statements used in financial analysis

3.1. The statement of financial position

This statement summarizes the different classes of assets, liabilities, and shareholder's equity of a firm. The difference between a firm's assets and the liabilities is called the net assets. Net assets represent the claim shareholders (who are the owners of public companies). The statement of financial position presents its information in conformity with the following identity: assets – liabilities = shareholders' equity. The statement of financial position is used extensively by analysts because it provides critical information useful for decision-making. Most importantly, the statement presents information about a firm's wealth or control over economic resources (the assets of the firm); financial structure, or the relationships between assets, liabilities, and equity (e.g. firms with high levels of debt relative to total equity are generally regarded as more risky than firm's with smaller ratios); capacity for adaptation, or the capacity of entities to modify or change the composition of its resources under its control to meet changing market conditions; and liquidity, or the availability of assets to meet financial commitments.[a] For instance, if total liabilities exceed its total assets, a firm is said to have "negative" shareholder's wealth. A disproportionate level of liabilities to assets usually signals liquidity problems for a firm.

As can be seen in the Appendix, it is customary for assets and liabilities to be classified as "current" and "non-current" in the statement of financial position. This is also required by legislation and by accounting standards (discussed below). For example, Approved Australian Accounting Standard AASB 1040 mandates the current and non-current distinction for assets and liabilities. The standard defines a current asset as assets that are sold, consumed, or realized within 12 months of the reporting date (the date when financial statements are prepared), or in the normal course of an entity's operating cycle. For example, accounts receivables are rights to receive cash from customers following the sale of goods and services. Current assets may also include assets that are held for trading purposes (such as short-term investments or inventory held for resale) or assets that are cash or cash equivalents (such as cash at bank and deposits). Likewise, current liabilities are defined by AASB 1040 as assets that are expected to be settled within 12 months of the reporting date or in the normal course of the entity's operating cycle. Examples include short-term interest-bearing liabilities, such as loans, accounts payables, and tax liabilities. Non-current assets are all assets other than current assets, such as property, plant and equipment, and intangible assets. Non-current liabilities are all liabilities other than current liabilities, and include items such as long-term debt and provision accounts. The current/non-current distinction is important for assessing the financial position, particularly the financial structure and liquidity of firms. For example, the relationship between current (or short-term) assets to short-term liabilities is an important indication of a firm's short-term liquidity.

3.2. The statement of financial performance

This statement reports the revenues, expenses, and net profit (or the surplus of revenues over expenses) that a firm has made during a given period. The statement shows how an entity's revenues were obtained and how the expenses were incurred. As the title of the statement suggestions, this document provides an important indication of the operating performance of the entity. Disclosures in the statement of financial performance are governed by AASB 1018, "Statement of Financial Performance." The profit is what remains after the expenses of operations have been deducted from the revenue earned. Sales revenue

[a]See Statement of Accounting Concepts No. 2, *Objectives of General Purpose Financial Reporting*, AARF, Melbourne, August, 1990.

represents amounts earned during the reporting period, through the sale of goods and services. Expenses represent outflows of resources that were necessary in order to earn the revenue for the period. They include items such as depreciation and amortization of non-current assets, bad debts expense, wages and salaries, and interest expense. The net profit is the difference between revenues and expenses, and is the net surplus from ordinary business activities during the period. This surplus increases the net assets of the entity, and it therefore increases the shareholders' stake in the enterprise. AASB 1018 requires a number of separate disclosures in the statement of financial performance, including all sources of revenues and expenses from ordinary operations, profit and loss from ordinary activities before income tax expense, income tax expense, net profit after tax, and profit and loss from extraordinary items. "Extraordinary items" are items of revenue and expense that are outside the ordinary or regular operations of a firm and are not of a recurring nature (such as loss of capital equipment through fire or flood). Hence, analysts usually disregard extraordinary items from the analysis of operating performance – they are items that appear "below" the profit line rather than "above" the line that is the focus of analyst attention.

3.3. *The statement of cash flows*

The statement of cash flows was introduced in Australia in 1992, when AASB 1026, "Statement of Cash Flows," was released. AASB 1026 is unique in the world as it requires the direct method of reporting cash flows. This method requires detailed disclosures for all major classes of cash receipts and payments, whereas the indirect method (preferred in the USA and UK) only provides an estimate of operating cash flow by adjusting net income for working capital accounts (Jones and Hensher, 2003). Hence, the indirect method does not provide the same level of detailed disclosures needed to assess a firm's operating performance. The usefulness of detailed cash flow disclosures (particularly the direct method of reporting operating cash flow) in economic decision-making has now been extensively demonstrated (Lee, 1986; Livnat and Zarowin, 1990; Neill et al., 1991).

This statement of cash flows summarizes all the cash receipts and payments of an entity, which are then classified into three major categories: operating, financing, and investing activities (see statement of cash flows in the Appendix). Operating cash flows relate to cash flows generated primarily from the sale of a firm's goods and services. Analysts generally regard it as a crucial measure for evaluating a firm's sustainable operations, as well as overall operating performance. Investing cash flows relate to the disposal and/or acquisition of non-current assets, such as property, plant and equipment, and investments. Financing cash flows relate to changing the size and/or composition of the financial structure of the entity, including equity issues and borrowings. The introduction of cash flow statements was an important development in Australia, as it was in the USA (Heath, 1978) and UK (Lee, 1986). The importance of cash flow statements has been closely associated with certain limitations with the statement of financial performance and the statement of financial position – both of which are prepared on the "accrual" basis of accounting. Under the accrual basis, revenues (expenses) are recorded when there is reasonable certainty of receiving (or dispensing) cash, not necessarily when the cash comes in (or out). Furthermore, the accrual basis of accounting involves a number of allocations and deferrals that, while affecting the determination of profit, do not involve any cash outlays by the firm. For example, depreciation expense is the allocation of the cost of an asset over its useful life. As an expense item, it affects the determination of profit, but has no impact on the cash flows. Allocations such as depreciation can also be subjective, as they rely on numerous assumptions relating to the useful life of the asset as well as the method of depreciation adopted. As a result, cash flows are now widely regarded as a more objective measure of performance than earnings figures, because they are less subject to managerial manipulation or subjective estimate. Cash flows are also regarded as simpler and more interpretable than "accrual"-based numbers for the same reasons (e.g. Rappaport, 1982; Healy, 1985; DeAngelo, 1986; Neill et al., 1991).

4. Background: Transurban and the Hills Motorway Group

The financial statements for TCL and HLY, for the period 2000–2002, are provided in the Appendix. Financial statement information includes a statement of cash flows, a summary statement of financial position, and a statement of financial performance. The Appendix provides 3 years of financial data for both firms, so that an assessment of the trend in financial performance can be determined. To assist the analysis, each financial statement item in the Appendix is numbered numerically, and will be used in the computational exercises. A brief background to TCL and HLY for the year 2002 – applicable to the financial data to be considered in this chapter – is provided in Boxes 1 and 2, and a detailed financial analysis follows.[a]

5. Financial analysis of TCL and HLY

Financial statement analysis includes the analysis of financial statement variables at a point in time (cross-sectional analysis) and relationships over time, which is

[a]Source for the background material comes from Aspect Financial Pty Ltd.

Box 1
Background to HLY

HLY is involved in the construction, financial management, and ongoing operations of the M2 motorway in Sydney, Australia, which provides a fast and efficient link for Sydney's rapidly growing North West sector. The motorway was officially opened for traffic on May 26, 1997. Average daily traffic for the M2 is between 60 000 and 70 000 vehicles. In early February 2003 the HLY introduced a new CEN-compliant electronic tolling system (ETC) that is interoperable with all other Sydney toll roads and the Sydney Harbour Bridge and Tunnel. HLY has achieved improvements in traffic flows, particularly through the main toll plaza, since the introduction of the new system, and further efficiencies are expected as motorists become familiar with the dedicated ETC lanes. HLY posted an improved performance for the year to June 2002 with an after tax loss of Aus. $1.5 million compared with an Aus. $11.4 million loss in 2002. Toll revenue was up 8.5% over the period to Aus. $65 million, despite a slight drop in average daily traffic from 63 822 vehicles in 2001 to 63 345 in 2002. Hence, increases in traffic tolls did not have a significant impact on revenues. The drop in traffic growth was attributed to a price increase in the first half, which appears to have been absorbed by motorists. Traffic growth was up 3.4% in the 6 months to June compared with the previous year. Two infrastructure projects on the horizon (M2 eastern link and the Western Sydney Orbital) have the potential to increase traffic volumes in the medium term. Construction of the Western Sydney Orbital, which will connect to the western end of the M2, has commenced, and the project is forecast to be completed by August 2006. Four consortia have lodged tenders for the construction of a link between the eastern end of the M2 and the Gore Hill Freeway, to be known as the Lane Cove East Tunnel project. It is expected that the Road Traffic Authority will select a shortlist of consortia before the end of the financial year; however, award to the successful proponent is not anticipated until mid-2004. HLY is a participant in the TunnelLink consortium; however, it appears that the company is not exposed to any construction or financial risks of the TunnelLink bid. Revenues for the M2 are likely to increase, with plans to implement a toll increase for Class 1 vehicles at the main toll plaza, North Ryde, on July 1, 2003. The toll at the Main Toll Plaza will increase from Aus. $3.30 to Aus. $3.80 (including GST) for all cars, motorcycles, vehicles with two axles under 2.8 m high, and vehicles with three axles under 2.0 m high.

Source: HLY 2002 annual report.

often referred to as trend analysis or time series analysis (Foster, 1986). Both techniques are widely used in the financial analysis of companies. Two useful cross-sectional techniques of financial statement analysis are common-size statements and financial ratio analysis.

Common-sizing is a useful analytical technique because it can deal with the problems associated with firms of unequal size. For example, referring to item 51 in the Appendix, it can be seen that TCL reports Aus. $1.628 billion in total debt commitments in 2002, whereas HLY only reports Aus. $423.8 million in the same year. Can it be concluded that TCL is more indebted than HLY? On closer analysis, it would be misleading to draw this conclusion because TCL is a much larger firm in financial and market capitalization terms than HLY. One way of dealing with size differences is to express the components of the balance sheet as a percentage of total assets (and the components of the income statement as a percentage of total sales). The reconstructed statements are termed common-size

Box 2
Background to TCL

TCL holds a concession granted by the Victorian government in Australia for the operation and maintenance of the Melbourne City Link. TCL has been a pioneer in the development and operation of advanced electronic toll roads. It has become a leader in the global market for intelligent transport solutions. TCL both developed and operates the electronic tolling on CityLink, a 22 km urban motorway in the Australian city of Melbourne. TCL will operate the toll road for 25–34 years, at which time ownership will revert to the government. Over 650 000 electronic toll transactions are processed daily on CityLink. The complex technology and systems integration needed to achieve this have been central to CityLink's success and a key component of TCL's expertise. TCL listed on the Australian Stock Exchange in 1996, and rapidly climbed to its current position as a Top 100 company on the exchange. TCL's shareholder base includes financial institutions and superannuation funds. In August 2002, TCL had an enterprise value of Aus. $4 billion and a Standard & Poor's credit rating of A–. Until recently, TCL has been a "single-purpose entity," legally restricted to operating the Melbourne CityLink project. The company's latest project is the 39 km Western Sydney Orbital (WSO). When completed the WSO will be Australia's longest electronically tolled road. As a partner in the WestLink consortium chosen to develop the WSO, TCL will also deliver and operate the new road's tolling system. TCL is a 40% equity partner in the consortium, with the remainder owned by the Macquarie Infrastructure Group (40%), Leighton Holdings (10%), and Abigroup (10%). The WSO will be a 40 km, four-lane motorway linking the M5, M4, and M2, and will form the final link in Sydney's orbital road network. Tolls will be collected electronically at the estimated rate of Aus. $0.25 per km up to a cap of Aus. $5.00, with a concession period of 34 years. Construction will commence following the finalization of financial agreements with the Road Traffic Authority. Closure is expected by December, with work to begin in the new year. The motorway is expected to be completed by 2007. TCL is expecting to fund its equity investment in the project via the issue of reset preference securities.

Source: TCL 2002 annual report.

statements. By dividing TCL's debt by its total assets in 2000, (item 51)/(item 47), we get 40.4%. Whereas the same ratio for HLY is 48.8%. Hence, despite the absolute differences in the size of debt, HLY is more indebted than TCL relative to the size of its asset base.

The most widely used cross-sectional technique is a comparison of financial ratios across firms. Financial ratios also control for difference in size between firms. Common-sizing reduces all balance sheet and profit and loss components to ratios of total assets and total revenues; however, financial ratio analysis is not restricted to using total assets and total revenues in the denominator. Numerous individual ratios have been proposed in the literature. The following seven categories of ratios (and their variants) are the more commonly used ratios by expert analysts. There are several broad categories of analysis, which provides a useful framework to evaluate TCL and HLY. Key areas of analysis include: profitability, cash flow, cash position, liquidity, capital structure, debt service coverage, turnover, volatility, fundamental valuation, and investment returns (for more detailed discussion, see Foster, 1986; White et al., 1997). We will briefly compare TCL and HLY on all major categories to make some determinations of the relative performance of each company over the period 2000–2002.

In making determinations about individual company performance, it is important to compare the firm's performance against its peer group or industry. Analyzing a firm's performance in isolation from industry benchmarks can be misleading. For instance, a firm's debt or liquidity position may seem reasonable in the initial analysis, but prove quite weak compared with the performance of the industry as a whole. From an investment point of view, firms that are underperforming relative to industry performance benchmarks are less desirable than firms that are performing well. In calculating industry averages, all firms in the industry group are averaged across each indicator under examination. To get a valid measure of central tendency extreme observations or outliers are removed from the sample. For the purposes of this study, observations more than four standard deviations from the mean have been removed.

6. Financial ratio analysis

6.1. Assessing profitability

Profitability refers to the ability of a firm to generate revenues in excess of expenses. Investors are concerned with a firm's capacity to generate sustainable profit growth. In assessing the sustainability of profit performance, the analyst is particularly concerned with operating revenues and expenses – these are the revenues and expenses generated and incurred from the ordinary operations of a firm through the sale of its goods and services. When making comparisons across firms or over time, it is useful to compare a firm's profits levels with its total asset base and shareholders' equity. This provides a measure of managerial efficiency in using assets and shareholders' funds to generate profits. The following ratios illustrate alternative, though not exhaustive, ways of measuring a firm's relative profitability:

(a) $$\text{EBIT to total assets} = \frac{\text{EBIT (line 28)}}{\text{total assets (line 47)}},$$

(b) $$\text{rate of return on equity} = \frac{\text{net income after tax (line 36)}}{\text{shareholders' equity (line 55)}},$$

(c) $$\text{rate of return on assets} = \frac{\text{net income after tax (line 36)}}{\text{total assets (line 47)}}.$$

The ratio of EBIT to total asset is one of the most important profitability measures to use in financial analysis. EBIT is regarded as a better indicator of operating performance because it ignores the impact of capital structure and taxation issues that can vary substantially between firms and industries. He notes: "this ratio is a measure of the true productivity of the firm's assets, independent of any tax or leverage factors." For instance, a firm with higher debt levels is likely to have higher interest payments, which affects the computation of net profit after tax. However, interest payment on debt is a finance expense that does not relate to operating activity – hence, differences in debt levels between firms can distort evaluations of operating performance. EBIT can be contrasted with EBITDA (earnings before interest, taxation, depreciation, and amortization). EBITDA has become an increasingly popular measure in Australia because it brings the earning figure closer to the true cash flows of the entity (depreciation and amortization are non-cash expense items). Either EBIT or EBITDA can be used in profitability analysis.

The second ratio (net income to shareholders' equity ratio, also called the return on equity ratio) is also important because it measures the efficiency with which shareholders' equity is being employed within the firm (Foster, 1986). The net income-to-total assets ratio (also called the rate of return on assets or capital) measures the efficiency with which the asset base is employed by the firm. It is clear that the higher each of these ratios the more profitable the firm is, in a relative sense. The numerator of the ratios is the net profit after tax, which is available to ordinary shareholders – hence, we have to deduct any preference dividends in the computation (neither HLY nor TCL pay preference dividends at the moment, so this item can be ignored for now).

(a) EBIT to total assets (%):

Year:	2000	2001	2002
HLY	0.49	0.75	1.77
TCL	–4.07	–0.71	0.68
Industry average	7.19	0.29	–3.32

(b) Rate of return on equity (%):

Year:	2000	2001	2002
HLY	–3.38	–2.77	–0.41
TCL	–103.22	–95.90	–3.25
Industry average	1.75	–4.08	–2.45

(c) Return on assets (%):

Year:	2000	2001	2002
HLY	–1.65	–1.28	–0.18
TCL	–5.58	–5.83	–1.67
Industry average	5.03	–1.14	–5.32

Analysis. Overall HLY appears to be relatively more profitable than TCL across all three ratio measures. However, HLY is still not a particularly profitable company if rate of return on equity or assets is used (both measures were negative in 2002). TCL shows the strongest improving trend, despite its rate of return on assets and equity being negative. It is noteworthy that the overall profitability of the industry group has been negative over the preceding 2 years.

Having made these observations, the unique business nature of toll road entities needs to be considered carefully. Like many infrastructure businesses, they are long-term in nature and highly capital-intensive. Both TCL and HLY have made large capital outlays, and analysts expect that significant accounting losses will be reported in the earlier years of the life of both enterprises. In fact, most analysts do not think either HLY or TCL will become profitable for many years. As can be seen from item 29, a large portion of the loss for both concerns are "non-cash" depreciation and amortization write-offs, and, in the case of HLY, there was also a one-off company reconstruction cost. TCL also reported a number of significant one-off losses, including a payment of Aus. $235.5 million arising from a major debt refinancing, a payment of Aus. $153.6 million to settle an outstanding legal settlement arising from the CitiLink contract with Transfield Obayashi Joint Venture, and a change in accounting policy on valuation of its concession note, resulting in a write-off of a further Aus. $58.6 million. Because these items are largely one-off losses, they are not so important in assessing the continuing operating performance of TCL.

Given the nature of infrastructure businesses, it is clear that profitability analysis can only provide limited insight into the performance of HLY and TCL in the earlier stages. However, in the 2002 annual reports of both companies, the directors of HLY and TCL were quick to stress the importance of operating cash flows as a key indicator of their firm's performance. We now turn to this measure.

6.2. Assessing cash flow from operations

As stated previously, operating cash flow is now regarded as a crucial measure of financial performance. Under AASB 1026, "Statement of Cash Flows," all listed companies, including HLY and TCL, must disclose their operating cash flow results. While HLY and TCL have reported negative earnings, they can only continue to survive if they generate positive operating cash flow. Ultimately, a firm pays its debt commitments and operating expenses (such as wages and salaries) from its cash resources, not from "profit."

Two important ratios that indicate cash generating ability are:

(a) $\text{net operating cash flows to sales} = \dfrac{\text{net operating cash flows (line 8)}}{\text{sales (line 26)}}$,

(b) $\text{net operating cash flows to total assets} = \dfrac{\text{net operating cash flows (line 8)}}{\text{total assets (line 47)}}$.

The higher each of the ratios, the larger the cash flow generated by the firm in its operations.

(a) Net operating cash flows to sales (%):

Year:	2000	2001	2002
HLY	77.8857	86.0613	86.1580
TCL	–444.875	44.4398	34.4598
Industry average	–12.7297	13.0450	31.2942

(b) Net operating cash flow to total assets (%):

Year:	2000	2001	2002
HLY	5.0977	6.1323	7.0842
TCL	–4.8658	3.4490	1.7374
Industry average	1.8996	0.2505	–0.2367

Analysis. HLY has the stronger net operating cash flow performance relative to TCL, whereas TLC shows a deteriorating trend over time. Despite a decline in HLY's traffic levels (following recent toll rate increases) the company has maintained a stable cash flow position. TCL's poorer operating cash flow results can be attributed to many potential factors. It could be the result of poor sales performance, or a failure of the firm to collect cash on its receivables. It could also be a result of increasing operating expenses that outstrip revenues generated from sales. In the case of TCL, its total sales revenue (item 1) declined from Aus. \$152.334 million to Aus. \$114.916 million from 2001 to 2002, which had the affect of reducing the ratio of cash flow to sales. This might have resulted from loss of traffic following increases in toll rates (the Sydney M2 motorway appears to be less sensitive to toll rate increases than CityLink). While net operating cash flows actually improved (mainly because operating expenses declined), the increase in TCL's asset base nearly doubled over the same period, which lowered the ratio of cash flow to total assets. It is also noted that HLY's performance is much higher than the industry average, and its cash flow generating performance has been relatively stable over the last 3 years.

Analysts are also concerned with the free cash flow performance of firms. Free cash flow is gross operating cash flow less gross investment. Free cash flow is often viewed as the discretionary expenditure capacity of firms, for example to pay dividends and finance expansion of operations (such as the development of new infrastructure projects). To facilitate comparisons between firms, free cash flow is often expressed in per share terms. To achieve this, we must divide the free cash flow by the total number of shares outstanding:

$$\text{free cash flow} = \frac{\text{funds from operations (line 3)} - \text{gross investment (line 8)}}{\text{shares outstanding (line 62)}}.$$

The higher the amount of free cash flow per share, the stronger is the free cash flow position of the firm.

Free cash flow per share (Aus. $):

Year:	2000	2001	2002
HLY	0.25	0.34	0.30
TCL	–2.9	0.16	0.093
Industry average	–0.854	–0.341	0.38

Analysis. In 2002, HLY had a free cash flow of Aus. $0.30 per share, whereas TCL only had a free cash flow position of Aus. $0.093. The industry average is Aus. $0.38 per share. TCL again reveals deteriorating performance on this measure since 2000, despite a significant decline in its capital expenditure levels. A poor free cash flow position can restrict the capacity for toll road entities such as HLY or TCL to finance new infrastructure projects, maintain existing ones or pay dividends. It is noteworthy that only HLY is paying a dividend to its shareholders in 2002, whereas TCL paid no dividends to its shareholder up until June 2000.

6.3. Assessing cash position

The cash position of a firm is measured by its immediate cash reserves, which includes cash at bank and deposits, and highly liquid investments that can be quickly converted into cash. Cash and liquid investments represent an important source of funds that can be used to meet a firm's operating commitments as they fall due. Note, however, that cash position is not the same as cash flow – the latter is the cash generated from operating activity. Cash position is affected not just by operating activities but by financing and investing activities as well. Low cash reserve ratios could signal potential financial distress, as all firms need to maintain an acceptable level of cash reserves to continue its day to day operations. Two commonly used ratios are:

(a) cash position to current liabilities $= \dfrac{\text{cash (line 37)} + \text{current investments (line 38)}}{\text{current liabilities (line 48)}}$,

(b) cash position to total assets $= \dfrac{\text{cash (line 37)} + \text{current investments (line 38)}}{\text{total assets (line 47)}}$.

The first ratio shows the amount of cash reserves for each dollar of current liabilities. The second ratio shows the amount of cash reserves for each dollar total assets. The higher these ratios, the higher the cash reserves of the firm.

(a) Cash position to current liabilities (%):

Year:	2000	2001	2002
HLY	155.1144	159.2582	163.2168
TCL	20.4952	49.0899	67.9805
Industry average	152.3343	62.2514	84.9700

(b) Cash position to total assets (%):

Year:	2000	2001	2002
HLY	4.2289	3.6344	4.5836
TCL	0.8556	5.3309	3.2768
Industry average	14.4112	12.9814	15.2462

Analysis. It can be seen from the tables that HLY has a stronger cash position than TCL relative to current liabilities. HLY is also doing significantly better than the industry average on this ratio. While TCL's trend is looking stronger than HLY, it is still underperforming the industry norm and HLY on the cash to current liabilities ratio. HLY also has a stronger cash position ratio than TCL relative to total assets, though both companies are significantly below the industry average.

This could signal that both companies may need to consider the adequacy of their current cash reserves relative to the industry benchmark.

6.4. *Assessing liquidity*

Liquidity refers to the ability of a firm to meet its short-term financial obligations when and as they fall due. The cash position ratios discussed above captures only one dimension of liquidity. Two further liquidity ratios that are frequently used are:

$$\text{(a)}\quad \text{current ratio} = \frac{\text{current assets (line 41)}}{\text{current liabilities (line 48)}},$$

$$\text{(b)}\quad \text{quick ratio} = \frac{\left(\begin{array}{l}\text{cash (line 37)} - \text{current investments (line 38)} + \\ \text{account receivables (line 39)}\end{array}\right)}{\text{current liabilities (line 48)}}.$$

Both ratios look at the ratio of short-term assets to short-term commitments. The current ratio indicates how many times each dollar of liabilities is covered by every dollar of current assets. Analysts look for a good "safety buffer" between current assets and current liabilities. While there are no hard and fast rules on what an ideal current ratio should be, a ratio of 2:1 is often regarded as an acceptable level of firm liquidity by analysts. The current ratio differs from the quick ratio in that the numerator includes inventories and prepaid expenses (i.e. all current assets). The quick ratio is often regarded by analysts as a better measure of short-term liquidity than the current ratio because firms can experience difficulty converting inventory quickly into cash, and prepaid expenses usually cannot be converted to cash at all (although both these items represent "current assets" to the firm). In the case of HLY and TCL, the difference between the two ratios should not be that great as toll roads typically do not hold inventory.

(a) Quick ratio:

Year:	2000	2001	2002
HLY	1.62	1.74	1.77
TCL	0.31	0.59	0.83
Industry average	3.82	1.9	4.37

(b) Current ratio:

Year:	2000	2001	2002
HLY	1.9540	2.1764	2.1424
TCL	0.3196	0.5930	0.8430
Industry average	4.6604	2.6551	5.1225

Analysis. While TCL's trend on the quick ratio and the current ratio is improving, it is still well behind HLY and the industry average. Both ratios for TCL indicate that current liabilities actually exceed current assets, which signals a weak short-term liquidity position. HLY has a significantly better liquidity position, though its performance is still below the industry average.

6.5. Assessing capital structure

Capital structure ratios provide insight into the extent to which loan or equity capital is used to finance the firm's assets. A firm can raise finance from two principal sources – by issuing shares or by incurring debt (Frino et al., 2001). Debt incurs interest charges, whereas shareholders expect to receive dividends and/or capital gains in the share price. Analysts regard firm's with high proportions of debt relative to shareholders equity as higher risks than firms with lower ratios. Why? The higher the proportion of debt incurred by a firm, the greater the degree of control that can be potentially exercised by outside interests such as a bank. Higher debt levels also imply higher debt repayment burdens, which, in extreme cases, can exhaust the cash flows of a firm to the point of financial collapse. Interest payments are fixed commitments, and must be paid at specified dates according to contract. Failure to make an interest payment can result in a bank calling in the loan. However, a firm has more discretion in the payment of dividends to shareholders, even though reductions or elimination of dividends can still have very detrimental consequences to a firm's share price. It should be noted that the debt to equity ratios of different industries can vary greatly, hence it is important to examine industry averages closely when evaluating individual firm capital structure ratios.

Two representative ratios are:

(a) $\text{total debt to equity} = \dfrac{\text{total debt (line 51)}}{\text{shareholders' equity (line 55)}},$

(b) $\text{total liabilities to total equity} = \dfrac{\text{total liabilities (line 53)}}{\text{shareholders' equity (line 55)}}.$

The second of these ratios includes all liabilities of the firm, whereas the debt to equity ratio only includes total interest-bearing debt. The higher these ratios, the greater is the financial risk of the firm.

(a) Total debt to equity (%):

Year:	2000	2001	2002
HLY	92.5300	102.1300	112.9000
TCL	802.3500	1266.580	78.7100
Industry average	77.6346	54.8875	64.6495

(b) Total liabilities to total equity (%):

Year:	2000	2001	2002
HLY	104.66	115.96	131.24
TCL	–1948.58	1546.00	94.84
Industry average	109.20	–78.21	31.20

Analysis. Many analysts regard a debt to equity ratio of around 50% as being the "acceptable" level of debt that any firm should hold in order to maintain an adequate level of financial safety. In 2002, HLY had the more highly leveraged balance sheet, with a debt to equity ratio of almost 113%. This is nearly double the industry average. TCL had only recently improved its capital structure ratios, and is in a relatively stronger position than HLY when it comes to risk associated with indebtedness. However, it should be noted that both companies are well above the industry average on both ratios.

6.6. Assessing debt servicing capability

Debt service coverage provides an indication of the ability of an entity to service its debt commitment from earnings (or cash flows, if the cash flow cover ratio is used). The annual interest payments component of both financial ratios is calculated as total interest payments made to debt providers (notwithstanding whether the borrower expenses or capitalizes interest payments, i.e. records interest expense as an asset in the statement of financial position).

Two ratios useful in making inferences about coverage are:

(a) $$\text{interest cover} = \frac{\text{EBIT (line 31)}}{\text{net interest payments (line 32)}},$$

$$\text{(b)} \quad \text{cash flow cover} = \frac{\text{net operating cash flows (line 8)}}{\text{net interest payments (line 32)}}.$$

Because analysts have moved more to a cash flow basis in financial analysis, cash flow cover is now an important measure for assessing debt servicing capacity. The higher these ratios, the greater the ability of the firm to service interest payments on its debt commitments. Again, while there are no hard and fast rules on what the ideal ratio should be for interest cover in practice, a ratio of around three times is usually considered a satisfactory level of interest cover by analysts.

(a) Interest cover:

Year:	2000	2001	2002
HLY	0.37	0.40	1.07
TCL	–3.37	–0.20	0.62
Industry average	7.26	–3.25	0.41

(b) Cash flow cover:

Year:	2000	2001	2002
HLY	2.80	1.78	2.77
TCL	–3.00	0.00	1.22
Industry average	–1.4	–7.23	–1.19

Analysis. While HLY has significantly more indebtedness than TCL, the firm appears to have a greater capacity to service its debt repayments, whether using the interest cover or cash cover basis. Both ratios for HLY are higher than the industry average, though the EBIT cover ratio (barely 1) is not particularly strong, which reflects generally weak earnings performance of HLY. TCL is the weakest performer on both ratios, which reflects weak cash flow and earnings performance, coupled with quite high debt levels.

Another useful way to look capacity to service debt is to divide the gross debt of a firm by its operating cash flows, often seen as a key measure for predicting financial distress:

$$\text{total debt to gross operating cash flow} = \frac{\text{total debt (line 51)}}{\text{gross operating cash flow (line 3)}}.$$

Higher ratios indicate reduced capacity to service debt commitments.

Total debt to gross operating cash flow:

Year:	2000	2001	2002
HLY	8.87	7.71	6.89
TCL	–8.92	22.31	23.25
Industry average	8.36	0.58	5.70

Analysis. From this ratio it can be seen that HLY has a much better capacity to service its debt commitments than TCL from its operating cash flows. HLY is also improving its performance on this ratio (which is indicated by a declining trend) whereas TCL has a declining trend. It should also be noted that TCL is doing particularly poorly on this ratio, being nearly four times higher than the industry average. Generally, these results present a similar story to the interest and cash cover ratios.

6.7. Assessing turnover

Various aspects of the efficiency with which assets are utilized can be gleaned from turnover ratios as well as from several of the previously examined ratios.

One such ratio is the total asset turnover ratio defined as

$$\frac{\text{sales (line 26)}}{\text{total assets (line 47)}}.$$

Asset turnover (%):

Year:	2000	2001	2002
HLY	6.2904	6.7136	8.2224
TCL	1.0935	7.7610	2.8513
Industry average	35.1253	32.0515	36.7463

This is a standard ratio indicating the sales-generating capacity of a firm's assets. He says: "It is one measure of management's capacity in dealing with competitive conditions." The higher this ratio, the greater the sales generating capacity of the firm.

Analysis. It can be seen that both TCL and HLY are well below the industry average, though HLY has a significantly higher turnover ratio than TCL. TCL's poor showing on this ratio can be partly explained by a significant decline in its sales revenue in 2002.

A second turnover ratio is the accounts receivable turnover ratio:

$$\frac{\text{accounts receivable (line 39)} \times 365 \text{ days}}{\text{net interest payments (line 32)}}.$$

Accounts receivables turnover (days):

Year:	2000	2001	2002
HLY	10.3	17.7	16.6451
TCL	146	50.3	93.4
Industry average	123.8924	62.0377	134.1215

Account receivables are rights to receive cash from customers through the sale of goods and services. If customers buy on credit instead of paying cash up front, the firm records a revenue, but a corresponding asset (account receivables) must be created to reflect the right to receive cash. Strictly speaking, only credit sales should be included in the denominator of this ratio. However, in many cases, total sales are used because a detailed breakdown of cash and credit sales is often not provided in the published annual reports of companies. By mulitplying 365 by the accounts receivable turnover ratio, an estimate of the average collection period of credit sales can be made. In other words, how many days does it take, on average, to collect on receivables? The higher the ratio, the longer the collection period. Collection periods that are excessively long can adversely impact on a firm's cash flows, and may signal a requirement for management to tighten up collection procedures.

Analysis. It can be concluded that HLY collects on its receivables every 16 days, where TCL has a much longer collection period of 93.4 days. Differences in the collection periods between HLY and TCL may partly explain differences in operating cash flow results between the two companies.

6.8. Assessing variability

So far we have only discussed cross-sectional ratio techniques, or financial performance at a point in time, notwithstanding that we have made mention of trends in financial ratios over the last 3 years for both HLY and TCL. An approach that is gaining popularity in the literature is to compute variability measures for financial ratios and other variables over time (Dambolena and Khoury, 1980). Variability is also referred to as volatility. For instance, firms with highly variable earnings or cash flow streams from year to year are considered more volatile by analysts. Firms with higher volatility in their earnings or cash flow streams are generally harder to predict by analysts, which make them higher risks when making in investment decisions. One common measure of variability is calculated as follows (Foster, 1986):

$$\frac{\text{maximum value} - \text{minimum value}}{\text{mean}}.$$

Table 1 presents this measure of variability for the net operating cash flow to total assets of ratio for both HLY and TCL. It can be seen that HLY has the lowest variability in the operating cash flow ratio, even though it has been listed on the ASX for a longer period than TCL. Another method for calculating variability is the coefficient of variance, which is the standard deviation divided by the mean (expressed as a percentage). Standard deviation is one of the most common measures for assessing volatility and risk in the finance literature (Frino et al., 2001). Standard deviation is a measure of the average variance of observations from the mean. It measures how tightly clustered (or how dispersed) observations are around the mean. Higher standard deviation indicates higher volatility and vice versa. It is calculated as the square root of the variance, where the variance is calculated as

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N},$$

where μ is the mean and N is the number of observations. The standard deviations for the TCL and HLY operating cash flow to total assets are 4.39 and 2.89, respectively. Dividing by the mean gives a coefficient of variability of 4100% for TCL and 72.9% for HLY (Table 1). Once again, the volatility of firm's cash flow streams need to be closely examined relative to the volatility of the industry benchmark. In the case of infrastructure and utilities, the average coefficient of variation is close to 90%. Needless to say, TCL's volatility, at least in terms of cash flow, is much higher than HLY and the industry average. Notwithstanding this, TCL only has 3 years of financial data, and some analysts might reasonably regard this period of time as too short to conduct a reliable volatility analysis of cash flow.

6.9. Assessing valuation fundamentals

Fundamental information on a firm's financial performance, such as earnings, cash flows, and sales performance, is also useful for making assessments about the relative value of a firm's shares on issue. This kind of evaluation can be very useful in assessing whether a firm's shares are relatively over- or underpriced. Taking the case of real estate, rational investors do not wish to buy properties that are overpriced – overpriced properties generally provide lower return yields and can be subject to market corrections, and consequently lower capital gains. The analogy of real estate is similar to share purchases. The higher the share price, the

Table 1
Net operating cash flow to total assets, 1995–2002

Year	TCL	HLY
1995		–2.0776
1996	0.0000	1.9728
1997	0.0000	3.7637
1998	0.0000	5.4669
1999	0.0000	4.2611
2000	–4.8658	5.0977
2001	3.4490	6.1323
2002	1.7374	7.0842
$\frac{\text{max.} - \text{min.}}{\text{mean}}$	519.62	2.31
$\frac{\sigma}{\text{mean}}$	4100%	72.9%

lower the dividend yield (dividends per share divided by share price), and the greater the risk that the price will correct unless the firm can meet market expectations about future growth.

Price–earnings ratio

One measure of valuation frequently used by analysts is the price–earnings (PE) ratio. This ratio is calculated as follows:

The price-earnings (PE) ratio is a frequently used figure in this analysis:

$$\text{PE ratio} = \frac{\text{market price share (line 66)}}{\text{earnings per share (EPS)}},$$

where

$$\text{EPS} = \frac{\text{NPAT (line 36)} - \text{preference dividends (line 64)}}{\text{total weighted number of shares outstanding (line 62)}}.$$

All other things being equal, the higher the PE ratio, the higher the expectations of firm's future performance. This is because share prices are often act as a "barometer" of future performance of the firm – firms with good track records, strong growth prospects, and good management are usually more attractive to investors, which creates buying interest in the stock, thus lifting the price and therefore the PE ratio. Generally, lower PE ratios represent better buying opportunities for investors, assuming a firm has good quality earnings and solid growth prospects. Once again, it is crucial to look at industry averages when assessing an individual firm's PE ratio, as PE ratios can differ very substantially between industries.

PE ratio:

Date: July 2003	EPS	PE
HLY	–0.008	–
TCL	–0.13	–
Industry average	–	39.8

Analysis. Both HLY and TCL have negative earnings, which makes the PE ratio non-interpretable (i.e. a negative PE ratio is meaningless, and hence it is not computed). Given that HLY and TCL reported negative earnings, why do both firms have such high levels of market capitalization? One explanation is that the market has factored a "premium" into the share price that reflects positive market expectations about HLY's and, to a lesser extent, TCL's future growth potential (partly reflected in HLY's stronger cash flow performance). Furthermore, the market does not have high expectations for the profitability of toll road entities engaged in long-term capital investments and that typically have large depreciation and interest charges in the earlier phases of business activity. It is noteworthy that the industry average PE ratio is 39.8, which is quite high relative to other industries in Australia (the overall market average is between 15 and 20). The high industry PE ratio could signal that the market expects strong future growth from many of the firms in this industry group. It was mentioned previously that cash flow analysis is more relevant than profitability analysis for toll road entities (or other infrastructure companies) in the earlier stages. We now turn to cash flows as a measure of valuation.

A note on the price earnings to growth (PEG) ratio. Peter Lynch, one of the most successful fund managers of all time, argued that "The p/e ratio of any company that's fairly priced will equal its growth rate." The PEG ratio is calculated by dividing the PE ratio (using the expected EPS figure in the denominator) by the historical growth in the EPS. If the PEG ratio is equal to one, it signifies that investors are pricing the share fully (at the expected EPS growth rate). This makes sense, because in an efficient market the PE ratio should reflect a share's future earnings growth. If the PEG ratio is greater than one, this could suggest that the share is potentially overvalued, or that the market expects future EPS growth to be greater than what analysts currently forecast. Growth shares (such as new economy shares) typically have a PEG ratio greater than one because investors are willing to pay more for a share that is expected to grow quickly. By way of example, analysts forecast that HLY's EPS would be Aus. \$0.01 in 2003, the year following the period analyzed in this chapter. The PE ratio based on the 2003 share price of 5.7 gives a PE ratio of 570 times. Historical growth in the EPS over the preceding year, 2002, was approximately 50%. Thus, 570 divided by 50 yields a PEG ratio of 11.4, which indicates either a very high level of potential overvaluation, or that the market expects HLY's earnings to grow rapidly in the future.

Cash flow per share

Analysts now pay closer attention to cash flow per share. As mentioned previously, earnings (and therefore earnings per share) are more subjective than cash flows, and can be subjected to managerial manipulation. Cash flows can offer a more objective and reliable benchmark for assessing a firm's overall operating performance.

$$\text{cash flow per share} = \frac{\text{net operating cash flow (line 8)}}{\text{total weighted number of shares outstanding (line 62)}}.$$

Price to cash flows:

Date: July 2003	Cash flow per share (Aus. cents)	Price to cash flow ratio
HLY	0.215	26.5
TCL	0.105	11.42
Industry average	–	17.85

Analysis. HLY's price to cash flow ratio is quite high at 26.5, whereas TCL's ratio of 11.42 is actually below the industry average, and less than half of HLY's price to cash flow ratio. One interpretation of this result is provided above – that is, the market has built a substantial premium into HLY's share price because there is stronger expectation of future growth. However, on a cash flow basis, HLY is looking overvalued, which could result in a market correction unless HLY can generate strong future cash flow growth to justify is current share price.

Sales per share ratio

Another valuation measure that is becoming more commonly used by analysts is the sale per share ratio. This ratio express sales performance in per share terms. By dividing share price by the sales per share figure, we get the price per share ratio:

$$\text{sales per share} = \frac{\text{sales (line 26)}}{\text{total weighted number of shares outstanding (line 62)}}.$$

Price to sales:

Date: July 2003	Sales per share (Aus. cents)	Price to sales per share
HLY	0.38	15
TCL	0.22	21.4
Industry average	–	11.4

Analysis. On a per share basis, TCL is more overvalued, with a ratio of 21.4, nearly double the industry average. HLY has a more modest ratio of 15 times. Hence, on the basis of sales revenue performance, HLY is valued closer to the industry norm. Again, the ratio reflects the relatively stronger performance of HLY on sales growth figures in 2002.

Price to book value

Another valuation measure used by analysts is the price to book value ratio. Book value represents what the shareholders own in the firm, which is determined by subtracting total liabilities from total assets. It includes both tangible and intangible assets. It is measured by dividing the shareholders' equity by the number of shares outstanding, as of the year end balance date:

$$\text{book value per share} = \frac{\text{shareholders' equity (line 55)}}{\text{total weighted number of shares outstanding (line 62)}}.$$

Price to book value:

Date: July 2003	Book value per share (Aus. \$)	Price to book value ratio
HLY	2.03	2.81
TCL	4.05	1.16
Industry average	–	1.63

Analysis. HLY has a price to book value ratio of 2.81. Once again, TCL is relatively undervalued by the market relative to HLY and the industry average.

6.10. Assessing investment returns

Analysts are keenly interested in the investment returns of companies. Investment returns can take two forms: dividends received on shares and capital growth in share value. Dividends can be made in the form of cash, or additional shares.

While past investment returns are not always a good indication of future returns, analysts nevertheless pay close attention to the dividends paid by firms, as well as current stock price performance. For instance, a downward momentum on share price could signal a revision in the markets expectations of a firm's future performance that is not immediately captured in the current financial performance results. Furthermore, stock prices reflect not just the financial performance of firms but also economy- and industry-wide information that may have an impact on perceptions of a firm's future financial performance.

One of the most useful measures of investment return is the dividend yield:

$$\text{dividend yield} = \frac{\text{dividend per share (line 65)}}{\text{share price (line 66)}}.$$

Dividend yield (%):

Date: July 2003	Dividend yield
HLY	3.42
TCL	–
Industry average	4.3

Analysis. It can be seen HLY has a dividend yield of 3.42%, which is a little below the industry average. However, up until June 2002, TCL was not paying a dividend at all to its shareholders. Hence, on a dividend per share and yield basis, analysts would regard HLY as the more attractive investment. Another key measure of investment performance is shareholder return, which captures two elements of investment return: share price growth and dividends. Over the period 2000–2002, TCL's shareholder return has been 14.2%, while HLY's has been significantly higher at 19.7%. In short, HLY has proven the better investment over these 3 years.

7. Conclusions

This study has illustrated some basic analytical techniques used by financial analysts in assessing the performance of firms. These techniques are equally applicable to toll road entities in other jurisdictions, including North America and the UK. Furthermore, given dramatic financial management reforms in the Australian public sector over the last decade, including the widespread introduction of "private sector" accrual accounting throughout all levels of the public sector (Jones and Puglisi, 1997), most of the techniques discussed in this chapter are expected to have broad application to public sector agencies.

It can be seen from the analysis that HLY appears to be performing significantly better than TCL across a number of measures, particularly cash position, liquidity, profitability, cash flows, turnover, and dividend paying performance, over the period considered (2000–2002). Given some inherent difficulties in applying profitability analysis to infrastructure companies, which tend to be long-term and highly capital-intensive, cash flow analysis provide better overall performance measures in the short term. HLY's superior financial performance appears to be reflected in its shareholder returns, which capture both stock price performance and dividend paying capacity. However, the analysis reveals that HLY's valuation fundamentals indicate some level of potential overvaluation, particularly on price-to-cash flow and PEG ratios. At least, the firm's share price appears "fully valued" in terms of current expectations of future performance.

While financial ratio analysis is the prevalent technique used by analysts to interpret company financial performance, readers are cautioned about the limitations of ratio analysis. One problem is that financial ratio analysis draws on conventional financial statements, which suffer certain limitations that can undermine their relevance, interpretability, and comparability. Some of these issues relate to the broader question of the measurement basis of accounting. For instance, in the annual report of TCL in 2002, under Note 1, "Summary of significant accounting policies," the company states under "(b) Historical Cost Convention" that: "The financial statements are prepared on the basis of the historical cost convention and, except where stated, do not take into account current valuations of non-current assets." It is noted that the financial statements of HLY are also prepared under the same convention. Under the historical cost convention, a non-current asset is carried in the accounts at its original cost (less accumulated depreciation), notwithstanding that its current market value (i.e. what it can be bought or sold for) may bear no resemblance to its historical cost value.

Changes to its market value can be created by a number of factors, such as inflation, wear and tear, and obsolescence (Sterling, 1970). Furthermore, the written down values of assets (i.e. cost less accumulated depreciation) are likely to differ significantly between entities because they are brought into the accounts over different stages of the firm's life, which can also undermine inter-firm comparability. Hence, ratios based on total asset values and book values (such as rate of return measures) do not necessarily provide valid or comparable depictions of a firm's true financial performance. Another limitation is that accountants have developed different valuation rules for different kinds of assets and liabilities. Sterling, for instance, identifies at least 11 different methods used by accountants to value different items in the statement of financial position. This has resulted in the "additivity" problem, where the statement of financial position effectively sums "apples and oranges," rendering interpretations of a firm's financial position very troublesome if not impossible (Walker and Jones, 2003).

There is one further word of caution. When comparing the performance of two companies (such as HLY and TCL), it is important to examine the accounting policies that firms use as the basis for preparing financial information. Accounting policy disclosures appear in the notes to the accounts, which provide more detailed information on how financial statement items are calculated. Accounting policy disclosures cover such issues as the basis of accounting, valuation of property plant and equipment, principles of consolidation, treatment of borrowing costs, methods of depreciation, and so on. For instance, prior to July 2000, the notes to HLY's financial statements indicate that the firm measured property plant and equipment at revalued amounts. This affects how non-current assets, and therefore total assets, are reported, as well as net profit reported in the statement of financial performance (through depreciation expense). Where there are significant accounting policy differences between firms, accounting adjustments need to be made to put the financial statements of firms on a comparable footing. In the case of HLY and TCL, the accounting policy disclosures appear to be similar in most material respects, hence significant accounting adjustments were not undertaken.

Appendix. Financial statement summaries for TCL and HLY (2000–2002)

All the information in the table overleaf is correct for 1 July 2003. Figures are in millions and Australian dollars, unless indicated otherwise. The source of the financial statement data is *Huntley's FinAnalysis Database* (2003).

	Company	TCL	HLY	TCL	HLY	TCL	HLY
	Year	**2000A**	**2000A**	**2001A**	**2001A**	**2002A**	**2002A**
	Balance date	**30/6/00**	**30/6/00**	**30/6/01**	**30/6/01**	**30/6/02**	**30/6/02**
	Months	12	12	12	12	12	12
	STATEMENT OF CASH FLOWS:	**30/6/00**	**30/6/00**	**30/6/01**	**30/6/01**	**30/6/02**	**30/6/02**
	OPERATING CASHFLOWS:						
1	Receipts from customers	26.945	61.411	158.112	69.967	109.451	78.371
2	Payments to suppliers	–169.296	–13.847	–90.415	–15.269	–39.429	–16.874
3	**Funds from operations**	**–142.351**	**47.564**	**67.697**	**54.698**	**70.022**	**61.497**
4	Dividend income	0.000	0.000	0.000	0.000	0.000	0.000
5	Interest expense (net)	35.305	–12.527	–70.788	–16.789	–43.940	–14.381
6	Tax expense	0.000	0.000	0.000	–7.955	0.000	–1.865
7	Other	1.212	0.000	3.111	0.000	27.711	–5.349
8	**Net operating cash flows**	**–105.834**	**35.037**	**0.020**	**29.954**	**53.793**	**39.902**
	INVESTING CASHFLOWS:						
9	Cash paid for PP&E	–1310.770	–0.817	–17.514	–0.920	–8.663	–3.080
10	Asset sales	0.000	0.000	0.000	0.000	0.000	0.000
11	Acquisitions & investments	0.000	0.000	0.000	0.000	0.000	0.000
12	Net loans to entities	0.000	0.000	0.000	0.000	0.000	0.000
13	Other	0.000	0.000	0.000	0.000	1361.215	0.000
14	**Net investing cash flows**	**–1310.770**	**–0.817**	**–17.514**	**–0.920**	**1352.552**	**–3.080**
	FINANCING CASH FLOWS:						
15	Equity raised	5.091	0.000	0.000	0.000	0.000	0.000
16	Debt	2389.737	436.000	0.000	0.000	8.000	2.000
17	Bonds, notes, commercial paper	–454.090	–319.418	0.000	0.000	0.000	0.000
18	Debt repayment	–180.195	–48.100	–7.570	–36.075	–13.330	–31.450
19	Dividends	0.000	0.000	0.000	0.000	–19.952	0.000
20	Other	–418.708	–115.464	1353.668	0.000	0.000	0.000
21	**Net financing cash flows**	**1341.835**	**–46.982**	**1346.098**	**–36.075**	**–25.282**	**–29.450**
22	**Net increase in cash**	**–74.769**	**–12.762**	**1328.604**	**–7.041**	**1381.063**	**7.372**
23	Cash at beginning	99.801	52.220	25.032	39.458	1353.636	32.417
24	Exchange rate adjustments	0.000	0.000	0.000	0.000	0.000	0.000
25	**Cash at end**	**25.032**	**39.458**	**1353.636**	**32.417**	**2734.699**	**39.789**

		30/6/00	30/6/00	30/6/01	30/6/01	30/6/02	30/6/02
	STATEMENT OF FINANCIAL PERFORMANCE SUMMARY:						
26	Trading revenue	31.990	58.692	152.334	59.883	114.916	71.377
27	Expenses	108.417	20.920	105.604	21.651	96.589	21.972
28	EBITDA	–76.419	40.149	46.730	41.906	106.610	49.405
29	Depreciation	42.568	5.211	60.697	33.227	79.019	34.047
30	Amortization	0.065	30.355	0.000	2.021	0.248	0.000
31	EBIT	–119.052	4.583	–13.967	6.658	27.343	15.358
32	Net interest expense	44.306	17.656	100.387	17.441	94.507	15.980
33	Gross interest expense	115.384	54.127	254.438	54.941	170.553	54.504
34	Pre-tax profit	–163.358	–13.073	–114.354	–10.783	–67.164	–0.622
35	Tax	0.000	2.331	0.000	0.640	0.000	0.914
36	Net profit after tax	–163.358	–15.404	–114.354	–11.423	–67.164	–1.536
	STATEMENT OF FINANCIAL POSITION SUMMARY:						
37	Cash (incl. short term depreciation)	25.032	39.458	104.636	32.417	132.063	39.789
38	Current investments	0.000	0.000	0.000	0.000	0.000	0.000
39	Account receivables	12.800	1.664	20.996	2.905	29.416	3.250
40	Inventory	0.000	0.000	0.000	0.000	0.000	0.000
41	Total current assets	39.031	49.707	126.396	44.300	163.762	52.228
42	Property, plant and equipment	2886.495	867.018	1836.411	833.362	3856.090	803.575
43	Intangibles (excl. goodwill)	0.000	0.000	0.000	0.000	0.000	0.000
44	Goodwill	0.000	0.000	0.000	0.000	9.752	0.000
45	Non-current investments	0.000	0.000	0.000	0.000	0.000	0.000
46	Other assets	0.000	16.320	0.000	14.299	0.660	12.279
47	**TOTAL ASSETS**	**2925.526**	**933.045**	**1962.807**	**891.961**	**4030.264**	**868.082**
48	Current liabilities	122.136	25.438	213.152	20.355	194.266	24.378
49	Current debt	0.137	0.000	20.672	0.000	8.000	0.000
50	Non-current debt	1269.653	421.820	1489.683	421.820	1620.169	423.820
51	Total debt	1269.790	421.820	1510.355	421.820	1628.169	423.820
52	Other liabilities	1691.995	29.896	140.725	36.768	147.368	44.477
53	**TOTAL LIABILITIES**	**3083.784**	**477.154**	**1843.560**	**478.943**	**1961.803**	**492.675**
54	**NET ASSETS**	**–158.258**	**455.891**	**119.247**	**413.018**	**2068.461**	**375.407**

	Company	TCL	HLY	TCL	HLY	TCL	HLY
55	**SHAREHOLDERS' EQUITY**	**–158.258**	**455.891**	**119.247**	**413.018**	**2068.461**	**375.407**
56	**NET TANGIBLE ASSETS**	**–158.258**	**455.891**	**119.247**	**413.018**	**2058.709**	**375.407**
	GROSS INVESTMENT						
57	**Increases in working capital**	27.779	0.000	–31.798	–9.823	13.671	3.145
58	**Capital expenditure**	1310.770	0.817	17.514	0.920	8.663	3.080
59	**Gross investment**	**1338.549**	**0.817**	**–14.284**	**–8.903**	**22.334**	**6.225**
	ISSUED SHARE CAPITAL						
60	**Opening**	0.910	185.000	510.000	185.000	510.000	0.000
61	**Closing**	510.000	185.000	510.000	185.000	510.028	185.000
62	**Total weighted average**	486.192	185.000	510.000	185.000	510.028	185.000
	MARKET DATA						
63	**Market capitalization**					**2473**	**1073**
	DIVIDEND INFORMATION						
64	**Preference dividends**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
65	**Ordinary dividends**					**NA**	**19.5 cents**
66	**Current share price**					**$4.71**	**$5.70**

References

DeAngelo, L.E. (1986) "Accounting numbers as market valuation substitutes: a study of management buyouts of public stockholders," *Accounting Review*, 61:400–420.

Dambolena, I.G. and S.J. Khoury (1980) "Ratio stability and corporate failure," *Journal of Finance*, 35:1017–1026.

Foster, G. (1986) *Financial statement analysis*. New York: Prentice-Hall.

Frino, A., A. Cusack and K. Wilson (2001) *Introduction to corporate finance*. Sydney: Pearson.

Healy, P.M. (1985) "The effect of bonus schemes on accounting decisions," *Journal of Accounting and Economics*, 7:85–107.

Heath, L.C. (1978) "Financial reporting and the evaluation of solvency," in: *Accounting Research Monograph 3*. New York: American Institute of Certified Public Accountants.

Jones, S. and D.A. Hensher (2003) *Predicting financial distress using reported cash flows: an ordered mixed (random parameters) logit design,* Working paper. Sydney: University of Sydney.

Jones, S. and N. Puglisi (1997) "The relevance of accrual accounting in the Australian public sector: a cause for doubt," *Abacus*, 33:115–132.

Jones, S., C. Romano and K. Smyrnios (1995) "An evaluation of the usefulness of cash flow statements by Australian reporting entities," *Accounting and Business Research*, 25:115–129.

Lee, T.A. (1986) *Towards a theory and practice of cash flow accounting*. New York: Garland.

Livnat, J. and P. Zarowin (1990) "The incremental information content of cash-flow components," *Journal of Accounting and Economics*, 12:25–46.

Neill, D., T. Schaefer, P. Bahnson and M. Bradbury (1991) "The usefulness of cash flow data: a review and synthesis," *Journal of Accounting Literature*, 10:117–150.

Rappaport, A. (1982) "Let's give the shareholders the figures they need," in: S. Zeff and T. Keller, eds, *Financial accounting theory: issues and controversies*. Singapore: McGraw-Hill.

Sterling, R.R. (1970) *Theory of the measurement of enterprise income*. Lawrence: University Press of Kansas.

Walker, R. and S. Jones (2003) "Measurement: a way forward," *Abacus*, 39:356–374.

White, G., A.C. Sondhi and D. Fired (1997) *The analysis and use of financial statements*, 2nd edn. New York: Wiley.

This Page Intentionally Left Blank

Chapter 20

FINANCING TRANSPORT INFRASTRUCTURE: PUBLIC FINANCE ISSUES

PETER ABELSON
Macquarie University, Sydney

1. Introduction

This chapter discusses methods of financing transport infrastructure and the implications for public finance. "Financing" in this chapter refers to the raising of financial capital to fund infrastructure. However, financing infrastructure is not a self-contained issue. As will be seen, the optimal way to raise capital cannot be separated from the question of how capital is serviced and repaid. Also, financing has implications for the ownership, organization, and management of infrastructure.

The distinction between raising capital and servicing and repaying capital is important. The term "financing" is often used to refer not only to raising capital for infrastructure but also to payments for the services provided by the infrastructure. The latter use of the term can produce confusion. Although the concepts of raising and paying for capital are often linked, they are not synonymous. In general, different parties are responsible for financing transport infrastructure and for paying for it.

Likewise, the raising of finance is linked to the ownership and performance of the infrastructure. The provision of capital often confers, or should confer, ownership or property rights, which in turn have implications for the management of assets. The optimal method of raising finance is the method that delivers overall the best value for money inclusive of the performance of the infrastructure. It is not necessarily the method that provides finance at least cost.

This chapter takes as its starting point that the proposed transport infrastructure represents an efficient use of resources, generally as determined by cost–benefit analysis. We assume that the expected social rate of return generated by the proposed capital formation exceeds the rate of return available from other uses of scarce capital. This is a necessary condition for the efficient allocation of publicly or privately financed resources. It is of course possible that the social viability of a project depends on how the project is delivered. However, discussion of cost–benefit analysis and related issues of resource allocation are outside the scope of this chapter.

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

This chapter is divided into four main parts. Section 2 describes instruments for raising and servicing financial capital. This provides a structure for the rest of the analysis. Section 3 discusses some general issues concerning the relationships between raising capital, the nature of infrastructure, user charges, risk, and ownership. Section 4 discusses the advantages and disadvantages of the main methods for raising finance for new infrastructure, including public–private partnerships and full private ownership and finance. A key issue is whether the financial structure for each major piece of infrastructure should be determined by the merits of each case or whether there are broader macroeconomic considerations, for example with respect to national net debt, which should determine how capital is raised. This chapter argues that the financial structure should generally be determined by the needs of the case rather than by macroeconomic considerations. Section 5 discuses the re-financing of infrastructure that has been financed by the public sector, which is generally referred to as privatization of publicly owned assets. Some issues are similar to those that arise in the raising of capital for new infrastructure. However, it may be desirable to develop an asset with public funds and later to sell the asset to private ownership. In any case, there is the question of how to deal with existing transport assets.

2. Instruments for raising and servicing capital

Table 1 shows the main instruments available for raising and servicing capital. There are three main ways to raise capital: by taxation, public sector borrowing, or private sector financing. For each main way, there are important subsets. Importantly, for each method of raising capital, there is often a prime instrument for servicing it.

Tax revenue may be derived from consolidated (general) tax revenue or from specific infrastructure-related taxes. Sometimes, part of general revenue, notably revenue from fuel taxes, is hypothecated to fund transport infrastructure. Capital may also be raised by development taxes levied on either developers or households, or both. For example, a value-capture tax may be levied on the estimated increase in land value due to the development of related infrastructure. Another infrastructure-related tax is a levy on developers; for example, in Sydney, Australia, land developers are required to pay the state government Aus. $15 000 per lot developed as a contribution to transport infrastructure. Of course, when capital formation is financed from any form of taxation, there is no capital servicing requirement.

The most common form of public sector borrowing is via long-term bonds. Bonds are interest-bearing certificates of debt. They entail the payment of interest to the lender as well as repayment of the principal at a nominated future date. Government bonds often have a maturity of 10 years or more. In some cases, interest payments are indexed against inflation. These "inflation-indexed bonds" eliminate significant uncertainty for potential investors about the impact of inflation over the long life of a project. Government bonds are usually serviced out of general taxation spread over the life of the bond, although user charges may provide supplementary revenue.

The government may also raise capital via specific infrastructure funds or infrastructure revenue bonds. These instruments are similar in that in both cases the government raises money from the public via bonds for the purpose of constructing infrastructure. However, an infrastructure fund may allocate its capital to a variety of infrastructure projects. It may not be project- or revenue-specific. Also, the capital may be serviced from general tax revenue or from a hypothecated tax source, such as a fuel levy. An infrastructure revenue bond is more likely to be project specific. The revenue raised by the bond issue provides capital to a specific project, and this capital is serviced from project revenues, albeit with a government guarantee backed by the government's tax powers. However, there are no precise general definitions of these concepts.

Another feature of infrastructure funds or revenue bonds is that they may be tax advantaged. In such cases, a bondholder typically receives a tax concession on the interest from the bond. The Australian government offered tax concessions on infrastructure bonds for a period in the 1990s. In the USA, municipal bonds (from which funds can be applied to various uses) are tax advantaged.

The public sector may also raise capital via borrowing by public trading enterprises (PTEs), such as airport or seaport corporations. The PTE usually services and repays this capital from user charges for the services that it provides. However, when a PTE incurs a deficit, the government may be required to fund the deficit from tax revenue.

On the other hand, the private sector may finance transport infrastructure quite separately from the government by issuing private debt or equity, or by some complex composition of financial instruments. Private companies usually expect to service and repay most of this debt or equity from user payments for the services provided by the infrastructure.

However, the government may contribute to the servicing and repayment of capital. It may do so by paying for particular services or by paying shadow tolls, which are payments for services when no user charge is levied. The government may pay transparent subsidies for services provided, which are sometimes described as community service payments. In other cases it may make unconditional guaranteed repayments. For example, the New South Wales government guaranteed repayments to the companies that constructed the Sydney Harbour Tunnel.

When payments are guaranteed and unconditional, it is questionable whether there is any real difference between public and private borrowing. Ownership implies the right to the residual income stream of an asset. Other parties may have claims on part of the income stream, but the owner has the claim to the surplus (or

Table 1
Instruments for raising and serving capital

Capital raising	Capital servicing
Taxation	
Consolidated revenue (taxation)	No servicing required
Infrastructure levies	No servicing required
Public sector borrowing	
Borrowing – general bonds	Taxation/user charges (a) Sale of asset
Infrastructure funds	Hypothecated tax revenue
Infrastructure revenue bonds	Project revenue
Public trading enterprise borrowing	User charges (a)/taxation supplements
Private sector financing	
Private debt	User charges (a)
Private equity	Government contributions (taxation)
Mixed private instruments	Shadow tolls
	Subsidies – community service payments
	Guaranteed repayments

Note: (a) May include access charges.

bears the loss). When government guarantees surplus income and takes the risk for shortfalls in income, nominal ownership may be private but the real effects are borne by the public sector. Income shortfalls affect public net worth and borrowing requirements. Although the transport infrastructure may be financed by the private sector, the government bears the ongoing risk. The effect on the government's fiscal position is the same as if the government itself had borrowed the money (at private sector rates) and incurred the liability for servicing and repaying the capital. The instrument used is a *de facto* government loan.

Of course, there are many permutations of the instruments shown in Table 1 and described above, including public–private partnership options. For example, infrastructure revenue bonds may be securitized and sold to the private sector. Alternatively, the government may permit a private provider to issue revenue bonds that are entitled to the revenue stream that the asset provides.

Another mixed form of financing is user contributions and contracts. For example, in the USA, airlines sometimes contribute some of the finance required for an airport development, such as a runway upgrade, and the airport agrees a long-term contract with them for charges that reflect the contributions made. The users have a direct input into the choice of standard and the capacity of the facilities, and the contract protects the interests of the users.

In any event, the policy objective is to find the preferred capital raising/capital servicing option, including public–private partnership combinations.

3. Capital raising, user charges, risk, and ownership

Before considering the merits and demerits of each capital-raising instrument, some general issues can usefully be examined.

Traditionally, public sector involvement in the financing of infrastructure was justified by the characteristics of the infrastructure (Access Economics, 2003). Infrastructure assets are generally long-lived, capital-intensive, and part of a wider network. The pay-back periods are long. Also, infrastructure assets often provide essential services, whose disruption can impose significant costs on the community. However, in recent years, increasing sophistication in public–private arrangements, especially in relation to contract design and pricing, has facilitated more financing options.

3.1. Capital raising and user charges

A key issue is the relationship between the capital-raising instrument and user charges. As we have seen and would expect, where capital is due to be serviced by taxation in the short or long run, capital is usually raised by the public sector. Other things being equal, this is appropriate. The party that has the financial obligations (liabilities) to service and repay the capital also has the power to ensure that these obligations are met. Private firms do not have the power of taxation, and so would be reliant in any case on public subsidy and support.

On the other hand, where a high proportion of the capital is serviced from user charges, there are stronger arguments for private sector financing. The private parties that have raised the finance and incurred the liabilities have both the means and the incentive to ensure that all or most of the capital is repaid. As we have seen, the government may contribute to capital servicing and repayment costs. However, where the government is responsible for servicing a high proportion of these costs, it bears the real liability for the capital, and may be regarded *de facto* as the real financier of the infrastructure.

Given that the methods available to service financial capital, especially user charges, are a significant determinant of the method of raising capital, it is important to consider when user charges are feasible and appropriate. When user charges are feasible and there are few externalities, there is a strong case for them. User charges, especially charges based on marginal costs, encourage an efficient allocation of resources because they provide signals to producers and consumers about the cost and value of services provided. They also provide incentives to management efficiency. When charges based on marginal costs fail to cover full costs, mark-ups or two-part tariffs can be applied to meet cost recovery financial objectives. When charges are inconsistent with equity objectives, the charges can be combined with transparent public subsidies. In such situations there is a prima facie case for private finance combined with user charges and limited and precisely targeted public subsidies.

User charges also have the advantage that they minimizes the tax revenue required and the related problem of the excess burden of taxation. Excess burdens are the costs associated with taxation that distorts voluntary behavior, for example income taxation that causes people to reduce work hours. The total burden of taxation is the tax paid plus the excess burden. Australian studies suggest that the excess burden of taxation is in the order of Aus. $0.20–0.40 in marginal dollar of tax raised (Abelson, 2003).

On the other hand, there are many situations in which user charges are either not feasible or not appropriate. For example, user charges may not be technically feasible at a reasonable cost in congested urban areas with limited technology. User charges may not be appropriate when there is excess capacity and the marginal cost of use, of, say, a bridge in a rural area, is very low. Full marginal cost pricing may also be inappropriate because of positive externality or safety equity reasons, for example when public transport is safer than private vehicle use. In such circumstances, public financing of transport infrastructure is usually preferred.

3.2. Capital raising and risk

Another general issue relates to risk. Here again there are two sub-issues. The first is whether the government or the private sector can generally handle risk better than the other party. This is a contentious area. Some economists, for example Arrow and Lind (1970) and, more recently, Quiggin (1997), argue that the public sector can handle risk better than the private sector because it can pool risk over a large number of projects and spread risk over a large number of taxpayers so that there is little risk in the long run to any individual taxpayer. Thus, the real cost of risk is lower for government borrowing than for private borrowing. This is reflected in the differential between the government bond rate and the market rate for private borrowing, where the latter is often several percentage points higher.

The contrary argument, put for example by Domberger (1995), is that the private capital market can and does diversify risk efficiently. The comparison between the bond rate and the market rate of interest fails to allow for the risks that are implicitly borne by taxpayers. The bond rate is lower only because the government can resort to tax to repay loans. The risks of projects are borne by taxpayers who receive no compensation for the risks.

Both arguments have merit. It appears that the government can spread project risk better than can the private sector, and that part of the differential between the bond rate and the market rate of interest for a risky project reflects an advantage of government funding. However, following Domberger, the full interest rate differential may overstate the benefit of public funding. It is hard to quantify the real economic benefit of the lower nominal borrowing rate of the government.

The second sub-issue relating to risk is control over project specific risks. Some project-specific risks are specific market risks. They relate to traffic forecasts, marketing, construction costs, operating efficiencies, and so on. It appears that private firms generally deal with such risks better than the government does, especially when there is competitive tendering. The fundamental reason is that members of private firms have an incentive to confront these issues and deal with them by allocating risk and responsibility in effective principal-agent relationships. Public officials bear less responsibility for market risks. Thus, private firms are more experienced in handling such risks. For large and complex construction works, competitive tendering, which ensures that the builder bears the project-specific risk associated with construction, has usually proved superior to construction by government departments.

On the other hand, when enterprises are heavily regulated, because of monopoly power or because they generate significant externalities, public ownership may be preferable. In such circumstances an unregulated private owner cannot be expected to act in the public interest. When businesses are regulated, profits depend on the nature of government decisions. As Quiggin (1997) observes, the profits of, say, an airport depend heavily on decisions with respect to aircraft noise, international aviation agreements, transport links, and so on. Similar risks arise in government-managed transport networks. For example, the traffic flow on a toll road depends on government decisions on land uses, public transport, and other road projects. A private operator must demand a risk premium over and above the normal equity premium or obtain some guarantees of favorable treatment. In such cases, other things being equal, the optimal solution is for the risk associated with network management to be internalized through public ownership. Public ownership, in turn, implies some form of government finance for transport infrastructure.

4. Raising capital for new transport infrastructure

We now consider more specifically the advantages and disadvantages of the various ways of raising capital shown in Table 1.

4.1. Using current tax revenues

The first option shown is raising (and paying for) capital for transport infrastructure out of current consolidated tax revenues. Of course, most government borrowing is also paid for ultimately from tax revenues. However, the incidence of the payments is different. When capital for transport infrastructure is provided from current tax revenues, the tax payer pays for the full cost of the infrastructure while receiving only a small part, if any, of the benefits of the services that will be delivered, often many years later. Thus, funding infrastructure from current general tax revenues does not satisfy the beneficiary principle of equity – that the beneficiary of a service should pay for the benefit. Nor are there significant efficiency arguments for funding infrastructure from current consolidated revenue.

Similar objections may be raised to transport-related tax revenue, such as fuel tax revenue. As a revenue source, excise taxes on fuel are as much part of consolidated revenue as are taxes on tobacco, alcohol, and other consumption goods or, indeed, as income tax. Certainly, fuel taxes may be hypothecated to transport infrastructure, as they are to road construction in New South Wales (Australia). This hypothecation provides a stronger nexus between the tax base for the infrastructure and the beneficiary. Nevertheless, current taxpayers bear the burden for provision of future services, and again there are no clear efficiency benefits from financing transport infrastructure from excise taxes on fuel. (There may be other efficiency benefits from fuel taxes operating as a correction for pricing failures.)

There are stronger equity arguments for raising capital via special development taxes (or levies) that are specifically related to the infrastructure development. As we have seen, these taxes may be levied on landowners or on developers. As shown by Abelson (1999), taxes on developers are usually passed back to landowners, because development is a competitive activity. The assumption here is that the infrastructure will be provided in any case, so that property prices reflect the present or future availability of infrastructure services. If payment of a development levy enables the infrastructure to be provided when it otherwise would not be, the levy may be passed forward to house owners. However, this reflects essentially the provision of services that would not otherwise be provided. In principle, value capture levies on landowners, developers, or house owners for the provision of infrastructure that results in increased property services and prices are equitable. Even though many services are provided only years into the future, in most markets the expected benefits are capitalized into current property prices. Thus, if applied accurately, development taxes may be construed as user charges.

However, caution is needed. It is easier to identify the beneficiary of a bus or car trip when it occurs than to identify the proportion of the benefits occurring over the next 30 years or so that will accrue to a specific property or development before the infrastructure is even constructed. Nor is it easy to define the range of beneficiaries. Value capture levies are inevitably based on judgements that may turn out to be inaccurate.

Moreover, there are significant inefficiencies associated with development levies as a method of raising capital to finance infrastructure. The authority that receives the levies typically has little incentive to provide the infrastructure in a cost-effective or expeditious manner. The incentive to cost-effective provision is reduced when an authority has coercive tax powers with which to raise finance. Importantly, the authority has no incentive to spend the funds quickly when the authority can accrue interest from the funds and when there is no obligation to service the capital. At the time of writing, local government authorities in Sydney, Australia, hold Aus. $750 million in developer levies that are pledged to the provision of facilities but which have not been provided. In this situation, there is no discipline from external financiers on the delivery and performance of infrastructure.

4.2. Public borrowing

As we have seen, general government borrowing has the advantage that the government can borrow at a relatively low nominal and real cost. Private finance is generally several interest points more expensive. On loans of Aus. $100 million, public borrowing may save Aus. $4 million or Aus. $5 million. Also, by spreading out repayments over many years compared with funding from current taxation, borrowing spreads out the burden of repayment in line with the receipt of services.

However, in recent years many governments have been very cautious about borrowing to fund infrastructure expenditure. The UK introduced the Private Finance Initiative in the early 1990s primarily to circumvent strict fiscal constraints. The EU has an arbitrary and not strictly enforced rule that fiscal deficits should not exceed 3% of GDP. A fiscal deficit is the excess of all expenditure, including current and capital expenditure, over current revenue. In Australia, the Commonwealth government and several state governments have made debt reduction a major policy objective. Indeed, in 1995 the New South Wales parliament passed legislation requiring the government to aim at debt elimination.

The EU rule dealing with deficits may be justified as a short-term policy against inflation when economies are buoyant. When demand for goods is high, there may be a macroeconomic case for limiting deficits or indeed for running a fiscal surplus. However, the EU rule is an expedient one that has no long-run justification. There is no general economic reason why capital expenditure should be funded out of current revenues. The golden rule of public finance is that governments should achieve a net operating balance over the business cycle. The net operating balance is current revenues less current expenses including depreciation. The golden rule requires current revenue to cover the use of capital but permits borrowing to finance productive investment. Indeed, over the business cycle of typically about 5 years, public borrowing will equal public investment.

The golden rule implies maintenance of net public worth. There is no microeconomic problem with financing productive capital investment by borrowing. Public borrowing increases net debt but, if the investment is productive, borrowing has no effect on net public worth. Public debt *per se* is not relevant to social welfare. The variables that matter to social welfare are public sector net worth and public sector contingent risk. Moving debt off the current budget by borrowing off-budget does not reduce the real liabilities of the public sector if the public sector has the obligation to service the capital in any case – it merely disguises real liabilities of the public sector.

The two main arguments put forward for reducing debt are a high debt may raise public borrowing rates and the debt may not be sustainable. These arguments may apply when there is a high level of indebtedness. A sustainable level of liabilities is one that can be readily serviced when there is a sharp downturn in economic activity without seriously compromising the provision of regular public services. However, they are exceptional arguments for avoiding excessive levels of debt. They are not general arguments for debt reduction in normal circumstances.

The issuing of special-purpose public infrastructure bonds may be attractive to private lenders. They may provide a catalyst for mobilizing private sector money and for some public–private partnerships. Also, the greater accountability and the hypothecated nature of such funds may make them more attractive to lenders. However, the issuing of such bonds does not relieve the government of any obligations that would attach to other borrowing arrangements. Investors are more interested in the nature of the risk than in the purpose for which the bond is used. Such bonds usually do not provide any cost savings or efficiency benefits to the government, and may duplicate funding processes for conventional public projects.

As we have seen, infrastructure bonds may also be tax advantaged. When a public agency issues tax-advantaged bonds, it should achieve cost savings. It can raise funds with a lower interest rate because lenders equalize their expected after-tax rates of return. However, overall the public sector does not gain because it loses tax revenues.

The government may also allow private firms to issue tax-advantaged infrastructure bonds. The aim is to compensate investors in infrastructure projects for the long lead time before receipt of income, especially when the tax system is biased against infrastructure projects. In Australia the income tax system does not allow claiming of losses incurred in generating accessible income, including depreciation and interest, until there is other taxable income against which such losses may be offset. This means that losses may be deferred for many years with declining present value levels. On the other hand, such enterprises usually benefit because tax is not levied on accrued capital gains.

However, tax subsidies such as tax-advantaged infrastructure bonds have significant disadvantages. They are a crude way to deal with biases in the tax system or any project related externalities. They do not reduce real economic costs and they may increase investment distortions. Moreover, such tax concessions often provide large opportunities for tax minimization for individuals on high marginal tax rates.

The idea behind infrastructure revenue bonds is that they authorize and hypothecate an income stream to service and repay the debt. However, when the government issues such bonds, it is responsible for repaying the debt. Revenue bonds may be valuable where private investors are unwilling to accept project risks. As a rated instrument, revenue bonds can be attractive to institutional investors unable to accurately assess or manage project risk. Also, it may be possible to sell the debt to the private sector (securitize the debt).

There is little difference between borrowing by general government and borrowing by state-owned trading enterprises (SOEs). Although SOE borrowing may appear to quarantine the effect of the borrowing from the general government, when an enterprise is wholly publicly owned the change in net worth of the SOE is part of the balance sheet and net worth of the government. In most cases, the general government is responsible for meeting any financial deficiencies of a SOE. The main advantage of SOE borrowing is that the risks are sheeted home more directly to management and this tends to engender greater management discipline.

4.3. Private sector financing

The arguments for and against private financing of major transport infrastructure are essentially the mirror image of previous comments. Financing should be a function of ownership and the allocation of ownership.

The core issue is whether the benefits of the incentive structure provided by private construction and management offset the higher real cost of finance. The New South Wales Treasury (2002) notes that integrating design, construction, operation, and maintenance over the life of an asset within a single project finance package can encourage maximum innovation from the private sector to improve the design and performance of the infrastructure and to reduce the whole of life costs. This presumes that clear and explicit contracts can be written for all aspects of the work. Moreover, external financiers can provide an essential discipline for the efficient delivery and operation of infrastructure. However, these general benefits of private ownership need to apply in particular cases and to offset the higher costs of privately raised finance.

The allocation of risk is a related issue. Transferring risks (and financial responsibility) to the private sector is desirable when the private sector is better placed than the government to manage those risks and can thereby improve the cost and quality of infrastructure. As we have seen, the private sector is better than the public sector at managing many kinds of project risks. However, if the private sector requires the public sector to underwrite the operation of the infrastructure and effectively to insure it against risks of various kinds, private finance may be an expensive option.

5. Privatization: re-financing transport infrastructure

The transfer of assets from the public to the private sector, which is generally described as privatization, has been a common phenomenon over the past 20 years, and the subject of a great deal of analysis. This chapter does not attempt to enter into the whole debate. However, privatization can be viewed as the refinancing of infrastructure by replacing public finance with private finance. Indeed, this financial motive for privatization is often the pre-eminent one. Thus, some of the main economic and financial implications of privatization will be discussed.

Specifically, we will discuss the implications of three scenarios:

- privatization is efficient and the infrastructure assets are sold at the market price;
- privatization is efficient and the infrastructure assets are sold at below market price;
- privatization is not efficient and the infrastructure assets are sold at the market price.

The main reason for selling publicly-owned infrastructure assets is efficiency. Economic theory and evidence suggest that, when a market is effectively competitive, assets are managed more efficiently under private ownership than under public ownership (Megginson and Netter, 2002; Abelson, 2003). Efficiency here means that private firms will provide more of the goods and services that people want at lower cost than will a publicly owned operation. There are therefore benefits to consumers or producers, or both. When the assets are sold at market price, the existing owner can capture much of this improvement in value. Financially the government gains because the capital income obtained from the sale exceeds the present value of the revenues that the government foregoes as a result of the sale. Thus, the public net worth rises as a result of the sale. This net worth can be used to finance other infrastructure, to reduce government debt, or for various other causes.

In our second scenario, privatization is still efficient. The total benefits from privatization exceed the total cost, and the net social benefit is positive. However, if the government sells assets at below the market price, it passes some benefits of the privatization to future owners of the asset. This does not reduce the overall economic benefits from privatization, but it does change their distribution. In this case the capital income from the sale may be less than the present value of the surpluses that the government foregoes as a result of the sale. Thus, the public net worth falls as a result of the sale. However, the public debt may not fall. The government can still use the income from asset sales to reduce debt. What has fallen is the equity component of the public net worth.

It should be noted here that there are transaction costs of selling assets. The costs of sales include the real resource costs associated with the transaction, for example marketing costs, and the payment to underwriters for taking the risk of picking up shares that are not sold on the market. Total transaction costs can run to between 2 and 5% of revenue from the sale.

In our third scenario, privatization is not efficient. For one or other reason, most likely in a monopolistic market, services after privatization may be less attractive to consumers or produced at higher cost, or both. In this case the primary economic arguments favor continued public ownership. Moreover, even if the assets were sold at the market price, the public net worth would fall unless the sale exploits a monopoly market power that the government would not exploit if it were the provider of ongoing services. These implications suggest that privatization should not occur when it is inefficient.

But, as we have seen, although the public net worth may fall as a result of the sale, debt may not fall. The government can still use income from asset sales to reduce debt or to provide short-term electoral presents. Undoubtedly this happens. Moreover, there are second-best situations in which sale of assets would be justified in this kind of situation. A second-best situation exists when the government cannot adopt optimal fiscal policy because of some constraint, and is forced to choose between second-ranked options. If the government faces major constraints on the revenue that it can raise from taxation or from borrowing, the sale of assets may be justified as a revenue measure to provide, say, more schools or more public transport than would otherwise be possible. However, the optimal policy would be to remove the constraint, say to public borrowing, rather than to sell the asset.

6. Conclusions

Transport infrastructure can be financed by current taxation, general or infrastructure-specific public borrowing, or by privately organized finance. The preferred method of finance depends on overall value for money, equity, and

Table 2
Advantages and disadvantages of instruments for raising capital

Capital raising	Advantages	Disadvantages
Taxation		
Consolidated revenue		Limited source of finance Inequitable Lacks efficiency incentives
Infrastructure levies	Can be equitable	Lacks efficiency incentives
Public sector borrowing		
Borrowing – general bonds	Low cost of finance Suitable for public projects	Inappropriate if debt is high
Infrastructure funds/bonds	May attract private funds	No cost advantage
Public enterprise borrowing	May encourage efficiency	May be a cost penalty
Private sector financing		
Private debt/equity	May encourage efficiency	Often high cost option
Mixed private instruments	Can combine advantages of public and private sectors	Requires clear contracts and efficiency allocation of risks

efficiency considerations, not just on the cheapest source of finance. Thus, the choice of method depends partly on the cost of finance but also on the role of user charges and on the ownership and management of infrastructure assets associated with each form of finance.

Table 2 summarizes the main advantages and disadvantages of the principal methods of financing transport infrastructure.

Current taxation measures are generally not effective, efficient, or equitable methods of raising finance for infrastructure. Current taxation can raise a limited amount of funds, imposes high tax rates (with consequent economic distortions) on current households, and imposes a large burden on the current generation of taxpayers.

General public borrowing is a low cost method of borrowing that spreads the burden of payment equitably. It is appropriate when user charges are not feasible (and so private financing is difficult) and when the government carries a high proportion of project cost and risk. It may also be appropriate when the market is not competitive. Specific infrastructure bonds have few advantages, although revenue bonds may be useful in some situations.

Privately financed infrastructure is a high-cost form of finance. However, when user charges are feasible, private ownership and management may produce goods and services more efficiently than does the public sector. These ownership and operational efficiencies may offset the relatively high cost of finance.

Similar arguments apply to refinancing existing transport infrastructure (privatization). This is worthwhile economically and financially when there are efficiency benefits and the assets are sold at market prices. Privatization is more difficult to justify in the absence of economic benefits, though there may be occasions when the financial benefits justify it.

In general, financing instruments should be chosen on the microeconomic merits of each case rather than on macroeconomic considerations. In parlous economic times, avoiding fiscal deficits and reducing debt are desirable objectives. At other times they are not sensible aims.

References

Abelson, P. (1999) "The real incidence of imposts on residential land development and building," *Economic Papers*, 18:85–90.

Abelson, P. (2003) *Public economics, principles and practice*. Sydney: Applied Economics.

Access Economics (2003) *Financing infrastructure for residential development*, Report. Canberra: Housing Industry Association.

Arrow, K.J. and R.C. Lind (1970) "Uncertainty and the evaluation of public investment decisions," *American Economic Review*, 60:364–378.

Domberger, S. (1995) "What does privatisation achieve – a comment on Quiggan?" *Australian Economic Review*, 95:43–48.

Megginson, W. and J. Netter (2002) "From state to market: a survey of empirical studies of privatisation," *Journal of Economic Literature*, 39:321–389.

New South Wales Treasury (2002) *Private provision of public infrastructure and services*, TRP 02-3. Sydney: NSW Treasury.

Quiggin, J. (1997) *Inquiry into road funding*. Submission to House of Representatives Standing Committee on Communications, Transport and Microeconomics Reform. Canberra: Parliament of Australia.

This Page Intentionally Left Blank

Chapter 21

THE WORLD BANK AND TRANSPORT[a]

KEN GWILLIAM
University of Leeds

1. The World Bank Group

The International Bank for Reconstruction and Development (IBRD) was established by the 1944 Bretton Woods Agreement on lending to governments to assist the reconstruction after the Second World War. Its sister institution, the International Monetary Fund, was the instrument for securing a stable and increasingly liberal world financial system.

The early efforts of the IBRD focused on post-war reconstruction, including part financing of the Japanese Shinkansen railway. But as the process of reconstruction was completed, the bank shifted its focus to that of assisting the development of lower-income countries. During the last decade, under the presidency of James Wolfensohn, the focus on poverty alleviation has sharpened, while the balance of attention has shifted from traditional economic growth (associated with the trickle-down philosophy) to more direct attention on the general quality of life, including environmental quality and the elimination of social deprivation. Human and social capital is emphasized rather more, and physical capital rather less. Institutionally, the emphasis has shifted from assisting the development of public agencies as service suppliers to that of securing effective supply through a competitive private sector, with public agencies acting as facilitators, procurers, and regulators.

The organization has also developed. The World Bank Group now consists of five linked institutions. The original IBRD continues to lend to governments on terms closely linked to the cost at which it can borrow capital on the international commercial markets. This is beneficial to its client countries with poorer credit ratings than the World Bank, and which thus would have to pay much more for borrowing directly. However, as countries become richer and financially stronger

[a]All judgements and interpretations in this chapter are those of the author and do not necessarily represent the views of the World Bank, its executive directors, or the countries they represent.

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

this advantage declines. In addition to the IBRD the most important of the agencies in the group are the International Development Association (IDA) – which lends money subscribed by the richer industrial countries to governments in the very poorest countries on extremely favorable terms – and the International Finance Corporation – which lends to private enterprises in the developing countries.

1.1. The policy context for World Bank lending

As an international development institution, managed by a board representing the interests of both borrowing and donor member countries, the World Bank Group has an obligation to engage in activities that are designed to serve the best long-term interests of all its members, and not just to engage in such lending as represents the most profitable, or even the most secure, use of the resources available. The implication of this is that it does not engage explicitly in a process of ranking all of the investment opportunities that are open to it. Instead, it has to seek a political balance in directing its resources to assisting member governments to achieve their national objectives, primarily by making the best use of their own national resources, to which World Bank lending is a supplement. For that reason, much emphasis is placed on the institutional and regulatory environment into which the World Bank is lending.

To some extent the pattern of World Bank lending reflects high-level priorities that the board of the bank itself has adopted. These include the avoidance of environmentally damaging investments, the widespread distribution of the benefits of projects throughout the national recipient community, and the avoidance of uncompensated losses by virtue of spatial or occupational displacement resulting from a project (World Bank, 1996). Strict standards are applied both to the environmental design of projects, which have to have formal environmental clearance before they are submitted for board approval, and to the resettlement and involuntary employment severance. Most recently, the adoption of the Millennium Development Goals (MDGs) emphasizes the social objectives to which development aid should contribute.

In a formal sense transport is unrecognized in the MDGs, which cover such things as standards of health, education, sanitation, and housing. But in practice it is fundamental to the achievement of virtually all of them through its effects on national income and its effects in determining accessibility to social services. It contributes to growth by facilitating trade, both nationally and internationally, and by improving the access of people to jobs, education, and other services. Both sector studies, such as that of crop production in Egypt (Esfahani, 1987), and country studies on economic policy and development, such as those undertaken for Indonesia (World Bank, 1992), and for India (World Bank, 1994), stress the importance of infrastructure to country growth at the stage when private sector growth is beginning to strain available infrastructure capacity. At the macro-economic level, cross-country studies have confirmed that inadequate transport infrastructure is an important constraint on aggregate agricultural productivity (Antle, 1983), and that investment in this sector raises growth by increasing the social return to private investment without "crowding out" other productive investment (Easterly and Rebelo, 1993).

These arguments are now once again being recognized, not least because of the emphasis given to them by the governments of the developing countries themselves. Hence, after a period during which the Bank's attention to infrastructure investment wavered, investment in basic infrastructure is again seen as one of the fundamental planks of a development strategy.

1.2. The World Bank process

The process through which that development strategy is pursued is essentially one of progressive refinement of the understanding of national priorities. At the start of the process has been the identification, jointly by the country and the World Bank, of a Country Assistance Strategy (CAS). This document identifies the borrowing capacity of the country, in both fiscal and administrative senses, and given the definition of such an envelope, the areas of high priority where the World Bank is thought able to contribute best to achieving national development objectives. This is reflected at the project appraisal stage by a requirement to demonstrate what value is added particularly by the World Bank's involvement in project finance. For the highly indebted poor countries (HIPCs), the country level Poverty Reduction Strategy Paper (PRSP) is now the context within which sector strategies and interventions are presented.

Where the transport sector is identified as a priority for assistance there is usually a sector review, within the World Bank, which focuses further on the transport sector, including consideration of the institutional and regulatory impediments to its effective development, and from which individual projects are in principle derived. This is referred to as economic sector work (ESW). Individual projects are then initially identified and presented for internal consideration in the World Bank in a Project Concept Document (PCD) or Project Concept Note (PCN). The full economic justification, including that of the selection between alternative actions, has then to be more fully worked up for inclusion in a Project Appraisal Document (PAD), now called the Staff Appraisal Report (SAR).

Within the SAR, which is the basis for presentation for board approval, a logical framework ("logframe") indicates the way in which project objectives, project components, and outcomes are related, and identifies the main perceived risks and the elements of the project design to mitigate such risks. A summary of the full economic evaluation must be included. Demonstration of the consistency of the project with the CAS objectives must also be formally shown in the appraisal report. The final details, including any covenanted conditions, are agreed in a negotiation between the borrower and the World Bank, and the agreed package presented to the board.

1.3. The lending instruments

As the objectives of the World Bank have developed, so have the range of instruments that it uses. In addition to the traditional project investment loan, which finances specific physical investments, there is a wide range of instruments tailored to specific needs. These include the structural adjustment loan (SAL), which offers general financial assistance to governments undertaking significant macroeconomic reforms, the sector adjustment loan (SECAL), which finances a range of expenditures within a sector undertaking specific reforms, and the technical assistance loan, which finances the reform or technical improvement process itself.

The transport portfolio

Since it started lending for transport in the late 1940s, the World Bank has provided nearly US $55 billion in loans and credits through over 1000 projects in direct support of transport sector development. Additional funding, amounting to roughly US $15 billion, was provided through the transport components of lending operations for agricultural and industrial development. The share of transport lending in the World Bank's overall lending has varied substantially. Up to the end of 1955, it averaged 18%, increasing to 40% between 1956 and 1965, and then declining to 30% between 1966 and the early 1970s. Since then, it has fluctuated between 13 and 16%. The composition of transport lending has also changed substantially, with the share of road (now 60%) and urban transport lending (now 10%) increasing at the expense of railway and port lending.

The World Bank has used a variety of instruments in the sector. The majority have been specific investment loans. Sector loans, such as the highway sector loans to Thailand and Indonesia in the early 1990s, have committed the bank to a program of expenditure over a number of years on the basis of a prior agreement on the objectives, general composition, and priorities of a subsector policy and related multi-annual investment program. SECALs, such as those made to Burkina Faso in 1992 and Senegal in 1994, have covered a wide range of sector policy and management issues, usually including several transport modes within a single lending operation. World Bank financing of transport vehicle fleets has been decreasing, as the bank has concentrated on assisting the transfer of enterprises from the public to the private sector (as in the case of buses in Barbados), and in reducing the constraints on private initiative (as in the case of the Mexican trucking industry). This has been pursued through narrowly targeted policy reforms in the context of infrastructure projects (as in the case of road freight haulage privatization in Yemen) or, more usually, as an integral part of broader SALs (e.g. the privatization of air transport in Peru). In other cases, such as trucking privatization in Hungary, Poland, and Russia, the World Bank has had an indirect influence through the general policy dialog on transport sector reform.

2. Investment project appraisal at the World Bank

2.1. The role of formal cost–benefit analysis

Transport investment projects are routinely subjected to economic appraisal by the World Bank, with net present value and internal rate of return estimated, and the standard comparisons between alternative project designs made. However, the primary concern is not to compare returns on investments in different countries but to make sure that resources are used sensibly in each recipient country, in particular avoiding "white elephants." Hence, the general requirement is to achieve an internal rate of return of at least 10–12%. Although it may be argued that this is too high, the general view appears to be that funding is sufficiently scarce, and the real opportunity cost of capital so high, that it is appropriate to set such a relatively high cut-off rate. The effectiveness of even this device for avoiding wasteful investment is, of course, limited by the underlying fungibility of finance, as countries may substitute World Bank funding for domestic funding of "good" projects in order to release domestic funds for more politically motivated schemes. Covenanted loan conditions are frequently employed to defend against this, and are particularly likely to be effective in very poor countries that are heavily dependent on external funding.

Benefit estimation usually follows generally accepted procedures of estimation of flows and link performances with and without the project. For road projects these are converted into aggregate benefits using standard national vehicle operating cost and benefit valuation conventions for such "intangibles" as time and pain and grief costs of accidents or loss of life, if available. Where such national standards are not available operating cost values may be constructed synthetically using Highways Development and Management System (HDM) operating cost model parameters, and time values constructed following suggested default values contained in a World Bank infrastructure note.

The general recommendation in the World Bank *Handbook on Economic Evaluation* is that economic evaluation should be undertaken at domestic prices, in real terms. This implies that all market prices should be adjusted to be net of taxes or subsidies; that the non-labor resource cost of non-tradable commodity or service inputs need no adjustment; and that border prices (cost, insurance, and freight (CIF) for imports, free on board (FOB) for exports) for internationally traded goods need adjustment in proportion to the ratio of the official exchange rate and the market rate. Market prices may also be adjusted where current wage rates substantially misrepresent the real cost of labor, either because of substantial underemployment of unskilled labor or because at prevailing remuneration rates many developing countries have severe shortages of skilled labor, which should therefore logically be assigned a premium value. The issue of the value of labor is particularly relevant to the selection between labor-intensive or capital-intensive production technology such as in road maintenance or rehabilitation. Otherwise, the tendency is increasingly to use market wage rates.

There has been a tendency to treat savings in operating costs in transport projects as more "real" than savings in time, with rates of return initially excluding time values, and enhanced rates including time valuation given as an extra. That is changing. Time savings are increasingly accepted as a legitimate element of benefit, and there is a sector guidance note that recommends the estimation of local values, but also suggests some default principles of evaluation in the absence of such local values.

The valuation of accident cost savings has been even more controversial. While there has been no objection in principle to the inclusion of the non-human resource cost benefits of accident reduction, the question of valuation of pain and grief and, in the extreme, loss of life has been very controversial. The minimum position, consistent with the view that the World Bank is assisting client countries to make effective use of their own resources, is to accept valuation of accident savings at whatever rate is currently adopted internally in the country. The implicit judgement is then clearly that of the country, and not of the World Bank. It is now a common requirement that safety audits are performed on project designs.

Environmental impacts are always considered. All projects are pre-classified according to whether they have zero, small, or large environmental impact. Those with non-zero impacts are required to have environmental impact assessments, and to contain mitigating measures to counter any adverse effects. This mandatory requirement covers the more obvious, immediate consequences of projects. It does not, however, deal with more subtle effects, either positive or negative, associated with traffic generation or modal shift effects. While there is no objection in principle to the inclusion of such environmental effects within the central economic evaluation, it is rarely done, except in cases that are primarily viewed as environmental projects. This partly reflects the absence of adequate data on the physical impacts of specific interventions, as well as the absence of evaluation conventions. Only rarely, as for example in the appraisal of a metro project in São Paulo, are the impacts evaluated in monetary terms and included in the central appraisal.

Risk is also considered in the appraisal document, with a requirement that the major risks of the project are identified, and mitigating measures included in the project design. It is also addressed in the economic evaluation, albeit frequently in the relatively trivial form of showing the sensitivity of the rate of return to a number of separate eventualities (changes in assumptions about cost, construction period, traffic growth rate, etc) and demonstrating the switching point with respect to these variables, either singly or cumulatively. More recently, however, Monte Carlo simulation methods have been used in some projects to estimate a distribution of rates of return based on what are considered to be reasonable assumptions about the range and distribution of specific risks.

2.2. The limitations of formal appraisal

Any economic evaluation is only as robust as its inputs. The procedures used for forecasting traffic in World Bank-funded projects vary greatly between projects. In some cases, for example for some recent urban projects in Brazil, there is a well-established database and traffic model available locally. But growing attention has been given to traffic forecasting, especially in urban transport operations. For example, a great deal of effort has been made in Chinese urban transport projects to improve the application of network-based urban travel demand models in forecasting traffic under alternative scenarios. The most common systematic structure for benefit estimation is probably seen in the application of HDM to road maintenance and rehabilitation projects. More usually, however, traffic forecasts have to be constructed on an ad hoc basis for individual projects.

Structural effects of major inter-urban projects are a major problem. Particularly where major international flows of traffic are concerned (e.g. in recent trade facilitation projects in southern Europe) or where significant new inter-regional connections are being developed (e.g. in the case of the Jamuna Bridge in Bangladesh), this will require heroic, and generally untested, assumptions about the effect of new infrastructure on trade flows. In a recent evaluation of a port investment in Poland, the relative costs of routes not even within the country were critical, and the appraisal had to be based on some very *ad hoc* estimation of the costs and flows on alternative routes though different Baltic ports. Similar issues have arisen in the evaluation of urban rail projects in Brazil.

Low-volume rural roads, and particularly feeder paths and tracks programs, have created particular problems. First, where initial traffic volumes are very low, there are more than normally difficult problems of traffic estimation. Second, because much of the movement is non-vehicular, there are problems of evaluation of benefits to non-motorized transport (NMT), despite some recent efforts to address this in the context of the Sub-Saharan Africa Transport Program. Third, the infra-marginal nature of the impact on basic access to some activities – for example, school and clinic attendance, which became possible with the project and were impossible without – means that they seem to be less convincingly proxied by measured transport benefits than in the case of the more marginal economic activity changes stimulated by improvements in existing networks. Fourth, the cost of detailed appraisal of individual small projects tends to make less sense than that of projects for which the project appraisal costs are a smaller proportion of project costs. For all these reasons, the World Bank has tended to rely on appraisal of programs rather than projects, and to concentrate on the cost-effectiveness of the design of specific interventions rather than a full cost–benefit analysis in these circumstances (Gannon and Lebo, 1999).

3. The institutional and policy context

While the technical issues that have been discussed so far are important, the experience of the World Bank is that projects that fail to achieve the desired return usually do so not because of failures in the technical design of the project, or weaknesses in the appraisal methodology, but because of more fundamental weaknesses in the policy and institutional frameworks within which the investments are made. Hence, in the transport sector, as in all sectors of World Bank activity, great emphasis is put on institutional and policy issues. A number of prime examples can be mentioned.

3.1. Maintenance strategies

Projects are usually evaluated on the assumption that proper maintenance of new infrastructure will be assured. The cost of appropriately scheduled maintenance is included within the project appraisal. In the case of investments in national road rehabilitation or construction it is common for the HDM software package to be used both to identify and to cost the optimal maintenance program, and for that optimized maintenance plan to be assumed. The latest version, HDM-4, has several new features, such as the capability of evaluating more types of vehicles, including NMT and the motorcycle, and the capability of evaluating rigid pavement. In the case of other modes, the treatment of maintenance is more ad hoc. In all cases, failures to undertake the optimized maintenance schedules would increase total system costs in the long run.

In practice, failure to maintain assets has been a perennial problem for developing countries. It has been estimated that US $85 billion worth of road assets were lost in Africa due to lack of maintenance during the 1970s and 1980s. This was approximately equal to the sum of all multinational investment in roads on the continent over the same period.

Deferring maintenance until roads need major reconstruction results in total road agency costs that can, in total, be up to three times those involved in a policy of effective current maintenance. Even using a discount rate of 12% the net present cost of a policy of deferring maintenance has been estimated to be more than one-third in excess of that of a timely maintenance policy. Although deferral is usually the result of low management capability or political interference, at the simplistic level of government financing it may even seem rational. In many developing countries the fiscal position is so weak that there may be many potential uses of public funds that can yield economic (and even financial) rates of return well in excess of the conventional 10–12% rate of return cut-off for public investments. Hence for the government it may appear perfectly rational to direct expenditures elsewhere.

The policies pursued by many national aid agencies, as well as by some of the multilateral agencies and development banks may have contributed to this sense of rationality. If soft funding is available for new investment or major reconstruction, but not for current maintenance expenditures, the calculations are shifted even further in the direction of postponing maintenance. The more favorable the terms for funding reconstruction compared with those available for funding routine maintenance, the greater the incentive to the recipient governments to defer maintenance.

But even that does not make deferral of road maintenance truly economically rational. The essential problem is that vehicle operating costs are a high proportion of total costs of transport, and increase progressively as road conditions deteriorate. It has been estimated that for a road carrying 500 vehicles per day, which is already in poor condition, every US $1 saved in the annual expenditure on road maintenance will increase the total operating costs of vehicles on that road by between US $3 and US $6 in the same yearly period (Heggie, 1995). What may appear to be rational politically (because votes can be more easily won by highly visible new investments than by timely maintenance) or even fiscally (because external funding is more cheaply obtained for reconstruction than for maintenance), is therefore not economically rational, however low the discount rate.

Financial failure to maintain assets not only pre-empts the possibility of those assets contributing to economic growth but also, and consequently, excludes the achievement of social and environmental objectives. Improved maintenance budgeting for public infrastructure, cost recovery for operations, and the identification and proper maintenance of that infrastructure and service network that can sustainably be afforded given the priorities of the government concerned is therefore of the highest priority. Not all existing infrastructure or service should be maintained. Where mistakes have been made in the past, or where technological progress has made capacity redundant, emphasis should be put on the adjustment of capacity. That is inevitably problematic, as the policy involves loss of welfare to the residual users of the redundant or technologically superceded infrastructure. How that should be handled is more a matter of welfare distribution than of efficiency.

The logic of that position is often blurred by governments and misunderstood by critics. In the absence of adequate mechanisms for direct income transfers, well-targeted, properly funded, subsidies of basic consumption goods and services may be a perfectly sensible instrument of redistribution policy. But inadequately funded, or badly targeted, subsidies may be inefficient mechanisms for the desired transfers, and may be self-defeating if they accelerate the decline of the service provided. Governments will, of course, wherever possible take the opportunity of blaming an external agency for fare increases or service reductions, in circumstances where they are unwilling to take the actions and face the consequences of those actions necessary to avoid that outcome. Critics then naively hold the lender responsible for what, in effect, is the policy of the government.

There are actions that a lender can take to forestall the outcome of asset deterioration. But they are not easy either to identify or establish. Loan conditions can in some cases be used to prevent governments from investing in new infrastructure that they cannot afford to sustain or on which the return is much lower than that on increased maintenance expenditure. Commitments can be sought concerning adequate allocations for maintenance; this can even go as far as the introduction of earmarking of some tax revenues for this purpose. But commitments concerning the future actions of a government are notoriously difficult to enforce. Moreover, earmarking of taxation has some serious consequences for macroeconomic management, and is, in any case, vulnerable to changes in government policy. And of course, earmarking does not ensure that funds will be well spent.

For those reasons, attempts are now being made to develop the commercialization of infrastructure as a means of improving performance in infrastructure, particularly roads in Africa, and has been discussed by Heggie and Vickers (1998). This involves putting roads on a fee for service basis, under which the level of expenditure on maintenance (and ultimately on capacity creation) is determined by the willingness of users to pay for that service. This involves:

- creating ownership by involving road users in management of roads;
- stabilizing road financing by securing an adequate flow of funds from user charges;
- clarifying the location of responsibility for road provision;
- strengthening management by creating proper structures of management responsibility.

Road boards with user representation are being established in many African countries, with explicit road tariffs being set, consisting of vehicle license fees and a fuel levy.

The achievement of the benefits of second-generation road funds has some demanding institutional requirements. There must be a strong legal basis for users to be assured that extra charges will actually be devoted to improved maintenance. To that end, an independent executive must be managed by a strong user-controlled board of management. And there must be credible external review processes to ensure that management has not been captured by limited factional interests. Second-generation road funds with these general characteristics were established in several African countries during the 1980s and 1990s, and similar arrangements are being considered for countries as diverse as El Salvador, Guatemala, Jordan, Lebanon, and Pakistan. Although the implementation of these second-generation road funds has not been without difficulties, and ministries of finance have tended to maintain a strong position in the management of the funds, a review of experience with them by Gwilliam and Kumar (2002) has identified substantial improvements in maintenance performance associated with the more commercial processes that they have adopted.

3.2. The role of the private sector

Private participation in infrastructure finance is of great current political interest, not least in the former command economies such as Hungary and Poland. Major programs of concessioning of toll roads have been already been undertaken in Mexico, Malaysia, and Argentina. Thailand has a more limited experience, while China has extensive plans for infrastructure expansion employing various forms of public–private participation. But many governments are grossly over-optimistic about the extent to which they can obtain the infrastructure development that they would like off-budget simply by opening the sector to unsolicited private proposals. Not all of these have been successful as purely commercial propositions. In both Argentina and Mexico, contracts have had to be renegotiated to avoid collapse. In both Mexico and Malaysia, much of the debt finance came from state-owned banks. In Thailand, the opening of a privately financed urban expressway in Bangkok was delayed pending resolution of a conflict between the government and the concessionaire over the level of tolls to be charged.

Initial experience has thus demonstrated the enormous difficulty of internalizing a sufficiently large proportion of the benefits of economically justifiable road or urban rail infrastructure to make pure private developments commercially viable. That does not make private sector participation undesirable, but it does raise a new set of questions about the identification and evaluation of public sector contribution to public–private partnership (PPP) schemes, concentrating particularly on the need to evaluate the external or off-route benefits that cannot be appropriated by the private project investor, and relating the acceptable amount of public contribution to the magnitude of those benefits.

For transport service provision, emphasis is now increasingly placed on separating responsibility for operational management from that for social policy, and on developing contractual arrangements for the financing of unremunerative public service obligations. That approach, already widely applied to railway undertakings in the industrialized countries, is being recommended by the World Bank to its developing country clients.

Associated with that approach, the World Bank has strongly encouraged the development of private sector operations. In trucking that has not been difficult, as only in the transitional economies of Eastern Europe and the socialist economies of China and Vietnam is the public sector important. But the same is not true of rail transport where the perceived need for regulation has been associated with public sector operation. In that sub-sector the World Bank has encouraged the development of private sector concessions or franchises. Almost all of the rail freight systems of Latin America are now privately operated, as are a number of the smaller African systems. In addition, large suburban passenger systems have been successfully concessioned to the private sector in Argentina and Brazil. Similarly, in the port sector, the World Bank has assisted a number of countries to develop "landlord port" models in which only the basic infrastructure remains in public ownership but most of the landside activities are privately owned or concessioned.

In urban bus transport the World Bank has, in the past, attempted to improve the efficiency of public sector operators by technical assistance (as in Sri Lanka) and investment in vehicles (as in Russia and several other CIS countries) (World Bank, 2002). The view has been reached that such assistance is of fleeting impact. Consequently, the emphasis has shifted to assisting the development of a more commercial approach to bus transport through competitive tendering of route franchises (Gwilliam, 2001). This has been introduced in Kazakhstan, Uzbekistan, and the Kyrgyz republic, and is presently being developed through a training program in franchising for Russian cities. As Gwilliam (2003) has shown, there are still many problems to face, but the impact of mobilizing the private sector in this way has made urban public transport more sustainable.

A common feature of developing countries is the very large role played by the informal, non-corporate, sector in the supply of public transport services. This comprises a wide range of technologies, from the rickshaw pullers of Bangladesh, through the peculiar indigenous jeepney of the Philippines, to the air-conditioned special buses in Argentina. It also comprises a wide range of regulatory responses from route licensing for the Manila jeepneys, through total vehicle stock constraints as in Bangladesh, to effective free entry (which often occurs despite formal regulatory requirements). The general dislike of the informal sector by governments – and the dislike of free entry into the transport market more generally – is the lack of operating discipline that fragmentation engenders. Not only does this take the form of unsafe behavior such as racing and overtaking on the road (which should in principle be controllable by the normal processes of law) but also in the development of patently uneconomic operating practices. A common phenomenon in many countries is the development of the practice of controlling departures from terminals to secure full loads on departure. This may be enforced by associations (minibus/taxis in South Africa) or by terminal operators (Bangladesh, Sri Lanka). This has the effect reducing the proportion of vehicle time spent in revenue-earning business, making it impossible to board buses outside terminals, and hence adding walking distance to trip lengths, and undermining schedule reliability.

Some constraints on such chaos are often self-imposed by "operator associations," which flourish particularly in Latin America and West and South Africa. Such associations typically establish property rights – often over terminal positions or slots – by force of arms if necessary. Once entry is restricted by some means, it becomes in the interests of members of the association to establish some stability and discipline in operations in order to protect the market value of the product and to distribute costs and revenues "equitably" among members. The devices to achieve this (e.g. holding vehicles in terminals until they are fully loaded) do not necessarily maximize consumer or total social welfare. However, in the context of a more regulated situation such as that in Buenos Aires, where overlapping route monopoly franchises are granted to individual associations, operator associations may be effective in reconciling fragmented ownership with disciplined service management.

4. Focusing on poverty impacts

It is the declared objective of the World Bank to concentrate its efforts on the reduction of poverty. The most common response in the rural transport context is to emphasize local participation in the selection and design of activities, through which it is believed that both better initial selection and greater community commitment to maintain the improved infrastructure will be achieved. Certainly this is a way of overcoming the problems associated with roads being constructed by a non-local authority without any clear allocation of funding for subsequent maintenance.

It must be recognized, however, that this very pragmatic approach sits uneasily with traditional cost–benefit evaluation procedures. If one of the intentions is to choose "pro-poor" investments, then the traditional reliance on "willingness to pay" as the source of the valuation conventions seems to be logically less supportable. As a consequence, in addition to experiments with benefit weighting discussed by De Walle (2002), there have recently been some attempts to devise more systematic analyses of the distributional impacts of urban projects. For example, in an urban rail rehabilitation project in Fortaleza, Brazil, which included some restructuring of bus routes to act as feeders to the rail system and the introduction of multi-modal transferable ticketing arrangements, the impact of the project on travel times and trip costs has been disaggregated very finely by zones with different average income levels. As the zones are small and relatively homogeneous, this allows the distributional effects of projects to be much more clearly observed.

The World Bank's work on local participation in the feeder road area, driven by the perceived benefits of local ownership and decision-making, has focused primarily on building applicable tools for local communities and local governments to plan and undertake their own (simplified) analysis and planning process. These analytical tools are typically based on one of the following approaches:

- the application of simplified or modified versions of standard economic decision analysis;
- a locally adapted (calibrated) version of multi-criteria analysis – usually based on some combination of observables (e.g. traffic, social trips, and economic activity) and chosen social weights (typically related to population within the projects "area of influence"), or simply by undertaking an implicit weighting of investment benefits (e.g. using time-based weights such as in the integrated rural accessibility planning approach);
- a form of cost-effectiveness analysis – as in the basic access approach which used dollars per population served as a indicator of the relative value of competing investments.

The motivation for the development of these planning tools is to introduce some form of rational decision-making into a local planning process, loosely related to principles of economic analysis and the desire for social/human development. However, the introduction of these tools has mostly been an attempt to ensure that the World Bank's investment programs are productive against given development objectives – rather than ensuring ownership or local decision-making.

In the urban context, the World Bank emphasis on poverty alleviation has been embodied in its attempts to improve public transport in the major cities. This has included investments in buses in Russia and Central Asia, in segregated busways in Peru and Colombia, and in urban rail rehabilitation in Brazil and Argentina. In all of these contexts, the emphasis on ensuring the sustainability of the improvements has led the World Bank to concentrate as much on the institutional and regulatory bases for operation of the systems as on the physical investments themselves. Even in rural areas, Starkey et al. (2002) have demonstrated the need for, and possibilities of, improving transport services.

Some very special problems arise in a number of East Asian countries such as China and Vietnam, and South Asian countries such as Bangladesh, where the predominant form of mechanized personal transport has been the bicycle. In those countries there is very little motorized public transport, and what there is often more expensive and little quicker than the bicycle because of its relative inflexibility and inability to exercise its potential speed advantage due to (predominantly bicycle caused) congestion. However, governments tend to associate non-motorized transport with poverty, and therefore tend to ignore it or restrict it in the interest of motorized vehicles. This impacts particularly on the poorest people who have no alternative. In addition to this disadvantage, the most obvious trade-up as income increases is not to the bus but, as already apparent in Vietnam, to a motorcyle – and eventually to a car. In the most congested cities of East Asia there is insufficient road space to accommodate that trend without severe congestion and environmental impact. The World Bank has therefore attempted not only to support the development of public transport but also to finance measures to develop facilities for pedestrians and cyclists in a number of cities, such as Lima, Peru, and Accra in Ghana.

The high levels of morbidity and mortality associated with air pollution and road accidents in developing countries – usually an order of magnitude greater than those of the industrialized countries – impacts particularly on the poor, who are most vulnerable. They have therefore been of particular concern to the World Bank. For both air pollution and direct poverty reduction reasons, the bank has encouraged better provision for non-motorized transport, as well as supporting cleaner technologies, such as improved maintenance of motor cycles in Bangladesh and the replacement of two-stroke engines by four strokes. In the safety field it initiated a global road safety partnership to heighten attention to road safety throughout the world. Safety audits are now common in the design of infrastructure projects, and a first free-standing road safety project is now being prepared in Vietnam.

5. Summary

As part of its mission to reduce poverty, the World Bank has consistently made about 15% of its investments in the transport sector, believing that it contributes significantly to growth and to poverty reduction. These investments are subject to formal economic evaluation, and must be consistent with a country assistance strategy agreed between the bank and national governments at the highest level. For several decades these investments were concentrated on the provision of basic infrastructure – mainly rail (in the early years) and roads. But the bank has recognized that infrastructure investment is not enough, as failure to maintain that infrastructure and inefficiency in the organization of transport services on it have undermined the effectiveness of the investments made. The World Bank has therefore moved progressively to greater concern for the underlying sector institutions and policies, which it attempts to achieve through both loan conditions and direct investments in reform activities.

References

Antle, J.M. (1983) "Infrastructure and aggregate agricultural productivity: international evidence," *Economic Development and Cultural Change*, 31:609–619.

De Walle, D. (2002) "Choosing rural road investments to reduce poverty" *World Development*, 30:575–589.

Easterly, W. and S. Rebelo. (1993) "Fiscal policy and economic growth," *Journal of Monetary Economics*, 32:417–458.

Esfahani, H.S. (1987) "Growth, employment and income distribution in Egyptian agriculture, 1964–1979," *World Development*. 15:1201–1217.

Gannon, C. and J. Lebo (1999) "Design and evaluation of very low-volume rural roads in developing countries," *Transportation Research Record*, 1652:82–91.

Gwilliam, K.M. (2001) "Competition in passenger transport in the developing world," *Journal of Transport Economics and Policy*, 35:99–118.

Gwilliam, K.M. (2003) "Bus franchising in developing countries: some recent World Bank experience," in: *8th International Conference on Ownership and Competition in Passenger Transport*. Rio.

Gwilliam, K.M. and A.J. Kumar (2002) *Road funds revisited: a preliminary appraisal of the effectiveness of "second generation" road funds*, Discussion paper TWU-47. Washington, DC: World Bank.

Heggie, I.G. (1995) *Management and financing of roads: an agenda for reform*, Technical paper 275. Washington, DC: World Bank.

Heggie, I.G. and P. Vickers (1998) *Commercial management and financing of roads*, Discussion paper TWU-32. Washington, DC: World Bank.

Starkey, P., S. Ellis, J. Hine and A. Ternell (2002) *Improving rural mobility: options for developing motorized and non-motorized transport in rural areas*, Discussion paper TWU-48. Washington, DC: World Bank.

World Bank (1992) *Economic policy and development in Indonesia*. Washington, DC: World Bank.

World Bank (1994) *India – recent economic developments and prospects.* Washington, DC: World Bank.

World Bank (1996) *Sustainable transport; priorities for policy reform.* Washington, DC: World Bank.

World Bank (2002) *Cities on the move: a World Bank urban transport strategy review.* Washington, DC: World Bank.

Chapter 22

ASSET MANAGEMENT AND FUNDING: TRANSPORT AND NON-PROFIT INSTITUTIONS

JAMES H. MILLER, JILL HOUGH and LYN HELLEGAARD
North Dakota State University, Fargo, ND

1. Introduction

Non-profit organizations, as contrasted with for-profit corporations or government entities, are widely used throughout the transportation sector to own assets, coordinate services, and operate transportation systems. Non-profit organizations are used for a wide variety of purposes for both passenger and freight transportation; however, this chapter of the handbook will focus on the role of non-profit organizations in the passenger transportation sector. Further, the institutional/legal framework for this form of organization and the case examples are based on practice in the USA, though similar types of organizations perform transport functions in many nations.

2. The role of non-profit corporations in passenger transportation

Three types of organization are commonly used to coordinate and/or provide localpassenger transportation, including public agencies (governmental units), private, for-profit, and private, non-profit corporations. Whereas public passenger transportation in the USA was once primarily a private, for-profit corporate activity, since the 1960s, nearly all local passenger transportation has received significant public tax and charitable support, and is now provided by government agencies or non-profit corporations.

Nearly all large urban public transit operations are under the direction of a general government agency (a division of a local government) or are provided by a special-purpose entity such as a transit authority, district, or commission. Many of these public entities contract the actual operation of the transit service to private, for-profit corporations, but the responsibility for determining the amount of transit to be provided, the price charged (the fare), and arranging for the financial

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

support of both operating and capital expenditures falls on the public body. In these large urban areas, responsibility for general public fixed-route transit generally falls to public agencies, but specialized para-transit services that provide demand-responsive transportation for special market segments, including elderly, disabled, and low-income residents, often fall to non-profit corporations. These non-profit transportation providers often receive public funds through the public transit systems or directly from transportation and human service funding agencies.

Likewise, in non-urban areas, rural transportation services are also provided by a combination of public agencies and non-profit corporations. A 2000 survey of rural transportation providers receiving US Department of Transportation funding under the Section 5311 program found that about 35% of the grantees were private, non-profit organizations, and 62% were government entities; only 3% were private, for-profit organizations. The other 2% were tribal organizations (Community Transportation Association of America, 2001). These data probably understate the significance of non-profit corporations in providing transportation services, since many specialized human service transportation providers do not seek US Department of Transportation transit funding. The York County Community Action Agency described in the case studies later in this chapter illustrates this type of organization that owns assets and directly provides transportation services.

Direct provision of service is not the only role for non-profit transportation organizations. In many communities the role of the non-profit transportation organization is to coordinate and broker transportation services, not provide them directly. In these cases, the role of the non-profit corporation is to arrange rides for clients of specific human services agencies or riders with specific needs, and then contract for services with government, for-profit, and non-profit providers. A key role of the non-profit broker is to arrange funding and monitor service quality for users. Wheels of Wellness, Inc. of Philadelphia is one non-profit transportation organization that performs this brokerage role, and is described below, in the case studies section. The Bis-Man Transit Board of Bismarck, North Dakota, also provides this brokerage role, and coordinates both local and intercity passenger transportation.

Another commonly found role for non-profit transportation organizations is to encourage efficient, environmentally sound transportation choices, especially in large urban areas, through the operation of transportation management associations or organizations. Transportation management associations (TMAs) or transportation management organizations (TMOs) have been formed by developers, groups of businesses in a particular area of a large city, or as a region-wide entity by local governments to reduce single-occupant vehicle travel by promoting ridesharing, transit, and alternative work arrangements, including staggered work hours and telecommuting. The case study on the Missoula-Ravalli Montana TMA (see below) illustrates this type of organization.

3. Characteristics of non-profit corporations

A non-profit corporation is a legal entity formed for a purpose other than generating a profit. Further, the income derived from the activities of the corporation cannot be distributed to directors or officers. The exact legal structure and organization of a non-profit corporation in the USA are set by the state in which it incorporated, and therefore these characteristics will vary from state to state. These state laws govern the number and composition of the board of directors, meeting and reporting requirements, and other aspects of the governance and accountability of the organization. Nevertheless, the basic structure and functions of non-profit corporations providing transportation functions are similar throughout the USA.

A survey by Salamon in 1995 revealed that the non-profit sector in the USA included more than 1.6 million organizations (Salamon, 2001). Approximately one-quarter of the non-profit bodies were member serving, such as social/fraternal organizations, business and professional associations, labor unions, or political parties. The other three-quarters were public-serving organizations that included funding foundations, churches, action agencies, and service providers. Most of the organizations described in this section are of this later, service provider designation. However, TMAs and some brokerages might be considered member-serving agencies.

Two key features of non-profit corporations important to transportation organizations are the tax status of the organization and the ability of the organization to accept charitable contributions of cash, assets, or other in-kind goods and services. These attributes are controlled by the tax status of the organization as recognized by the US Internal Revenue Service (IRS). In addition to meeting the incorporating and filing requirements of the state in which a non-profit corporation is formed, the organization must apply for tax-exempt status with both the state and federal governments.

The IRS code recognizes more than 30 specific types of tax-exempt non-profit corporation. The specific classification depends on the purpose of the organization and the activities to be undertaken. These classifications determine the reporting requirements and whether a corporation can receive charitable contributions. The most common type of transportation non-profit corporation is one that conforms to the requirements of Section 501(c)3 of the IRS code that applies to organizations that are religious, educational, charitable, educational, or literary in purpose. Non-profit transportation providers that serve elderly, disabled, and low-income clients usually seek this designation as a charitable organization, thus enabling charitable contributions to support their operations. The two other most common IRS designations for transportation non-profit corporations are under the provisions of Section 501(c)4 and Section 501(c)6. These designations are often appropriate for brokerages and transportation management organizations that do not depend on charitable contributions and that may be formed by several private, for-profit businesses (US Internal Revenue Service, 2003).

As indicated earlier, the roles for private, non-profit corporations in transportation are varied, and include ownership of assets and/or coordinating or operating transportation services. Likewise, the characteristics of founders and incorporators of private transportation vary widely. In some cases, private, for-profit firms form non-profit entities to function as brokers or coordinators. TMAs often are formed by for-profit corporations. Private companies also form non-profits to provide other benefits for the incorporators. For example, in one large urban county a number of private taxi companies formed a non-profit brokerage to handle scheduling and billing for human transportation service contracted transportation services.

Government entities also form non-profit corporations to own particular assets used for transportation, such as vehicles or maintenance facilities. The separate non-profit corporation offers the benefit of joint governance over the use of the assets, such as in the case where a number of governmental and non-governmental organizations pool resources to accomplish a particular transportation function. Government agencies also use private, non-profit corporations to hold assets or employ personnel "off budget," i.e. to keep the personnel costs and counts as well as other expenditures from being counted in the cost/size of the government agency. Ferris (2001) presents a critique of this governmental practice of creating "off-line" agencies, and questions the loss of accountability for government funds that results from the use of non-profit corporations to provide governmentally sanctioned and funded services.

Private, non-profit transportation corporations are also formed by concerned individuals or other charitable non-profit organizations to advance a particular transportation cause. Wheels of Wellness in Philadelphia was formed by concerned citizens to provide medical transportation. Non-profit human service agencies often incorporate separate private, non-profit coordinated transportation systems to provide needed services for their clients. In all cases, these new agencies may own vehicles and fixed facilities and also engage in direct operations, or they may broker or contract for services.

4. Advantages and disadvantages of the private, non-profit form of organization

Community groups wishing to organize to meet a particular transportation need such as those described earlier can use any one of a number of legal structures to respond to the need, including the private, non-profit corporation. How, then, is the choice made to form a non-profit corporation versus a for-profit or governmental unit? The factors influencing this decision are both objective and subjective. First, at least in the USA, the choice of forming a private, for-profit organization is usually ruled out, because the transportation purposes for which the organization is being formed will not result in a positive net income over expenses. Success of the transportation venture will depend on receipt of government grants and other funds plus charitable contributions. Therefore, the primary choice faced by organizers of the transportation program is between a government agency providing the service and a private, non-profit corporation taking on these tasks. Often, for political reasons, a governmental agency may not want to expand into a new program area or increase the size of government, and therefore the private, non-profit option is the only logical option available to provide a needed service.

Objectively, the benefit of the private, non-profit form of organization is that it enjoys the limited liability afforded to all corporations, it is exempt from federal income taxes, and, perhaps most important, it can receive both public and private grants and accept tax-deductible donations. It can own assets, employ personnel, and enjoy other benefits afforded corporations. However, unlike governmental agencies, it cannot tax or exercise the right of eminent domain to acquire property. Further, some public grant programs are only available to public agencies. However, in many cases, non-profit transportation organizations accommodate this limitation by working with a governmental entity that receives grants and holds title to assets (see the Bis-Man Transit Board case, below). Likewise, governmental units are often exempt from certain taxes, such as those on motor fuels, and non-profit agencies providing the same services may not be exempt. However, this disadvantage is eliminated in some states that offer the fuel tax exemption based on the type of service provided (public transportation) as opposed to the legal characteristics of the entity providing the transportation.

The decision to form a private, non-profit transportation organization is also based on a number of subjective factors. First, most often the charitable purpose of the transportation entity is to meet the transportation needs of the disadvantaged. Other organizations present in a local community that meet other needs of these same individuals are non-profit corporations such as senior citizen clubs, mental health organizations, or other types of human service agencies. These organizations refer clients and often pay for at least part of the cost of rides and are most comfortable dealing with another non-profit corporation. They understand the culture and motivation of these types of entities as compared to a government agency. Often staff or board members for these organizations make up the board of directors of the transportation organization. Likewise, representatives of businesses forming transportation management associations make up the board of the TMA.

Finally, non-profit organizations are also more likely to attract contributions of funds and assets as compared with a government agency. In addition, individuals are more likely to volunteer their time and resources for a non-profit entity rather than for a government organization.

In summary, the private, non-profit organizational option is chosen for two reasons. One is that it is the default option after the governmental agency option is eliminated, because either the proponents of the transportation activity or the governmental policy-makers determine that the transportation purpose is not best suited for a public agency response. The other more proactive reason for selecting the private, non-profit option is that it is the best structure to meet the transportation purposes intended. The ability to include users and funding agency representatives on the board of directors, the ability to receive charitable and grant contributions, and the ability to attract volunteers all help contribute to the attractiveness of the non-profit option.

5. Case studies of non-profits

To illustrate how private, non-profit organizations have been structured to meet a variety of transportation needs throughout the USA, four brief case studies are provided below. These case studies illustrate organizations that directly operate transportation services or broker rides between users or their funding organizations and transportation providers, and a transportation management organization that promotes ride sharing through vanpooling and coordination of human service transportation.

5.1. York County, Maine

York County Community Action Corporation (YCCAC), in Maine, is a non-profit organization that works to help York County residents be self-sufficient by providing transportation options to those who need access to community programs in area of some 2570 km^2, with about 190 000 residents. Over 17% of the residents have some type of disability.

In the 1960s a war on poverty helped initiate the community action agencies. York County Community Action Corporation (YCCAC) was incorporated in 1965. Its focus is to "Help people to help themselves." Its written mission is to alleviate the effects of poverty and promote self-sufficiency, administer programs to bring relief to those who are poor, and respect and preserve the dignity and self-worth of the people it serves.

The YCCAC has six specific programs servicing the county, which include: Affordable Housing, Community Services (crisis situations), Energy Services, Head Start, Transportation, and WIC (a nutrition program for children and pregnant women). Transportation was added to the mix in 1969.

Today, the YCCAC continues to provide necessary services to the residents of York County. YCCAC's transit manager has served the residents of York County since 1980, and watched the transportation portion of YCCAC meet the growing needs of the residents. Four transportation services are provided to residents. These include Bus Transportation, My Bus – Sanford Transit, Volunteer Driver, and Wheels to Access Vocation and Education (WAVE).

Bus Transportation is a service offered to each of the towns in York County on a weekly basis. Routes are designed to carry passengers to the closest regional shopping and medical facilities. The residents must arrange their rides 24 hours in advance by calling the YCCAC. The rides are required to fit into the regularly scheduled routes. The residents are picked up at their door, and dropped off at their destination. The YCCAC requires all riders to complete a brief application, which is short enough to be completed over the phone. Fares are based on one-way service, and elderly, disabled, and children under 8 years of age pay half-fare; low-income individuals are asked to pay what they can afford; individuals covered by Medicaid who are going to Medicaid-eligible destinations pay no fare; and individuals covered by other contracts will be told what fare they will need to pay when they make their trip reservation, e.g. Division of Mental Retardation, or vocational rehabilitation.

My Bus – Sanford Transit is a fixed-route transit service that operates 5 days a week between Springvale and South Sanford. Anyone can ride this service on a fare-paying basis. The fares range between US $0.25 for elderly, disabled, and children under 8 years old, US $0.50 for the general public, and US $0.75 for commuters to South Sanford industries.

The Volunteer Driver program has approximately 85 volunteer drivers who help meet the transportation needs of those individuals whose needs cannot be met by other forms of transit provided by the YCCAC. The volunteers transport residents whose needs include medical services covered by Medicaid, child protective and foster care, elderly persons at risk, and other residents with medical needs. The volunteers operate their own vehicles, and are reimbursed for distance traveled and tolls only. The volunteers donate their own time and driving skills. To request a trip, residents need to contact the YCCAC's Volunteer Driver program at least 48 hours in advance of their trip.

The WAVE program transports residents to jobs and training sites as well as riders' children to day care. The vans used for this program operate for several shifts and run 7 days a week. Residents need to call 48 hours in advance if any changes are made to their schedule. The fares vary depending on location where the individuals are picked up (home versus van stop), in what town they live, and where they are going.

The assets utilized by YCCAC transportation include the building where they are located, vehicles, and employees. The YCCAC utilizes its assets to the maximum. They own 21 buses, of which 17 are used for routes with four spares, and six vans. They employ 17 bus drivers, six van drivers, and one driver/trainer. YCCAC transportation also employs 15 office staff. The staff includes a program director, assistant director, office manager, two data processors, and a receptionist. Bus operations include a coordinator, an assistant coordinator, and a scheduler. The Volunteer Driver program includes a coordinator, an assistant coordinator, and three schedulers. Another specialized program, the Job Access Reverse Commute (JARC) program, has an office manager.

The YCCAC experiences operating costs of US $3 018 670 (for buses and vans). The vehicle-kilometers traveled in 2003 totaled 1 006 733 km (for buses and vans). The costs break down to US $3.01 per vehicle-kilometer and US $78.10 per vehicle-hour. Passenger-kilometers traveled are 6 878 846 km, which includes buses, vans, and volunteer drivers but not taxis or JARC services. There were a total of 201 484 passenger trips (bus and van) and 12 702 Sanford Transit trips, which equates to 214 186 trips. The cost per trip for the YCCAC totaled US $14.09.

When interviewed about the advantages of operating as a non-profit provider, the transportation director of the YCCAC identified key advantages, including:

- greater level of autonomy;
- ability to move faster with decisions;
- more creative and able to do things on a lower budget than a for-profit organization;
- can be a direct recipient for government funds.

The main disadvantage for the YCCAC of being a non-profit organization is the shortage of funds to provide all of the services its clients want and need. Another disadvantage the agency faces is getting vendors to recognize non-profits so that they are able to receive tax rebates, particularly for fuel.

5.2. Bis-Man Transit Board, Bismarck, North Dakota

The adjoining cities of Bismarck and Mandan in North Dakota were the recipients of funding for public transportation; however, neither city wanted to provide the services, so they brokered out to the Bis-Man Transit Board. The Bis-Man Transit Board is a non-profit organization that was initiated in 1989. In turn, the Bis-Man Transit Board contracts with two entities, to provide the operational and management aspects of the services. The Central Nodak Development Corporation maintains the management of the transportation, while a taxi company maintains the operational agreement.

The mission of the Bis-man Transit Board is

> To identify, promote, coordinate, and establish transportation services for elderly and handicapped individuals in Bismarck and Mandan areas, to identify and meet transportation service needs of the elderly and handicapped individuals as they relate to employment, medical, recreational issues; to establish a transportation network serving elderly and handicapped individuals utilizing federal, state, county, and local funding sources.

The eligible riders of the city transportation include those aged 60 years and over, as well as certified disabled residents. Disabled children must be over 8 years old to ride independently. Residents are asked to make their reservations a day in advance, and are provided door-to-door service. The fares for these services are US $1.50 within city limits and US $2.25 to cross city lines.

The management budget totals US $2.0 million. Funding comes from a variety of sources including, federal (5311 funds), state, and county funds, and a local levy of three mills in Bismarck and Mandan, as well as funds from United Way.

The assets for the Bis-man Transit Board include a building housing the local taxi and bus service and the Greyhound bus depot. These three entities moved into the building in 1998. The city donated the land for the building. The funding for the building came from the Federal Transit Administration and local match. Sales tax played a lead role in the local contribution. There are 26 small buses, three vans, and two cars that provide transportation services to the adjoining cities.

The executive director of the Central Nodak Development Corporation, operator of the Bis-Man Transit Board, stated that the advantages of being a non-profit organization are that services can be provided to individuals that need them at a lower cost than a for-profit organization would charge. The major disadvantage is dealing with boards and having to jump through several hoops for decisions to be made.

5.3. Wheels of Wellness, Inc., Philadelphia, Pennsylvania

Wheels of Wellness, Inc. was formed in 1959, and is the oldest transportation non-profit corporation in the USA. Founded as a project of the Junior League of Philadelphia, it was originally an all-volunteer organization with the purpose of providing non-emergency medical transportation services. Its service area includes the city of Philadelphia and the surrounding four counties in Pennsylvania, and four additional counties in New Jersey. From a small volunteer organization, Wheels of Wellness has grown into an organization with more than US $22 million in annual revenue that provides 2.6 million rides.

Keeping with its founding purpose of providing free medical transportation to persons who cannot afford to pay for such necessary services, Wheels of Wellness raises about US $0.5 million per year that is not reimbursed by grants and service contracts to provide this free service. The United Way and a number of other organizations and individuals make charitable contributions to Wheels of Wellness to pay for these rides. In addition, Wheels of Wellness sponsors a number of fund-raising efforts throughout the year. Though it receives the bulk of its funding from service contracts, the contributions are a key to meeting its mission and a key benefit from being a private, non-profit, charitable corporation.

Wheels of Wellness is an example of a non-operating brokerage. Most of the nearly 10 000 daily rides are provided by seven for-profit carriers. In addition, Wheels of Wellness helps riders learn to use the extensive fixed-route transit system (Southeastern Pennsylvania Transportation Authority), and funds these rides. Wheels of Wellness employs just under 100 persons, most of whom are involved with call taking, dispatching, and customer service. Only recently has Wheels of Wellness begun to operate its own vehicles for a very small, specialized part of its effort.

The Wheels of Wellness executive director stated that the private, non-profit form of organization was key to his agency's success. The organization is governed by a 28-member board that includes representatives of funding agencies, representatives of user groups, and community leaders. The non-profit culture of the organization is key to its meeting its mission. The only negatives associated with being a non-profit entity, and the executive director stated that these shortcomings were minor, were the additional reporting requirements associated with being a charitable non-profit organization, and the requirements concerning retained income that are a part of the US Internal Revenue Service's regulations for the Section 501(c)3 type of organization.

5.4. Missoula Ravalli Transportation Management Association, Missoula, Montana

One special type of private, non-profit transportation organization that has become increasingly common in the USA is the TMA. These organizations can be controlled by a board of directors or their stakeholder members. Organizations that are member-driven typically provide services to a particular area, such as the downtown business district, medical centers, industrial parks, or government entities. Organizations governed by a board of directors are typically organized to address regional transportation demand management (TDM) issues.

TMAs can be formed by regional or local governments, chambers of commerce, clustered businesses, or management of a major employer such as a hospital or university. Most TMAs require subsidies, whether it is seed funding, which could be obtained through federal Congestion Mitigation Air Quality (CMAQ) funds, with matching funds coming from local entities or membership dues. It should be noted that CMAQ funding will only fund newly formed TMAs for 3 years with a match requirement. This short-term pilot project funding rather than long-term operational funding, the lack of support among stakeholders, and the perception that any benefits are small, as many TDM measures take time to reach capacity, create real barriers to the sustainability of TMAs.

TMAs provide an organizational structure to implement TDM strategies and services. Overall, TMAs are more cost-effective than a program managed by individual employers. An example would be a TMA organized in a business district that would allow all employers and employees within that district to benefit from a TDM program. These strategies and programs may be too costly for a small business owner to implement individually.

TMAs can provide a wide range of services that promote more efficient use of transportation and parking resources. Most often TMAs do not own significant transportation assets (other than office and administrative equipment), but they do provide a number of functions, including the following:

- rideshare matching for carpooling or vanpooling;
- guaranteed ride home services;
- commuter financial incentives and education;
- school outreach programs;
- telework support;
- marketing and promotion of bicycle, pedestrian, mass transit, and other alternative modes of transportation;
- special event transportation management;
- flextime support;
- parking management;
- land use planning support.

TMAs can support or provide parking management and brokerage services that result in the maximum use of parking resources. An example might be a movie theater, which normally operates during non-commute hours, allowing their parking facilities to be used as a park-and-ride location for a local shuttle service during the day. This can not only reduce the need to expand parking capacity, and the amount of paved land in a community, but also address single-occupancy vehicle issues. TMAs formed by clustered retailers are a means to reduce employee parking demand, thus allowing limited parking to be dedicated to shoppers. By metering parking the TMA can generate operating funds as a means of offsetting some or all of the program costs, which otherwise would have to be borne by its members.

As a case study, the Missoula Ravalli TMA (MRTMA) in Missoula, Montana, is a private non-profit organization that was established in 1997 with a mission to develop and implement creative solutions to single-occupancy vehicle use as an alternative to road widening of a major commuter route and to address US Environmental Protection Agency air quality attainment issues. This is the only TMA in Montana. MRTMA serves the five surrounding communities, involving employers, businesses, medical providers, government agencies, the university system, and individuals. It owns and operates nine 15-passenger vans, customized for the long commute. The TMA is governed by a board of directors, with representation from the government, local transit providers, the university system, the parking commission, job services, and stakeholders. In an effort to maximize local transportation resources, commuter vans are used weekdays from 9 a.m. to 3 p.m. to supplement the local transit service's paratransit program. MRTMA provides:

- vanpool services;
- ride-matching services for carpooling or vanpooling;
- a guaranteed ride home;
- seniors and adults with a disability transportation service;
- a school outreach program;
- transportation coordination program support;
- transportation information and referral;
- special events transportation.

Summary

Non-profit public transportation organizations have been in operation around the world for decades, although this chapter has concentrated on the US case. Non-profit public transportation organizations can be handled differently, as addressed by the four case studies presented. The case studies highlight successful, long-term programs where the non-profit organizational form has been used to coordinate and/or provide transportation services. The examples chosen are typical of the approaches used in the USA for transportation of the general public and special user groups, and for directing programs to improve the efficiency of existing facilities and services through transportation management organizations.

References

Community Transportation Association of America (2001) *Status of rural public transportation – 2000*. Washington, DC: CTAA.

Ferris, J. (2001) "The double-edged sword of social service contracting: public accountability versus nonprofit status," in: J.S. Ott, ed., *The nature of the nonprofit sector*. Boulder: Westview Press.

Salamon, L.M. (2001) "Scope and Structure: the anatomy of America's nonprofit sector," in J.S. Ott, ed., *The nature of the nonprofit sector*. Boulder: Westview Press.

US Internal Revenue Service (2003) *Tax exempt status for your organization*. Publication 557, Washington, DC: US Internal Revenue Service.

Part 4

REGULATORY ISSUES

This Page Intentionally Left Blank

Chapter 23

DEREGULATION OF US AIR TRANSPORTATION

STEVEN A. MORRISON
Northeastern University, Boston, MA

1. Introduction

This chapter deals with the "economic" deregulation[a] of US domestic passenger transportation by air. Accordingly, the chapter will not be concerned with regulatory reforms in domestic and international air freight, or with international air travel. Also, "non-economic" regulation of safety (the responsibility of the US Federal Aviation Administration) and security (the responsibility of the US Transportation Security Administration) will not discussed. Although the chapter pertains to the US experience only, those interested in the domestic passenger transportation of other countries may find the following discussion useful.

After 40 years of regulation, US domestic passenger airlines were deregulated in 1978. Airline deregulation has received considerable attention over the years for several reasons. It was the first of many industries in the USA to be deregulated, and as such served as a source of insights into the likely effectiveness of deregulation in other sectors. Second, the vast amount of data available – both during and after regulation – makes detailed analyses possible. Finally, many people find the airline industry interesting in its own right, and want to understand developments in that sector.

This chapter begins with a brief history of airline regulation in the USA, and outlines the forces that led to deregulation in 1978. Next, the paths of important measures of industry performance (e.g. fares, profits) are analyzed. Following that is a look at some trouble spots where performance could be improved. The chapter concludes with a look to the future and a quantitative assessment of the effects of deregulation.

2. History[b]

The first scheduled commercial passenger flight in the USA took place in 1914 when, for US $5.00, the St Petersburg-Tampa Airboat Line carried passengers

[a]The deregulation of the prices a firm is allowed to charge and the markets it can enter and exit.
[b]This section relies heavily on Morrison (2002).

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

18 miles (29 km; 1 mile = 1.6 km) between Tampa and St Petersburg, Florida. However, it was mail, rather than passengers, that was the primary cargo for the next several years. The US Post Office began operating a regular airmail service in 1918. By 1927 all carriage of mail had been contracted out to private carriers. Carriers were compensated by the Post Office based on the capacity (i.e. space-miles) of the aircraft they used. This gave carriers the incentive to use larger aircraft than were necessary to carry mail and to use the extra space to carry passengers. Passenger transportation without a mail contract was not financially feasible – thus, the Postmaster General had effective control over entry into the airline passenger transportation business.

Formal economic regulation of the transportation of passengers by air in the USA began in 1938, when Congress created what was to later become the Civil Aeronautics Board (CAB). In order to receive a certificate to provide airline service, an airline had to show that it was "fit, willing, and able" to provide the service and that the service was "required by the public convenience and necessity." The 16 carriers in existence when the Act was passed were granted certificates under a "grandfather" clause in the legislation. Although the board occasionally allowed entry of new classes of carriers (e.g. feeder carriers), until 1977, the year before deregulation, it never allowed entry onto a route that was already served by more than one carrier. In addition to regulating entry by new carriers into the airline business and entry by existing carriers onto new routes, the CAB also regulated the fares that airlines could charge. However, subject to meeting minimum frequency of service requirements, frequency of service was not regulated. One of the board's responsibilities was to foster the development of aviation. Thus, in order to increase the amount of air travel, the board set fares below cost for short-haul routes (where travelers have viable alternatives to travel by air) and set fares for long-haul routes above costs (where the alternatives to air travel are less attractive). The profits from high long-haul fares were to be used to internally cross-subsidize the losses from the low short-haul fares. However, because frequency of service was not regulated, carriers reduced frequency of service on unprofitable short-haul routes (as much as was allowed) and increased service (and other amenities) on long-haul routes in order to attract more passengers, thus reducing or eliminating the excess profits from long-haul transportation.

As early as the 1950s, economists began criticizing CAB regulation. Over time, more and more analysts came to believe that the US airline industry did not have characteristics that made economic regulation necessary. An interesting facet of airline regulation in the USA was that federal regulation by the CAB did not apply to intrastate airlines, which operated solely within one state. States such as California and Texas provided the setting for interesting "controlled experiments" of sorts, because these states had large cities far enough apart to warrant air service. By comparing fares and service patterns in California and Texas with the fares and service patterns that prevailed on similar routes that crossed state lines,

Table 1
A historical profile of the US airline industry

Year	Enplanements (× 1000)	Revenue passenger-miles (× 1 000 000)	Operating revenue (× 1000)
1930	418	93	NA
1940	2 966	1 152	135 301 (a)
1950	19 220	10 243	839 920
1960	62 257	38 863	2 884 779
1970	169 922	131 710	9 289 658
1980	296 903	255 192	33 727 806
1990	465 560	457 926	76 141 739
2000	666 150	692 757	130 838 619
2001	622 129	651 700	115 526 896
2002	612 877	641 102	106 987 095
2003P	646 523	655 850	115 449 719

Note: (a) Data arc from 1941.
Source: US Air Transport Association.

the effect of regulation could easily be seen. One particularly influential study pointed out that in 1965 the fare charged by the intrastate carrier Pacific Southwest Airlines (PSA) between San Francisco and Los Angeles (338 miles) was US $11.43, while the fare charged by CAB certificated carriers between Boston and Washington, DC (400 miles) was US $24.65 (Levine, 1965). This simple comparison provided compelling evidence to legislators that regulation of airlines resulted in higher fares.

In addition to the case for deregulation provided by academic studies, other events during the 1970s contributed to the eventual passage of the Airline Deregulation Act. Wide-body jets were introduced in 1970 during an economic recession in the USA; the combination of these two events resulted in overcapacity – too many seats chasing too few passengers. As a result, the CAB imposed a moratorium on new route cases and allowed carriers to make agreements to reduce capacity on major routes. These actions were widely criticized. Next, in 1973, fuel prices rose sharply due to the Arab oil embargo. Deregulation was seen as a way to fight the resulting inflation and reduce the role of the government in the economy. Finally, the public's attitude about the appropriate role of government in the economy had evolved since the 1930s – skepticism of unregulated markets had been replaced by skepticism of government regulation.

As a result of these forces, beginning in 1976 and continuing until 1978 the CAB began interpreting the regulatory statutes more liberally, and allowed some pricing and entry freedom. Finally, in October 1978, the Airline Deregulation Act was signed into law. The law provided for a phasing out of the CAB's authority over routes and fares. By 1983, provided they were fit, willing, and able, airlines were free to enter the domestic routes of their choosing and to charge the fares that they desired. In 1985, the CAB ceased to exist.

Table 1 offers a historical profile of the US airline industry, showing how enplanements (passenger boardings), revenue passenger-miles (each paying passenger traveling 1 mile counts as 1 revenue passenger-mile) and operating revenue have developed since 1930. What is striking in the table is the significant growth the industry has undergone since its beginnings – and the decline in travel and, especially, in revenue since 2000, in the wake of a soft economy and the terror attacks of 9/11.

During 2003, air travelers in the USA took nearly 400 million trips, involving 605 domestic airports and more than 45 000 unique origin–destination pairs.[a] Sixty-six% of them flew from their origin to their destination on a single aircraft. The remaining 34% changed planes during their trip, usually at a "hub" airport. Because there are economies of aircraft size (the cost of flying a larger aircraft is less, on a per-seat basis, than the cost of flying a smaller aircraft) on low-density routes (those with relatively few passengers) the combination of fare and departure frequency of the hub-and-spoke system is usually preferable to what is possible with non-stop service. This way, passengers from a given origin traveling to many different destinations are flown to a hub airport, where they change planes and are flown, along with others from many different origins, to their ultimate destinations.

3. Methodology

The appropriate methodology to assess the effects of a policy change such as airline deregulation is to construct a counterfactual. That is, the fares, service patterns, etc., under deregulation should be compared with estimates of what would have prevailed had airlines remained regulated. Nonetheless, the assessments of the effects of deregulation that follow are primarily factual in that they compare fares, service patterns, etc., before and after deregulation. However, the results of counterfactual analyses, when available, are reported to put the factual changes in perspective.

4. Effects

The effects of airline deregulation on key measures of industry performance are highlighted below, and assessed: route competition, load factors, fares, network structure, profits, and safety.

[a]These figures are based on calculations by the author using data in the US Department of Transportation's Origin–Destination Survey (Data Bank 1A).

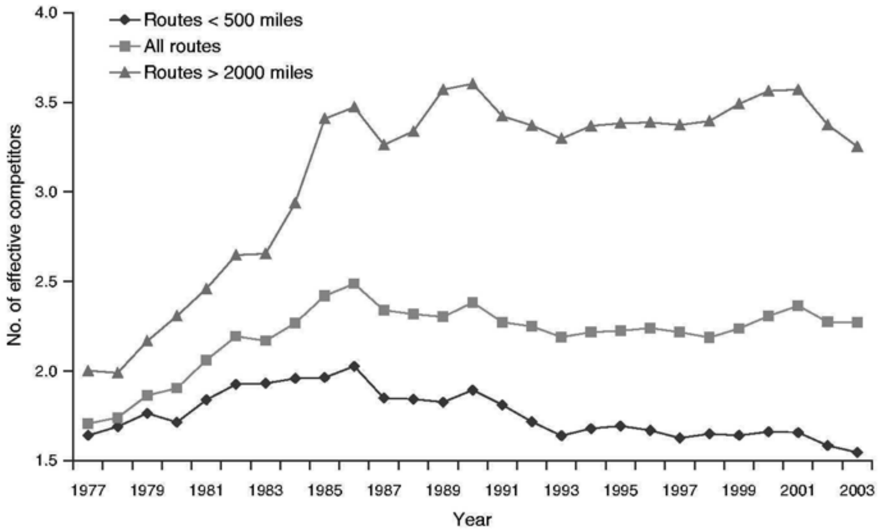


Figure 1. Domestic competition at the route level. (Source: author's calculations using data from the US Department of Transportation, Data Bank 1A.)

4.1. Route competition

In assessing the extent of competition in a market, it is not appropriate to simply count the number of firms serving the market, because firms with a larger market share exert more influence in a market than firms with a smaller market share. In Figure 1 the measure of the extent of competition is the number of "effective competitors," which is an inverted (numbers equivalent) Herfindahl–Hirshman index.[a]

Figure 1 shows the extent of competition at the route level for all routes, routes shorter than 500 miles, and routes longer than 2000 miles. For all routes, route competition increased from approximately 1.7 carriers per route in 1977 to a maximum of 2.5 carriers per route in 1986 before the merger wave that took place.[b] Since then, the number of effective competitors per route has fluctuated

[a]The number of effective competitors is the number of firms with equal market shares that would yield the same Herfindahl–Hirshman index (HHI). The HHI is the sum of each firm's market share squared. Thus, if two firms each had a 50% market share, the HHI would be $0.5^2 + 0.5^2 = 0.5$; inverting gives two effective competitors. Alternatively, if one firm had a market share of two-thirds and the remaining two firms each had a market share of one-sixth, the resulting HHI would also equal 0.5, resulting in two effective competitors.

[b]During 1986 and 1987, eight major airline mergers took place in the USA.

Table 2
Competitive profile of Southwest Airlines in 1998

Category	Variable	Percentage of domestic passenger-miles	
Actual competition			44.8
	Routes served	21.0	
	Nearby competitive routes	23.7	
Potential competition			49.4
Total			94.2

Source: Adapted from Morrison (2001).
Note: Totals may not add due to rounding.

between 2.2 and 2.4. By 2003, the number of effective competitors per route was nearly 2.3, an increase of 33% from the level in 1977. But the changes in the extent of competition on all routes masks what is happening at a more disaggregate level, particularly with respect to "short" routes and "long" routes. For routes longer than 2000 miles, the number of effective competitors increased from 2.0 in 1977 to 3.5 in 1986 before the merger wave. Although competition declined immediately after the merger wave to 3.3, since then it has fluctuated between 3.3 and 3.6 carriers per route. By 2003, there were nearly 3.3 carriers per route, an increase of 63% since 1977. The picture on short routes, however, is significantly different. Here, carriers per route increased from about 1.6 in 1977 to 2.0 in 1986. Since then, however, route competition has eroded to about 1.5 in 2003, a loss of about 6% since 1977.

The extent of competition on a route depends not just on the number of effective competitors on the route but also on the identity (i.e. business model) of the carriers serving the route, in particular the extent of competition from low-fare carriers. Figure 2 shows the percentage of domestic passenger-miles provided by low-fare carriers. This increased (from zero in 1978) to 7% in 1985, due primarily to People Express, which exited the industry through merger in 1986. Beginning in 1987, the presence of low-fare carriers has increased more-or-less steadily to the point where, in 2003, low-fare carriers accounted for over 20% of domestic revenue passenger-miles. This, however, tells only part of the story. The impact of low-fare carriers on airline competition goes beyond the fraction of traffic they carry because carriers competing against them lower their fares to be competitive. One measure of the extent of this impact is shown in the figure (the percentage of domestic passenger-miles (by all carriers) flown on routes where low-fare carriers have at least a 10% market share). This measure has followed the pattern of traffic carried by low-fare carriers, but clearly more passengers are

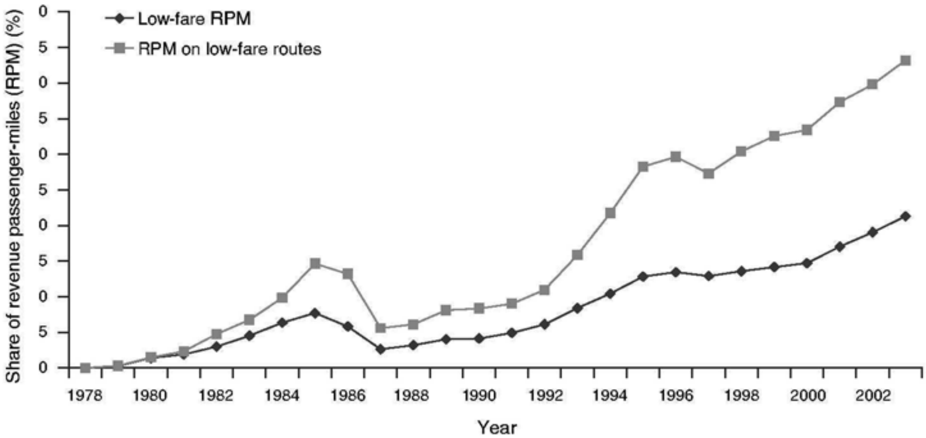


Figure 2. Revenue passenger-mile share of low-fare carriers and revenue passenger-mile share of routes with low-fare competition. Low-fare carriers are defined as: JetBlue, Frontier, Tower, AirTran, ValuJet, Morris, Kiwi, Carnival, Nations Air, Vanguard, Spirit, Pro Air, People Express, Reno Air, ATA, Western Pacific, Southwest, and Air South. Low-fare carriers had to have at least a 10% (collective) market share on a route for that route to be considered a low-fare route. (Source: author's calculations using data from the US Department of Transportation, Data Bank 1A.)

affected. By 2003, nearly 45% of passenger-miles were flown on routes with low-fare presence. And the trend is increasing.

Table 2 shows that the impact of at least one low-fare carrier, Southwest Airlines, goes well beyond the routes that it serves. Southwest's presence also affects fares on nearby routes and the possibility of their entry affects fares as well. In 1998, 21% of domestic passenger-miles were flown (by all carriers) on routes that Southwest Airlines served. But Southwest Airlines also influenced fares on nearby competitive routes affecting nearly 24% of domestic traffic.[a] In addition, potential competition from Southwest Airlines influences fares on routes that it does not serve, amounting to 49% of domestic passenger-miles.[b] In total, Southwest Airlines' actual competition and potential competition influenced fares on routes that make up over 94% of domestic air traffic.

[a]Suppose Southwest Airlines serves the route from airport A to airport B. Nearby competitive routes are defined as those routes with origins and destinations that are within 75 miles of airports A and B, respectively. For example, Southwest Airlines' presence on A–B also affects fares on routes A–D and C–D, where C is within 75 miles of A and D is within 75 miles of B.

[b]Consider a route A–B that Southwest Airlines does not serve. Fares on A–B will be influenced by Southwest Airlines if it serves airports A or B or nearby airports C or D (that are within 75 miles of A and B, respectively).

4.2. Load factors

Load factor, the percentage of available seats filled with paying passengers, is a very important measure that affects airfares and airline profitability. Other things being equal, a higher percentage of seats filled allows an airline to charge lower fares or to earn higher profits, or both. For example, as shown in Figure 3, in 1977 load factors averaged 55.9%. By 2000 they were 72.4%. Just this change alone would result in fares 23% lower, even if other factors remained constant. Before deregulation, load factors ranged from the high 40s to the mid-50s. Since then load factors have been increasing, more or less steadily, until in 2003, when nearly 74% of available seats were filled with paying passengers – a record high, with the exception of the Second World War and its immediate aftermath.

4.3. Fares

Interest in the extent of airline competition and in load factors exists because of their effects on fares. Trends in fares are typically expressed using the concept of "yield," which is the average fare per unit of distance traveled (e.g. airline passenger revenue divided by revenue passenger-miles). US domestic yield,

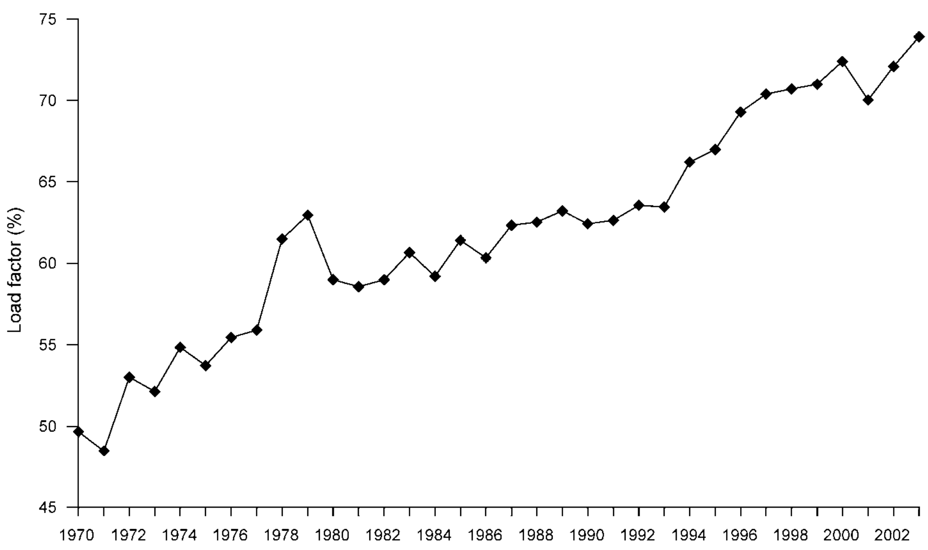


Figure 3. Percentage of seats filled with revenue passengers in the USA. (Source: US Air Transport Association.)

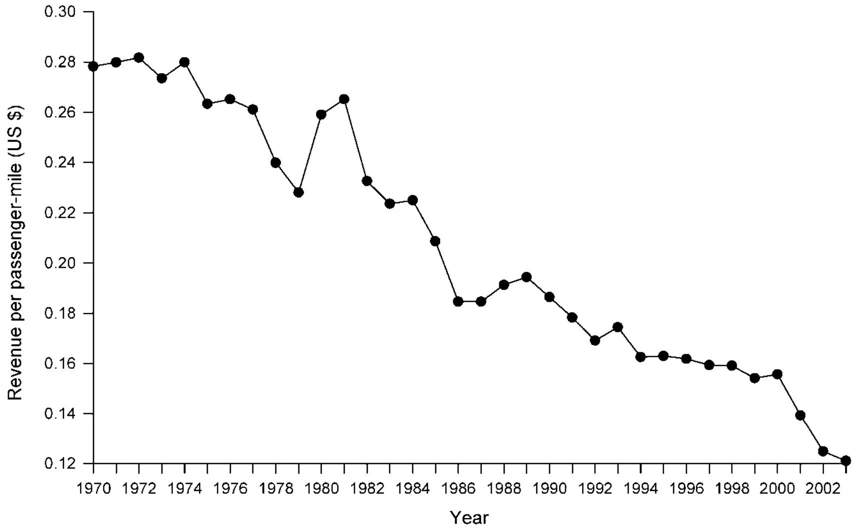


Figure 4. US domestic airline yield adjusted for inflation (US $, 2003 values). (Source: author's calculations using data from the US Air Transport Association.)

adjusted for inflation, is shown in Figure 4. The yield has been falling reasonably steadily. In 1977 the real yield was US $0.26. In 2000 it was US $0.156, a decrease of 40%. By 2003, in the aftermath of September 11, the sluggish recovery from the recession that began in early 2001, and the restructuring of the airline industry, the yield had fallen to US $0.121, a decrease of 54% since 1977. Of course, fares were falling before deregulation, and some of the observed decreases would likely have happened even without deregulation. A counterfactual analysis of fare changes under deregulation estimated that in 1998 fares were 27% lower than they would have been had regulation continued (Morrison and Winston, 2000). The same counterfactual analysis estimated that, in 1998, 80% of passengers (accounting for 85% of passenger-miles) paid lower fares under deregulation than the estimate of what regulated fares would have been in the same economic environment.

Figure 5 shows what has happened to fare dispersion in the USA since 1978. Here, fare dispersion is measured using the Gini coefficient – a larger Gini coefficient reflects greater dispersion in fares.[a] Dispersion increased slightly

[a]The Gini Coefficient multiplied by two equals the expected difference in price of two randomly selected tickets, expressed as a fraction of the average price in the market. Thus, a Gini coefficient of 0.20 means that two tickets selected at random would, on average, differ by 40% of the average fare in the market.

immediately after deregulation in 1978, leveled off for a few years, and then started a dramatic climb that ended in 1991, when the industry was experiencing weak demand and incurring significant losses. As the industry recovered, price dispersion again increased, reaching a peak in 2001, again followed by a decline as the industry again experienced weak demand and suffered significant losses.

4.4. Network structure

One enduring myth of airline deregulation is that the hub-and-spoke network structure is a product of airline deregulation. The hub-and-spoke system existed for many years in the USA before airline deregulation in 1978. For example, in 1977 29% of US domestic air travelers changed planes on their trips. In 2003, 34% of passengers changed planes. The percentage of people changing planes on their journeys has increased somewhat to be sure, but what has changed dramatically is the nature of those connections. In 1977 nearly half of all connections involved changing airlines (interline connection). In 2003, virtually all connections were on the same airline (online).

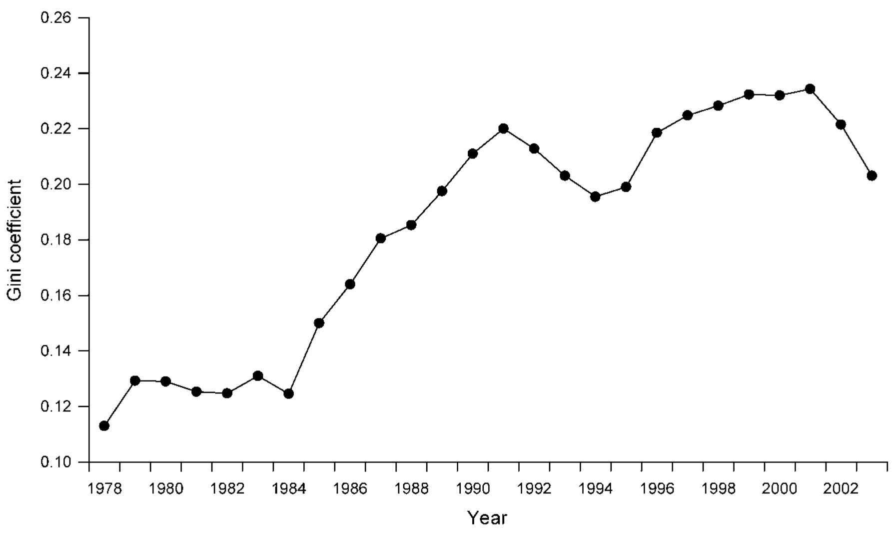


Figure 5. Fare dispersion in US domestic markets. (Source: author's calculations using data from the US Department of Transportation, Data Bank 1A.)

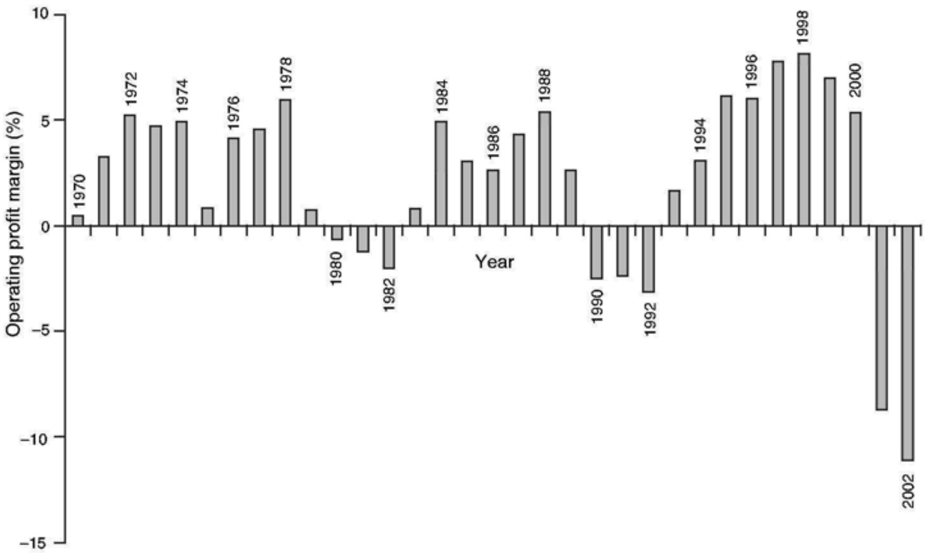


Figure 6. Operating profit margin (all services) for US scheduled airlines. (Source: Air Transport Association.)

4.5. Profits

The US airline industry is a cyclical one – profits vary with the business cycle. This is shown in Figure 6, which shows operating profit margin since 1970.[a] From 1970 to 2000 the operating profit margin averaged about 3%. Even given the cyclical nature of the industry, the losses that the industry has sustained in each year from 2001 to 2002 are quite large – indeed, each year's loss is larger than the industry has ever sustained in any one year since industry-wide profitability has been tracked, beginning in 1938. Nonetheless, a counterfactual analysis of the effect of deregulation on industry profitability suggests that in 1988 industry profits would have been US $2.6 billion lower if airlines had remained regulated. This is because, although fares would have been higher under regulation, costs would have been higher still and load factors lower (Morrison and Winston, 1995).

[a]Operating profit margin is operating profit (before taxes and interest on long-term debt) expressed as a percentage of revenue. An operating profit margin of 5%, for example, indicates that $0.05 out of every $1 collected in fares is available to pay taxes, to pay interest on long-term debt, and as profit to shareholders.

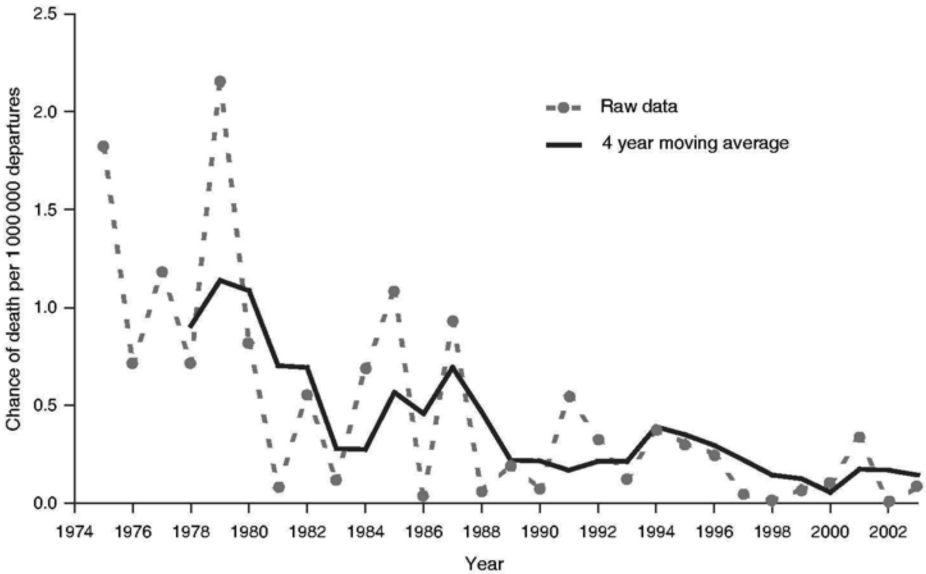


Figure 7. Air travel safety in the USA. (Source: author's calculations using data from the US National Transportation Safety Board.)

4.6. Safety

The first thing to note about safety is that it was not deregulated in the USA – safety continues to be regulated by the Federal Aviation Administration. Nonetheless, there was some concern that deregulation would force carriers to skimp on maintenance or hire less qualified pilots. Figure 7 shows the chance of death per million departures for the USA.[a] Because of the fluctuation in the accidents, the figure also shows a 4 year moving average, which has declined over the years.

5. Trouble spots[b]

Despite the success of airline deregulation in the USA, some trouble spots remain. Effective policies to deal with these remaining areas would enhance the already sizable benefits from airline deregulation.

[a]The chance of death is appropriately expressed on a per flight (departure) basis because the risk exposure in flying is related to take-off and landing more so than to the distance of the en route portion of the flight.

[b]This section relies heavily on Morrison (2002).

5.1. Gates

Although deregulation radically changed the domestic environment for US airlines, the system of providing airport services has changed much more slowly. Airports and airlines in the USA still enter into long-term exclusive-use leases of airport gates, just as they did during regulation. The US General Accounting Office (1990a) reports that 88% of the gates at 66 large and medium-sized airports are leased, and 85% of these are leased under exclusive-use terms, where the gate is effectively the property of the airline to use (or not use) as it sees fit. These contracts, and the difficulty of adding new capacity at airports, may make it difficult for new carriers to enter airports. One study found lack of availability of gates due to exclusive and preferential use lease provisions cost US domestic travelers US $3.8 billion annually in higher fares (Morrison and Winston, 2000). A move away from long-term exclusive-use gate leases to shorter-term non-exclusive use should result in greater access to gates and result in lower fares.

5.2. Perimeter rules

Perimeter rules refer to rules on the allowable distances that aircraft may fly non-stop to or from a particular airport. These rules are in effect at New York's LaGuardia airport (non-stop flights longer than 1500 miles prohibited in order to shift long-haul traffic to Kennedy Airport), Washington's Reagan National Airport (the limit is 1250 miles, and is designed to shift long-haul traffic to Dulles Airport), and Dallas Love Field (airlines using aircraft with more than 56 seats are prohibited from flying from Dallas Love Field to states other than Texas, Louisiana, Arkansas, Oklahoma, New Mexico, Mississippi, and Alabama, in order to shift long-haul traffic to Dallas-Ft Worth Airport). These restrictions have been estimated to cost travelers nearly US $1 billion per year in higher fares (Morrison and Winston, 1997). Eliminating these restrictions would lead to more competition and lower fares.

5.3. Dominated hub airports

Although the hub-and-spoke system of consolidating traffic allows passengers on low-density routes to receive greater frequency of service (via a hub) than would be possible with a point-to-point service, it results in some airports being dominated by one or two airlines. Indeed, the US General Accounting Office (1990b) found that fares in 1988–1989 at 15 dominated hub airports (where one carrier had 60% or more of passenger enplanements or two carriers combined had 85% or more of passenger enplanements) were 27% higher than at a control group of 38 unconcentrated airports. A subsequent study found, using the same definition of a concentrated hub airport, that in 1998 there were 12 concentrated hub airports, and fares were 23% higher than fares at a control group of airports (Morrison and Winston, 2000). However, that study also found that when the control group of airports did not include airports served by Southwest Airlines, the hub premium vanished. Thus, what is called a hub premium appears to be a premium that airlines charge in any market in which they do not have competition from Southwest Airlines or other low-fare carriers. Policies, such as access to gates, that encourage more competition at hub airports, particularly from low-fare carriers, would reduce or eliminate the premium that carriers are able to charge at hubs they dominate.

5.4. Congestion

Although the number of flights has increased by more than 80% from 1977 to 2002, the growth in airport and air traffic control capacity has not kept pace. As a result, in 2002 the average domestic airline flight took nearly 13 minutes longer than the same flight did in 1977. Apart from the growth in system capacity not keeping up with demand, the delay problem is exacerbated by the method used to assess aircraft landing fees at airports. Fees are based on aircraft weight, and are uniform throughout the day. For many years economists have advocated charging landing (and take-off) fees that vary with the amount of congestion at the airport. In particular, it has been estimated that if air carrier airports in the USA charged aircraft take-off and landing fees equal to the extra congestion cost created by the take-off or landing, economic welfare would increase by nearly US $4.0 billion in 1988.

6. The future

The US airline industry is in the midst of what many observers believe will be profound and lasting change. As indicated in Figure 2, the presence of low-fare carriers has increased significantly. This, combined with reductions in traffic due to the events of 9/11 and the sluggish economic recovery, has significantly reduced airline industry profits, especially for the so-called "legacy" carriers, i.e. those formerly regulated carriers that were in business at the time of deregulation in 1978. This stress has already put considerable pressure on airline labor in the USA to accept pay cuts, which is likely to continue. Industry consolidation is also likely to occur as carriers exit the industry, either through merger or bankruptcy/liquidation.

Table 3
Annual gains to travelers from airline deregulation in the USA (US $ billions, 1993 values)

Category	Gain in consumer welfare
Fares	12.4
Travel restrictions	–1.1
Frequency	10.3
Load factor	–0.6
Number of connections	–0.7
Mix of connections (online or interline)	0.9
Travel time	–2.8
Total	18.4

Source: Morrison and Winston (1995).

7. Summary and conclusions

Table 3 summarizes the results of a comprehensive study of the effects of airline deregulation through 1993 in the USA. After 15 years of airline deregulation, travelers were paying US $12.4 billion less in fares annually than they would have, had regulation continued. However, some of these fares (i.e. discount fares) were available only to travelers who were willing to be bound by travel restrictions (e.g. advance purchase and Saturday night stay requirements). It is estimated that these restrictions imposed a cost of US $1.1 billion annually on travelers. Travelers valued the increased frequency of service at US $10.3 billion, but more crowded flights (increased load factor) reduced passenger welfare by US $0.6 billion by making it more difficult to book a flight (after taking the increased frequency into account). The percentage of passengers utilizing connecting flights, which require a change of planes, has increased slightly, with an estimated cost to travelers of US $0.7 billion annually. However, the vast majority of those connections are now with the same carrier (online) whereas before only about one-half of connections were on the same airline (the rest being interline). The additional convenience of better coordinated flight schedules and more proximate gates is estimated to be worth US $0.9 billion annually to travelers. Due to increased congestion, average travel times by air have increased over the years, resulting in a cost to travelers of US $2.4 billion annually. Taking all these effects into account yields an estimate of the net effect of airline deregulation on traveler welfare of US $18.4 billion as of 1993. The increased presence of low-fare carriers in recent years suggests that these benefits will continue to grow.

Although the current situation in the industry is taking its toll on industry profits, analysis during another cyclical downturn revealed that the flexibility that airline management has under deregulation cushioned the profit decline relative to what might have happened were airlines still regulated. Arguably, the same could be true for this present downturn.

References

Levine, M.E. (1965) "Is regulation necessary? California air transportation and national regulatory policy," *Yale Law Journal*, 74:1416–1447.

Morrison, S.A. (2001) "Actual, adjacent, and potential competition: estimating the full effect of Southwest Airlines," *Journal of Transport Economics and Policy*, 35:239–256.

Morrison, S.A. (2002) "Airline service: the evolution of competition since deregulation," in: L.L. Duetsch, ed., *Industry studies*, 3rd edn. Armonk: Sharpe.

Morrison, S.A. and C. Winston (1995) *The evolution of the airline industry*. Washington, DC: Brookings Institution.

Morrison, S.A. and C. Winston (1997) "Foul regulatory weather grounds airline competition," *Wall Street Journal*, Dec. 3.

Morrison, S.A. and C. Winston (2000) "The remaining role of government policy in the airline industry," in: S. Peltzman and C. Winston, eds, *Deregulation of network industries: what's next?* Washington, DC: AEI-Brookings Joint Center for Regulatory Studies.

US General Accounting Office (1990a) *Airline competition: industry operating and marketing practices limit market entry*, GAO/RECD-90-147. Washington, DC: US General Accounting Office.

US General Accounting Office (1990b) *Airline competition: higher fares and reduced competition at concentrated airports,* GAO/RCED-90-102. Washington, DC: US General Accounting Office.

Chapter 24

REGULATORY ISSUES: THE ROLE OF INTERNATIONAL MARITIME INSTITUTIONS

WAYNE K. TALLEY
Old Dominion University, Norfolk, VA

1. Introduction

Laws to regulate international shipping are not enacted by an international legislative body nor are they enforced by an international court. Rather they are enacted and enforced by three different regulatory authorities: the classification society, the flag state, and the coastal state. The regulations may be safety and environmental, though the focus is on the former.

States have sought to standardize maritime law by convening conferences for the purpose of making maritime conventions – statements of objectives and regulations on particular issues that are acceptable to the majority of the states that have convened to discuss them. Three agencies of the United Nations (UN) have been active in the organization and drafting of maritime conventions: the International Maritime Organization (IMO), the International Labour Organization (ILO), and the UN Conference on Trade and Development (UNCTAD). Once a prescribed number of convening states have ratified a convention, each ratifying state is obligated to enact the convention into national law, thus standardizing the law among the ratifying states.

2. Classification societies

Classification societies provide for self-regulation (as opposed to state regulation) of the shipping industry. Classification societies, which are generally privately owned and non-profit-making, set ship-quality standards and inspect ships to ensure that they are seaworthy. They produce ship specification rules and supervise the design and construction of ships to see whether these rules are followed; and inspect ships to determine if they are in compliance with the classification societies' regulations (or rules), national laws, and international conventions. New and existing ships are classified. If a ship passes inspection,

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

the classification society issues a Classification Certificate; if the ship fails the inspection, a certificate is not issued, or an existing certificate is withdrawn. Classification societies number more than 50 worldwide, and some date back more than 200 years. Five of the largest classification societies (based upon the number of ships classed) are Lloyd's Register of Shipping, Nippon Kaiji Kyokei, the American Bureau of Shipping, Det Norske Veritas of Norway, and Bureau Veritas of France.

The initial clients of classification societies were marine underwriters who used information provided by classification societies to assess risks in insuring ships. By the late nineteenth century the role of classification societies began to change – assigning ratings to inspected ships and issuing Classification Certificates. Ship owners paid for these services. Rules were established to determine ship safety. Classification societies determined whether ships were designed, built, and maintained according to their rules. Societies, however, did not establish rules for the operation and manning of ships. Large classification societies supplement their incomes by inspecting ships for governments. Since societies and governments regulate similar aspects of ships, there are overlaps in their safety regulation of ships.

A ship owner is not required to have his ships classed by a classification society. However, since marine underwriters must be confident that ships are seaworthy, they will only insure classed ships. A ship owner without a Classification Certificate cannot obtain the necessary trading certificates required by ports of call. Also, parties that charter ships require that these ships be classed; in addition, some governments require the classing of ships. The Classification Certificate is the industry standard for establishing that a ship is seaworthy.

By developing independently, classification societies do not necessarily have common rules. Technical standards for ship construction for one classification society may not be acceptable by another society. The costs incurred in satisfying the latter may exceed those of the former by millions of dollars. The International Association of Classification Societies (IACS) was established in 1968 to address this problem. The IACS is an association of 11 classification societies, including the ten largest. The purpose of the association is twofold: to develop uniformity among classification-society rules and to represent classification societies in their calibration with other shipping rule-setting organizations (Stopford, 1997). Over 160 sets of IACS Unified Requirements have been developed, e.g. the use of steel grades for various hull members, cargo containment on gas tankers, a minimum longitudinal strength standard, loading and guidance information, and fire protection of machinery spaces. The IACS has a governing council that is supported by a general policy group and 23 technical-specialty working parties. IACS members classify over 90% of the world's merchant-ship gross registered tonnage and more than 50% of the merchant ships afloat (Ozcayir, 2001).

Classification societies have no legal authority. Consequently, they compete with each other for ship owner clients, thereby raising doubts about the their safety enforcement performance – given the insoluble conflict of interest between themselves and ship owners that arises when hired by ship owners to class their ships. If a society requires a ship owner to incur expenses to class his ships, the ship owner's profits will decline, all else held constant. In a competitive environment for ship owner clients, and when ship owners themselves are facing stiff competition, societies are under pressure to reduce their safety demands, possibly classing non-seaworthy ships.

By the end of the 1970s, ship insurers in general were highly critical of the performance of classification societies. The UK's P&I Club, for example, noted: extreme variations recorded in the quality of services provided; difficulty in obtaining ship inspection reports given the contractual links between societies and ship owner clients; unwarranted extensions of classification of older ships; and safety rules that did not consider operational aspects of safety on board, e.g. crew quality and operating standards (Boisson, 1994). Protection and indemnity (P&I) clubs are ship owners' organizations that provide liability insurance for the same ship owners. The UK P&I Club, responding to its concern that classification societies could no longer provide an accurate evaluation of ship quality, established its own ship appraisal system. In 1990 it inspected more than 1500 ships, revealing serious safety failings, some aboard classed ships. Marine underwriters themselves have also been criticized for insuring substandard ships.

In addition to insurers, ship charters and bankers have also been critical of the performance of classification societies (Talley, 1999). For example, large oil companies have responded to this mistrust by utilizing in-house surveyors to inspect tankers for charter; some companies inspect more than 800 a year. The inspections are then used to determine which tankers are to be chartered.[a] Bankers who require reliable ship condition information for making ship investment decisions are promoting the creation of a general databank on ship conditions. The IACS has responded to the criticism of classification societies by using the IACS Code of Ethics and Quality System Certification Scheme (QSCS) to address ship standards; IACS members are bound by the QSCS service standards that they render.

3. Flag states

The legal jurisdiction over a ship is connected with its nationality. The nationality of a ship refers to the state that has authority over and responsibility for the ship.

[a]The EU has expressed concern for the charter of substandard ships, and is considering plans to fine cargo owners that knowingly charter substandard ships.

The symbol of a ship's nationality is the flag that it flies, and a flag state is the ship's state of nationality. Prior to the concept of nationality, regulation conflicts arose among states when the ship of a given state entered the territory of another state. International law has established that the seas and oceans between states should not be regarded as part of any state's territory, i.e. the freedom of the high seas. Consequently, on the free seas every state possesses authority, though not absolute, over its own ships. "Under freedom of the high seas, vessels belonging to all nations have unrestricted access to all parts of the sea that are not included in the territorial sea or internal waters of a state" (Ozcayir, 2001). This freedom applies to all states, even if land-locked. If a ship were not to have a nationality, it would not have protection under international law, and may be refused permission to enter foreign ports and engage in commercial activities. The state that has granted a ship the authority to sail under its flag has exclusive legislative and enforcement jurisdictions over it on the high seas. Flag states have the legal responsibility to ensure that their ships comply with state accepted standards of international law and conventions to which they are signatory.

When a ship registers in a particular state, the ship and its owner become subject to the laws of that flag state. Flag states enact the laws that govern the commercial activities of its ships. A ship may be registered in a given state if it meets the state's registration requirements. Once registered, the ship becomes an extension, while at sea, of the state's territory, and qualifies for its protection. Also, the ship owner becomes subject to the state's commercial laws, such as taxation and liability, safety regulations related to the construction and operation of ships, and crew regulations related to the selection and terms of employment. These factors for one flag state may provide a commercial advantage to a ship owner relative to the factors of another flag state. If so, the ship owner may be motivated to change the flag state to that of the former.

There are two groups of registers, national and open. National registers require ship owners to adhere to national legislation that covers the operation of a business in the country. Open (flag of convenience) registers offer commercially favorable terms of registration to ship owners for the purpose of earning revenue for the flag state. Open and most national registers are open to any ship owner. Hence, ship owners have many choices for a flag state; the relative advantages and disadvantages of each must be weighed. Prior to the 1950s, open registers were insignificant in number. Today, over one-half of the world shipping fleet is registered under open registers. The principal open-registry flag countries include Panama, Liberia, Cyprus, and the Bahamas. Open registers are predominantly used by foreign ship owners; conversely, national registers are predominantly used by domestic ship owners.

Open registers generally do not tax the profit of ship owners, but do charge a subscription tax per net registered ship ton. Ship owners are not required to employ a national crew, but rather are given the freedom to recruit internationally. Since crewing costs can be 50% of ship-operating costs, ship owners have used open registers to employ lower-cost crews from less developed countries – lowering their crewing costs but also the maritime employment of developed countries. The enforcement of ship safety standards varies widely among open registers, from high enforcement to leaving safety entirely to the ship owner, which may explain the poor safety records of some open registers in comparison to the good safety records of most national registers.

4. Coastal states

Only the flag state has jurisdiction (with certain exceptions) over a ship on the high seas. The 1982 UN Convention on the Law of the Sea (UNCLOS) defines high seas as "all parts of the sea that are not included in the exclusive economic zone, in the territorial sea or in the internal waters of a state." The coastal state has jurisdiction (with certain exceptions) over a ship on waters other than the high seas. Coastal state jurisdiction (or lack thereof) over shipping has been defined in terms of distinct maritime water zones: the internal waters, the territorial sea, and the exclusive economic zone.

Internal waters include "ports, harbors, lakes, rivers, canals, and waters on the landward side of the baselines from which the breadth of the territorial sea is measured" (Ozcayir, 2001). The coastal state has sovereignty over internal waters, i.e. it is entitled to enforce its laws against all ships and those on board in these waters. Sovereignty exceptions include, for example, those that relate to foreign warships and trying persons for crimes on board a ship. Coastal state authorities cannot board a foreign warship without the consent of its captain or other authority of the flag state. The flag state may also try persons for crimes committed on board a ship. Foreign ships do not have the right of passage, innocent or otherwise, in internal waters except in the case of an international treaty or agreement. However, access granted to foreign ships may be restricted, e.g. states may prescribe conditions for access to their ports, ports that are open to international trade, and ports that are closed to international trade.

The territorial sea is the area of water adjacent to internal waters. The maximum width for this water as recognized by the 1982 UNCLOS is 12 nautical miles, but in practice states use a range of 3 to 200 nautical miles, with the most common being 12 nautical miles. The coastal state has sovereignty over territorial seas subject to certain limitations. Unlike that for internal waters, a coastal state cannot restrict innocent passage of foreign ships through its territorial sea. Ships exercising the right of innocent passage must comply with all coastal state laws and regulations, except to the design, construction, equipment or manning of foreign ships, unless there are accepted international rules and standards. In its territorial sea, the coastal state has the discretion to prohibit non-innocent ship passage and adopt anti-pollution laws and regulations (subject to restrictions) directed at foreign ships.

The 1982 UNCLOS established the new ocean zone, the exclusive economic zone, defined as "an area beyond and adjacent to the territorial sea." An exclusive economic zone is an area of water extending up to 200 miles from a state's shoreline and is claimed (or established) at the discretion of each state. The coastal state has sovereign rights over natural resources (to explore, exploit, conserve and manage) within the zone, but not over shipping. Thus, as for high seas, there is freedom of navigation. Unlike that for high seas, the coastal state has the right to protect and preserve the marine environment within the exclusive economic zone, enforcing pollution regulations related to the dumping of waste, oil spills, and other forms of pollution from ships.

The flag state has the primary responsibility to ensure that ships comply with internationally accepted standards. However, deficiencies in ship-standard enforcement by flag states, in particular open registers, have prompted the use of port state control (PSC). The 1982 UNCLOS permits coastal states to legislate navigation safety, i.e. to provide for PSC to inspect ships entering a port and detain any for necessary repairs that do not meet international safety standards. In 1982, 12 European countries signed the Paris PSC Memorandum of Understanding, arranging to inspect safety and other certificates carried by ships of all flags (including each other's) visiting their ports, and to insist, by detention if necessary, on deficiencies being rectified. In 1995, member countries inspected 8834 ships, of which almost half had deficiencies; ships were detained in port when deficiencies were regarded as so serious that the ship or those on board were in danger, or where the marine environment could be threatened (Porter, 1996). In the UK, approximately 60% of inspected ships have some type of deficiency, and 6% have such serious defects to prevent, until rectified, their sailing (Goss, 1994). Although PSC systems are also found in Australia, Canada, the USA, and elsewhere, much of the world remains unaffected by PSC. PSC has broadened the responsibility for enforcing ship standards beyond the flag state. "The willing acceptance by the flag state of a mechanism (PSC) for the direct intervention of other sovereign states in matters affecting its registered ships has been a significant departure from the principle that those ships are an extension of its own territory" (Farthing and Brownrigg, 1997).

5. The International Maritime Organization

The IMO is an UN agency that is responsible for improving the safety of international shipping and preventing pollution from ships. In 1948 the UN convened in Geneva an international conference to consider establishing an international organization devoted exclusively to maritime matters. It was believed that such maritime matters as maritime safety would be more effectively addressed at the international level rather than by individual countries acting unilaterally. The conference lead to the establishment of the UN agency, the Inter-Governmental Maritime Consultative Organization (IMCO). In 1982 its name was change to the International Maritime Organization. The IMO has 158 member and two associate member states. In the last 30 years the IMO has promoted the adoption of 30 conventions and protocols and adopted over 700 recommendations and codes for maritime safety, the prevention of pollution, and related matters.

The IMO organizational structure consists of the Assembly, the Council, five committees (which undertake most of the work) – the Maritime Safety Committee (MCS), the Marine Environment Protection Committee (MEPC), the Legal Committee, the Technical Co-operation Committee, and the Facilitation Committee – and the Secretariat. The Assembly, which consists of all member states and meets once every 2 years, is the governing (or legislative) body of the IMO. It approves the work program and budget, makes decisions regarding international conferences and adopts resolutions that will be recommended to member states for their action. The Council is composed of 32 member states, elected by the Assembly for 2 year terms. In between Assembly sessions, the Council acts as the IMO governing body.

The MSC is the IMO's senior committee for technical work, and addresses a wide range of safety-at-sea issues. Its subcommittees are concerned with the construction and equipment of ships, aids to navigation, the prevention of collisions, handling of dangerous cargos, safety in ship manning, safety procedures and requirements, hydrographic information, casualty investigation, navigation records, salvage and rescue, and other factors directly affecting maritime safety. The MEPC coordinates IMO activities in the prevention and control of pollution from ships, especially those related to oil pollution. The Legal Committee is responsible for any legal matters that arise within the scope of the IMO. The Technical Co-operation Committee is responsible for coordinating the technical assistance of the IMO, e.g. helping governments implement adopted technical conventions. The Facilitation Committee is responsible for IMO activities that facilitate international maritime traffic, e.g. reducing the formalities and simplifying the documentation required of ships when entering and leaving ports. The Secretariat (appointed by the Council with the approval of the Assembly) consists of the Secretary-General and nearly 300 international civil servants based at the IMO headquarters in London.

IMO conventions normally develop with a committee or a subcommittee undertaking the initial work. A draft instrument is submitted to a conference consisting of invited delegations from all UN (IMO and non-IMO member) states. The final text adopted by the conference, i.e. the convention, is then submitted to governments for their ratification. The convention comes into force (or is ratified) when a specified number of countries have ratified it and fulfilled any other requirements. Implementation of the convention is mandatory on countries that have ratified it. The IMO not only gives technical assistance to ratifying flag states that lack the means to implement conventions but also provides for a process of internal and external audits of flag state implementations (De Bievre, 1997).

5.1. IMO safety conventions

In 1960 the IMO organized its first conference, adopting the International Convention on Safety of Life at Sea (SOLAS), which came into force in 1965. The 1960 SOLAS Convention was designed to improve the safety of shipping. Its provisions cover navigational safety, ship design and stability, electrical and machinery installations, fire protection, radio communications, life-saving appliances, and the transport of dangerous goods. The 1974 SOLAS Convention modified the 1960 Convention by including the amendment procedure whereby SOLAS can be updated for changes in the shipping environment without having to call a conference, i.e. amendments adopted by the MSC enter into force on a predetermined date, unless objected to by a specific number of states. The 1974 convention entered into force in 1980, with subsequent amendments in 1981, 1983, 1988, 1989, 1990, 1991, and 1992. The amendments address ship construction and fire safety, life-saving appliances, passenger ship safety, a global maritime distress and safety system (GMDSS), watertight doors, and ship stability.

The problem of dangerously overloading ships was addressed in the 1966 International Convention on Load Lines. The convention, which entered into force in 1968, established standard load lines for different types of ships under different conditions. In 1972 the Convention on the International Regulations for Preventing Collisions at Sea was adopted, and came into force in 1977, making ship traffic separation schemes adopted by the IMO mandatory. The 1972 International Convention on Safe Containers, in force in 1977, was designed to facilitate container trade by introducing uniform container size and strength requirements. The 1976 Convention on the International Maritime Satellite Organization, in force in 1979, promoted the use of space satellites to aid ship operation and its safety. The training of crews for the promotion of ship safety was addressed in the 1978 International Convention on Standards of Training, Certification and Watch Keeping for Seafarers. The convention, in force in 1984, established, for the first time, internationally acceptable minimum standards for crews. The 1979 International Convention on Maritime Search and Rescue, in force in 1985, was designed to improve existing procedures for carrying out search and rescue operations following accidents at sea.

5.2. IMO pollution conventions

In 1969 two IMO pollution conventions were adopted. The International Convention relating to Intervention on the High Seas in Cases of Oil Pollution Casualties, in force in 1975, established the right of coastal states to intervene in incidents on the high seas that are likely to result in oil pollution. The International Convention on Civil Liability for Oil Pollution Damage, in force in 1975, specified the civil liability of the owner of a ship or cargo for damage suffered as a result of an oil pollution incident. The convention ensures that compensation will be paid by the ship owner to the victims of pollution. However, some states felt that the liability limits were too low and the compensations, in some cases, were inadequate. Consequently, another IMO conference was convened in 1971 which adopted the Convention on the Establishment of an International Fund for Compensation for Oil Pollution Damage, in force in 1978. Unlike the civil liability convention, this convention spreads the burden of the compensation between ship and cargo owners. The fund is designed to provide compensation to victims of an oil pollution accident beyond that provided by the civil liability convention, and is operated by the International Oil Pollution Compensation Fund Organization, headquartered in London.

In 1973 the IMO convened a conference to discuss the broad problem of pollution from ships, resulting in the first comprehensive ship anti-pollution convention, the International Convention for the Prevention of Pollution from Ships (MARPOL), in force in 1983. The convention addresses not only oil pollution but that from chemicals, other harmful substances, garbage, and sewage. It greatly reduces the amount of oil that ships can discharge into the sea, and eliminates these discharges in certain areas. It requires ship inspections, the issuance of ship inspection certificates, the adherence of ships to specific anti-pollution rules, and the enforcement of these rules via inspection, and reporting of incidents involving harmful substances. Rules include the requirement that tankers have slop tanks and be fitted with oil discharge and monitoring equipment. Tankers that are new and over 70 000 dead-weight tons in size must have segregated ballast tanks of sufficient size to enable them to operate without using oil tanks for water ballast except in very severe weather conditions. In 1990 an IMO conference adopted the International Convention on Oil Pollution Preparedness, Response and Cooperation, in force in 1995, which enhances the ability of states to cope with sudden oil pollution emergencies such as those from tanker accidents.

Although the IMO is primarily concerned with safety and the prevention of pollution, it has also addressed other areas. In 1965 the IMO Convention on Facilitation of International Maritime Traffic was adopted, in force in 1967. The convention was in response to delays in maritime traffic from the lack of internationally standardized documentation procedures. It established uniformity in formalities and procedures related to ship arrival, stay, and departure at ports.

The IMO Convention for the Suppression of Unlawful Acts Against the Safety of Maritime Navigation was adopted in 1988, in force in 1992. It addressed measures for dealing with terrorist attacks on commercial shipping. In 1989 the IMO International Convention on Salvage was adopted, not yet in force. A salvor who prevents pollution damage but does not save the ship can expect no compensation.

5.3. *IMO recommendations and codes*

In addition to promoting the adoption of conventions, the IMO Assembly adopts recommendations and codes for practices that are not generally suitable for regulation by formal treaty instruments. Although the recommendations and codes are not usually binding on governments, many governments adopt them by their incorporation in national legislation and regulations. Codes become mandatory when they are part of adopted and ratified conventions.

Recommendations and codes often provide guidance for implementing provisions of conventions and clarify questions that arise in their interpretation. Also, they clarify questions related to specific maritime measures and attempt to ensure their uniform interpretation and application by governments. IMO codes adopted over the years include: the Code of Safe Practice for Solid Bulk Cargoes (1965); the Code for the Construction and Equipment of Ships Carrying Dangerous Chemicals in Bulk (1971); the Code of Safety for Fishermen and Fishing Vessels (1974); the Code for the Construction and Equipment of Ships Carrying Liquefied Gases in Bulk (1975); the Code for the Construction and Equipment of Mobile Offshore Drilling Units (1979); the International Gas Carrier Code (1983); the Code of Safety for Diving Systems (1983); and the International Code for the Safe Carriage of Grain in Bulk (1991).

An example of an IMO mandatory code is the International Safety Management (ISM) Code. It was adopted by the IMO in 1993, and brought into force at the 1994 SOLAS Conference, when its provisions became mandatory under a new Chapter IX to the SOLAS Convention. The ISM Code recognizes that environmental pollution and loss of life at sea are influenced by how companies manage their ships. The code seeks to insure safe practices in ship operation by improving ship management, and requires shipping companies to develop, implement, and maintain a safety management system (SMS) that includes:

- a safety and environmental protection policy;
- procedures and instructions to ensure safety and environmental protection;
- defined lines of communication and levels of authority between and among shore and shipboard personnel;
- procedures for reporting accidents;

- procedures for responding to and preparing for emergencies;
- procedures for management review and internal audits.

By July 1, 1998 all commercial passenger ships, oil tankers, chemical tankers, gas carriers, bulk carriers, and cargo high-speed craft of 500 gross tons or larger are to comply with the code. By July 1, 2002 all other commercial cargo ships and mobile offshore drilling units of 500 gross tons or larger are to comply with the Code (O'Neal, 1998).

6. The International Labour Organization

The ILO was created in 1919 under the Treaty of Versailles, to advance satisfactory working conditions and pay for workers. In 1946 it became an agency of the UN. The ILO consists of the General Conference, the Governing Body, and the Permanent Office. The General Conference normally meets once a year. Each of its 174 member states may send four delegates – two from government and one each from employer and worker organizations. The conferences may result in conventions and recommendations. The Governing Body meets three times a year; it is responsible for ILO programs, budgeting, and the work of the Permanent Office. The latter collects and distributes information on labor, and assembles materials for future conventions. Since 1920 the ILO has adopted 32 international conventions (which are mandatory for ratifying states) and 25 recommendations (which are not mandatory) dealing exclusively with maritime labor conditions. Their provisions are primarily concerned with working conditions on board ocean-going ships, addressing such issues as hours of work, minimum age, medical examinations, recruitment and placement, vacation, crew accommodation, pensions, sick pay, occupational safety and health, minimum wages, seafarer identity documents, social security, seafarer welfare in port and at sea, vocational training, certificates of competency, and continuity of employment. The 1936 ILO Officers' Competency Certificates Convention requires navigating officers, masters, engineering officers, and chief engineers to hold certificates of competency. In 1946 the Certification of Ships' Cooks Convention and the Certification of Able Seamen Convention were adopted. The 1946 Food and Catering (Ships' Crews) Convention established regulations to ensure that adequate catering facilities and sufficient supplies of food and water were available on board ship. The 1970 Prevention of Occupational Accidents to Seafarers Convention established minimum safety standards for the ship, its machinery and equipment, with procedures for inspection and enforcement of these standards. The 1976 Minimum Standards Convention requires ratifying states to have laws or regulations for registered ships that include safety standards ensuring safety of life on board ship that encompass competency, manning and hours-of-work standards, social security measures, and conditions of shipboard living arrangements and employment.

7. The UN Conference on Trade and Development

UNCTAD was established in 1964. During the same year, UNCTAD's Committee on Shipping was created, to develop maritime conventions on the economic and commercial aspects of ocean shipping. This committee has been particularly concerned with the maritime interests of developing countries, especially those related to liner conferences. Liner conferences are shipping line cartels that provide scheduled vessel service over specific trade routes and collectively discuss and set rates, usually only port-to-port rates. Liner conferences have immunity from anti-trust legislation from most OECD (Organization for Economic Cooperation and Development) countries.

During the 1960s, developing countries faced difficulties in establishing liner shipping operations for the transportation of their trade. Their specific criticisms of liner conferences included (Farthing and Brownrigg, 1997):

- freight rates were fixed unilaterally without regard to the problems of developing countries;
- participation by lines of developing countries was difficult;
- discrimination was practiced against developing countries;
- maritime law was developed by industrialized countries for their benefit.

The issue of liner conferences and developing countries was debated at the 1964, 1966, and 1972 UNCTAD Committee on Shipping conferences. In 1974 the Code of Conduct for Liner Conferences Convention was adopted. The code gives national shipping lines of the countries served by a liner conference the right to conference membership and the right to participate in trade volumes generated by mutual trade. Further, national shipping lines have the right of consent on all major liner conference service decisions affecting their countries. The code came into force in 1983. The code was the first international attempt to address closed liner conferences by opening liner conferences to new members (Stopford, 1997).

Today, shipping issues addressed by UNCTAD are much diminished. Within UNCTAD "there is a new mood of recognition of the importance of free-market principles, not just for the developed world, but also for the developing nations" (Farthing and Brownrigg, 1997). In 1992 the UNCTAD Committee on Shipping was suspended. As a consequence, shipping no longer has a separate identity within UNCTAD. The focus of the remaining shipping issues addressed by UNCTAD is technical cooperation and evaluation, e.g. identifying conditions to facilitate regional shipping cooperation and ways to promote shipping competition.

References

Boisson, P. (1994) "Classification societies and safety at sea," *Marine Policy*, 18:363–377.

De Bievre, A. (1997) "IMO brings regulation to its logical conclusion," in: D. Hughes, ed., *Lloyd's nautical year book 1998*. London: Lloyd's of London Press.

Farthing, B. and M. Brownrigg (1997) *Farthing on international shipping*. London: Lloyd's of London Press.

Goss, R. (1994) "Safety in sea transport," *Journal of Transport Economics and Policy*, 28:99–110.

O'Neal, W.A. (1998) "Implementation in the ISM era," in: D. Hughes, ed., *Lloyd's nautical year book 1999*. London: Lloyd's of London Press.

Ozcayir, Z.O. (2001) *Port state control*. London: Lloyd's of London Press.

Porter, J. (1996) "Majority of ships detained in Asia in '95 were bulkers," *Journal of Commerce*, Oct. 13:8B.

Stopford, M. (1997) *Maritime economics*. London Routledge.

Talley, W.K. (1999) "Determinants of ship accident seaworthiness," *International Journal of Maritime Economics*, 1:1–14.

This Page Intentionally Left Blank

Chapter 25

FOSTERING INLAND WATERWAYS

JACOB B. POLAK
Formerly Universities of Amsterdam and of Groningen

1 Introduction

The "fostering" of inland waterways, namely canals, lakes, and navigable rivers, as transport arteries is fairly well established in many countries. The justifications for this vary but are almost inevitably it is to meet some larger social or economic goal. This may involve such things as influencing industrial location or as an instrument to minimize the environmental impacts of transport. In this sense the fostering of inland waterborne transport can be seen as designed to meet intermediate goals that serve as an instrument for achieving a larger ulterior goal.

One common goal in many countries for fostering inland waterway transport is environmental protection, although this is less frequently vioced today. Waterborne transport has been seen as more environment-friendly than most other forms of freight transport such as railways and, in particular, road transport. Technical innovations have, however, in recent years placed the competitors to inland waterways transport virtually on the same environmental footing as inland waterways.

This chapter is not about inland waterborne transport in its very broad sense. It does not, for example, deal in detail with important themes such as performance, capacity utilization, costs, demand, characteristics of markets, and the externalities of inland waterways.[a] Instead the focus of the chapter is on regulatory issues and on alternative regulatory regimes. The aim is to survey what is known about the effects of alternative regulatory arrangements for inland waterways on economic efficiency.

[a]Information on the performance, cost structure, capacity utilization, and forecasting of inland waterborne transport with respect to the EU can be found at http://europa.eu.int/comm/transport/iw/en/site_map_en.htm#intro; for external costs and elasticities see Beuthe et al. (2002); and for details of market structures and the size of inland shipping firms, see the European Conference of Ministers of Transport (ECMT) (1999). For information on the US fleet and performance, see http://www.bts.gov/publications/national_transportation_statistics/2002/excel/table_water_transport_profile.xls, and on commodity groups see http://www.iwr.usace.army.mil/ndc/wcsc/pdf/preliminary02.pdf.

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

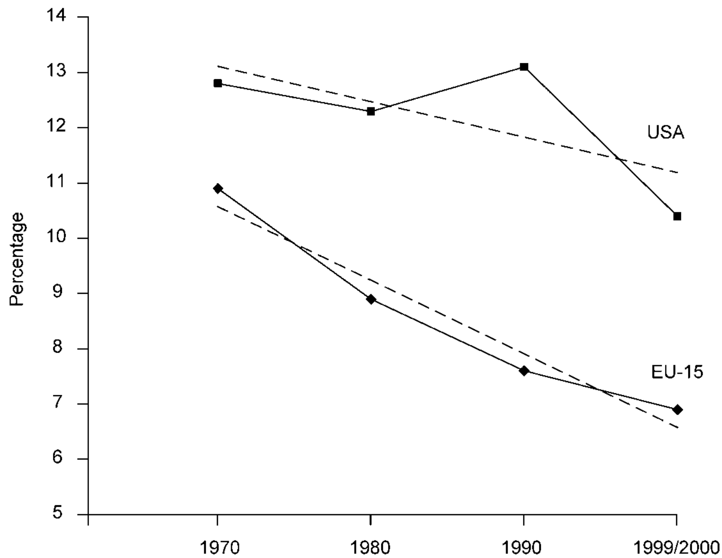


Figure 1. Relative modal shares (tonne-km) of inland waterway transport in the EU-15 and the USA, 1970 to 1999/2000. (Source: European Commission, 2003.)

2. Inland waterborne transport

Although inland waterborne transport is found in many parts of the world, it is only in two geographical areas that it plays a significant role; the USA. and a number of countries in western Europe. It is these areas that we focus on. Even in these markets, however, its share of total transportation as measured by tonne-kilometers has been fairly modest, only 12% in the USA and 9% in the EU. Furthermore, as shown in Figure 1, over the past 30 years the relative share of inland waterborne freight transport has been in steady decline in both regions.

There is at least one important reason, however, for suggesting that the future may not mirror recent past. The technological gap between inland waterborne freight transport and other forms of transport that has grown over past years is now beginning to rapidly close as improvements to vessels and enhanced information and communications technologies have emerged. One important example of this is the introduction of large containerships, such as the *Jowi*, with a capacity of 470 TEU, in the Netherlands.

It should not be forgotten, however, that the competitive modes to inland waterway transport – road, rail, and short sea shipping – are also seeing technological innovations. There are reasons to believe, however, that the relative past decline of inland waterborne freight transport has been halted and its role may become more important in the future.

Within this context, regulatory issues still remain important. Transportation supply and choices are driven as much by institutional and legal environments as they are by technology changes. This is as true with inland waterways as with other modes. Regulatory issues thus remain important and topical.

Although in the USA and western Europe inland waterway transport has largely been deregulated in the sense that many economic controls have been removed or considerably relaxed, this is not the case in many other countries. Even in the USA and in Europe there are still those that wish to reintroduce more stringent economic regulations or reform those that exist. There are also larger issues to consider given the reasons for interventions in the inland water transport market involving the links between transportation and overall economic efficiency. If the objectives of final policy change, for example the current interest in environmental protection increases, then this has implications for the various regulatory regimes applied to transportation.

Economic regulations are usually defined as all acts by governments that restrict parties involved in market activities from choosing those combinations of prices and quantities – their "decision space" – that they would prefer. Deregulation, to use the American term (liberalization is the more common British jargon), is the opposite of regulation and involves the removal or relaxation of restrictions on market practices. Regulation and deregulation are often discussed as if they are absolutes with nothing in between them. In reality, the situation has much more to do with the scale and nature of market intervention

The term "transport" is often used rather loosely but this is not always helpful when considering regulatory matters. Transportation can be approached from a number of ways and this applies to inland waterborne transport as to other modes.

From the production side there is the issue of what may be viewed in terms of horizontal domains. Essentially, should private transport ("transport on own account") play a role when considering regulatory issues, or should the discussion relate only to public modes ("transport for hire and reward")? In neoclassical economics, the combining of factors of production are considered a distinct economic activity if the result is intended for exchange in a market. This, by definition, does not apply to private transport where there is no public market. Because of this, transport on own account is not considered here. On a practical note, own account inland waterway transport has only ever attracted the attention of the regulatory authorities in a minor way.

Transport is also an industry that consists of two parts involving networks and activities that make use of the networks. This means that optimization of transport may be viewed as either the optimization of the entire structure or as the optimization of the two parts separately. Approaches differ and, for example, in the case of railways some countries adopt the integrated approach whilst others separate the regulation of track from the use of the track. In the case of roads and inland waterways, vertical separation has been dominant in the past. With roads, however, there has been some recent moves towards organizational integration (the building and subsequent exploitation of toll roads by private companies) but this trend has not yet manifested itself for inland waterborne transport. Whether consideration of regulatory issues regarding inland waterway transport should include the network supply part of the sector can be debated. From a purely theoretical perspective there may be reasons for doing so. The problems of network supply regulation are, however, different to those of most other activities (e.g. the granting of access to water networks). The focus here, therefore, is on the regulation of the use of waterways and little is said about the investment and maintenance of the waterway networks.

3. Regulation

Before looking at the motives of governments for regulating inland waterway transport, some of the relevant economic theory for regulatory action are considered. As highlighted by Hensher and Brewer (2001) one can distinguish between neoclassical and Austrian viewpoints. The neoclassical view contends that the existence of a competitive equilibrium in markets is a necessary requirement for allocative efficiency. In the Austrian view, on the other hand, dynamics hold a central place with particular emphasis on whether markets perform their function of inducing innovation.

In terms of the current regimes of regulation affecting inland water transport, during the Great Depression of the 1930s, many countries lost confidence in the self-regulatory powers of market processes and competition. In western Europe, as industrial production fell, a significant disequilibrium between supply and demand in inland waterborne transport emerged and the rates charged for the carriage of freight fell significantly causing severe financial difficulties for operators.

Governments in the main European inland waterborne transporting countries – Belgium, France, Germany, and the Netherlands – felt the situation of the industry and those involved in it would only be improved if there was direct public intervention. Consequently, governments introduced a rota ("*tour-de-rôle*") system for the principal element of inland waterborne freight transport, tramp shipping in domestic markets. The main aim of governments was that by introducing this system all operators would have equal shares in meeting transport demand and could thus be guaranteed a minimum income (Button, 2000). The rota system was seen as an emergency measure designed to meet the challenges of an immediate economic crisis. Given the length of the Depression plus the power of vested interests, however, it is not surprising that after a number of years, more extensive regulations were introduced. The main stated aim of this further regulation was to ensure continuity in the provision of inland waterway transport industry services for those with a demand for them.

In the USA, although some earlier regulation existed, it was the Transportation Act of 1940 that introduced major regulatory controls over inland waterway transport akin to those found in Europe. Pegrum (1973) argues that, US domestic shipping had, after the First World War, been subject to very severe competition and declining earnings. The US government was of the opinion that the "chaotic conditions in the water-carrier industry produce[d] results ... contrary to the public interest." In particular, it judged that the prevailing "instability of rates ... was harmful to the conduct of commerce and industry." Furthermore, it was found that the absence of effective regulation prevented coordination between all forms of transport, which at the time was considered a task for the government.

Two general conclusions may be drawn from these developments.

First, it appears that the motive for regulating inland waterborne freight transport was not only one of promoting supply – the case in western Europe – but also involved demand-side considerations – the case in the USA . Secondly, both considerations of allocative efficiency (the USA) and of equity (the rota system in western Europe) came into the thinking of governments when deciding to regulate inland waterway transport.

The underlying motives, be they positive concerns with efficiency or normative concerns with equity, point to desires by government to correct market failures (Baum and Schulz, 2000). Market failure implies that an inefficient functioning of the economy would rise if the allocation of scarce goods is left wholly to market processes. Such failures are associated with such things as indivisibilities, externalities, and imperfect information.

It may be added, however, that market failure need not be the only motive of governments regulating transport. For example, Teske et al. (1995) state, regarding the regulation of transport in the USA, that market interventions are "more likely, the choice of regulatory solutions was the result of political processes" implying the capture of the regulatory processes by vested interests.

The two main instruments – or, more precisely, groups of instruments – of economic regulatory policy that are commonly used are the setting of limits on the prices (rates) a firm may charge and controls over the number of firms and their size in an industry. A government may in principle use both price and quantity regulation at the same time, but need not do so. What is done varies considerably between countries; as Guasch and Spiller (1999) have noted "The regulatory functions and the regulated elements depend by and large on the country's regulatory philosophy, endowments, and characteristics; the sector's characteristics; and the extent of competition allowed."

In the particular context of inland waterborne transport, a very important element of a regulatory philosophy is the extent to which a government is willing to restrict the autonomy of the parties involved. This applies not only to the supply side but also to the demand side of the market, but in general the greater the importance of the autonomy of producers is to a government, the more likely it is to choose a more rigid instrument – normally regulating rates rather than the firms allowed in the market.

In the USA, there has been no quantity regulation of inland waterborne freight transport. In western European, governments realized that the rota system that had been introduced to regulate inland water transport had not dealt with the real cause of problems in the sector, namely the mismatch between supply and demand. The system, therefore, was supplemented with a quota system that applied to all transport. This system remained in use until the mid-1990s.

Rate controls are a common means of affecting quasi-market outcomes. There are four main ways in which the authorities may regulate rates charged by barge and other inland waterway transporters:

- Maximum rates will be set if a government feels it has to protect buyers of waterborne transport services against the abuse of monopoly powers that suppliers might possess – essentially "rate or price capping."
- Minimum rates can also be set to protect suppliers that, because of excess competition, may be forced to set prices at such a low level that many would not be able to survive – this kind of behavior is likely to occur in times of depression in the business cycle.
- Forked or bracket rates are a combination of maximum and minimum rates aimed to meet the objectives of maximum and minimum rates at peaks and trough of business cycles in markets where there is considerable temporal instability.
- Fixed rates, prescribes one single rate. A government might choose such an instrument in cases where it believes suppliers are unable to make proper cost calculations on which to base their prices. Such a belief is likely to arise in industries with small-scale suppliers, of which, inland waterborne transport is an example.

In a number of countries in western Europe, fixed rates existed for inland tramp shipping as an element of the rota system. In some cases (Germany and in Belgium, until 1969) fixed rates had been prescribed to supplement the rota, and in others (Belgium, from 1969; the Netherlands) minimum rates were introduced. Under the umbrella of the EU (then the European Economic Community) a common system of forked tariffs for all of inland transport modes was discussed for a period in the 1970s and 1980s but ultimately only applied to the international carriage of goods by road (degli Abbati, 1987).

Turning to the USA rate controls, the distinction between common carriers and contract carriers becomes relevant. A common carrier supplies its services to any potential customer; a contract carrier, in contrast, only serves one or perhaps a few shippers under specific contracts with it (Pegrum, 1973). The Transportation Act of 1940 empowered US public authorities to act in a number of different ways regarding rates. For common carriers, analogously to existing regulations governing road transport, any of the three types of rates – maximum, minimum, and fixed – could be prescribed for inland waterborne traffic. Although this could give the impression of an extensive regulation of rates, this was not the case (Teske et al., 1995; Fruin, 1997). There were so many exemptions that this led the responsible authority – the Interstate Commerce Commission – to remark that its jurisdiction was "limited to a small portion of the total transportation by water performed in the United States" (Locklin, 1972). Locklin estimated that only about 10% of all tonnage shipped by water was subject to regulation (he, unfortunately, does not distinguish between sea shipping and inland navigation). In addition, for those categories of inland waterborne transport to which no exemptions applied, the practical use of the regulations was very limited. Thus, few reasons were found to prescribe maximum rates, and little was done to set minimum rates for common carriers according to Pegrum (1973). For contract carriers, the authority of the government with regard to rates was restricted to the prescribing of minimum rates.

Overall, when considering formal regulation, the main conclusion is, for the countries considered, both quantity regulation and rate regulation have been used but in different ways. The contrast is especially marked for quantity regulation. This difference may be explained in terms of the prevailing views at the time when regulations were introduced of the extent to which government should be allowed to steer the economic processes

4. Effectiveness of regulating inland waterborne transport

The aims of regulating inland waterborne transport have been diverse and have involved the guaranteeing of a certain minimum rate of return to firms in the industry and continuity supply of inland waterway transport services. From a theoretical point of view, it is possible to distinguish between the direct and indirect effects of deregulation. Thus, Morrison and Winston (1999) state that the deregulation of transport also affects such matters as a firm's inventory policies, the scale of its operations, and the location of its businesses.

There are, however, few empirical studies specifically concerning the effectiveness of the regulation of inland waterway transport in meeting its objectives nor on the secondary consequences of regulation. What is available relates to the regulation of inland waterway transport in the USA and to the regulation of transport in general. This paucity of evidence makes it difficult to come to a solid conclusion.

Regarding US regulation, while the literature is sparce, it largely follows Pegrum's (1973) argument that "The experience of 20 years of regulation under the Act [the Transportation Act of 1940] has little to commend it. The legislation was incapable of providing a resolution of the difficulties that have beset domestic shipping"[a] and also that "Restriction of competition within the industry has had little effect on its behavior, and even if vigorous attempts were made to broaden and enforce the law, they would be unlikely to solve any of the problems that exist."

The more general literature concerning the regulation of transport in general is larger and is succinctly summarized as "The empirical analysis of regulatory policy regarding the transport sector shows that the objectives of the intervention regime were not achieved" (Baum and Schulz, 2000).

5. Deregulation

In the USA, in the 1970s and 1980s, the federal government proceeded to deregulate a number of sectors including many transport industries (Banister and Button, 1991). Other countries soon followed suit. Where deregulation of inland waterborne freight transport has taken place, it needs to be examined against the background of a general movement toward market liberalization.

The motives for liberalization stem to a large extent from the practical experiences with regulation (Quinet, 1998; Guasch and Spiller, 1999) and include:

- A greater awareness of the relatively high costs involved in the regulatory process.
- The conviction that economic regulation acts as an obstacle to technical progress (firms are not stimulated to develop new products or new production processes) and thus stands in the way of dynamic efficiency.
- The feeling that regulation led to consumers' needs not being being fully satisfied.

In contrast to the earlier belief that government intervention is required to correct for market failures, it became accepted that there may equally well be "government failure." From this, the conclusion was drawn that allocative deregulating markets and reforming the regulatory structures that remained could in many cases, better serve efficiency. Two particular developments in economic theory were particularly instrumental in causing changes in the way regulation was viewed: the concepts of contestable markets and of regulatory capture.

[a]The term "domestic shipping" in the USA does not distinguish between shipping on the Great Lakes (which is more akin to deep sea shipping) and coastal shipping, and that on rivers and canals.

The central idea of contestability theory is that the existence of competition is not a necessary condition for an efficient outcome in any given market as long as entry and exit to that market are easy and without cost (Baumol et al., 1982). From this, it follows that there is no need for governments to regulate markets, as long as they are contestable. Regulatory capture refers to a situation where regulators, instead of identifying with the general public interest serve the interests of a limited section of society, notably producers, whose behavior in the market they are supposed to supervise (Laffont and Tirole, 1991). These theories have greatly undermined the belief that the regulation of an industry is a requirement for its efficient operation.

Turning now to the deregulation of inland water transport in practice, developments in economic theory seem to have played more of an implicit than explicit role. When the EU decided to deregulate inland waterborne modes, it did not make reference to any theoretical considerations but just stated that making inland waterborne transport more competitive would lead to "the better use of the transport potential offered by inland waterway" (European Council, 1996). The crucial development was the decision that by the year 2000 all contracts would have to be freely concluded between the parties concerned and prices freely negotiated. Only in cases where problems specific to the inland waterborne market were likely to cause "a serious and potentially persistent excess of supply over demand" may measures be taken to prevent any further increase in transport capacity. The countries most concerned – Belgium, France, and the Netherlands – adapted their legislation to conform to the EU's decision (Germany had already abolished its fixed rates in 1994).

In the USA, the statutes allowing rates to be regulated were abolished in 1980 (Fruin, 1997).

While explicit studies of the effects of effects of regulating inland waterborne transport are limited in number, there has been some work on the impacts of deregulation. These analyses need also to be put into the much larger body of work that has been conducted across the deregulation of all modes of transport. In the USA, for example, deregulation of airlines, trucking, and railways has generally been found to benefit innovation, industry efficiency, and consumer welfare (Morrison and Winston, 1999).

Regarading western Europe, there are a number of studies on the effects of deregulation on inland water transport. Table 1 summarizes the findings. It should be noted that some of the performance indicators in the table are not independent of each other, e.g. cost savings will, in principle, have implications for the competitive condition of inland waterborne freight transport. It can be concluded from the table that deregulation in a number of western European countries has had a favorable effect on this mode of transport, with the single exception of the effect of the abolishing of the rota system on the utilization of vessels.

Table 1
Effects of deregulation on IWT

Performance indicator	Deregulating measure: Abolishing minimum rates	Abolishing the rota system	Deregulation	Tariff liberalization
Competitive position	+ ("improves," Dullaert et al., 1999)	+ ("substantially improves," Dullaert et al., 1999)	NA	NA
Cost savings	NA	+ ("30%," Beuthe et al., 2002)	NA	NA
Growth	NA	+	+ ("strong," Blauwens et al., 2003)	+ (ECMT, 1999)
Utilization of vessels	NA	– (ECMT, 1999)	NA	+ (ECMT, 1999)
Tariff level	NA	NA	NA	+ (ECMT, 1999)
Fraudulent practices	+ (ECMT, 1999)	NA	NA	+ (ECMT, 1999)
No. of small operators working as subcontractors	NA	NA	NA	+ (ECMT, 1999)

Key: NA, not applicable; +, positive effect; –, negative effect.
Note: Findings by Dullaert et al. (1999) refer to Belgium, France, Germany, and the Netherlands; those by Beuthe et al. (2002) and Blauwens et al. (2003) to Belgium. Findings by ECMT (1999) refer to Belgium, France, and the Netherlands, with the exception of the effect of tariff liberalization on the tariff level, which applies to Germany.

6. Conclusion

The primary aim here has been to examine the evidence on the impact of alternative regulatory arrangements for inland waterborne transport in industrialized countries on economic efficiency. Governments first regulated and later deregulated inland waterways transport for efficiency and distributional reasons using both market entry and rate controls. The regulation of market entry, however, seems by far to have been the most important instrument.

Empirical evidence on the effectiveness of policies of regulation and of deregulation of inland water transport is scarce. A cautious case can, however, be made for regulation being ineffective. Considering deregulation, Dullaert et al. (1999) have shown that, at least in certain cases, economies of scale exist in inland waterborne transport. This means that there are certain advantages to a policy of deregulating inland waterborne transport. On the other hand, a possible consequence of such a policy would be that those employed in small-scale firms in inland water transport, that, until now, have been numerous, would lose their jobs.

References

Baum, H. and W.H. Schulz (2000) "Transport policy," in: J.B. Polak and A. Heertje, eds, *Analytical transport economics, an international perspective*. Cheltenham: Elgar.

Baumol, W.J., J.C. Panzar and R.D. Willig (1982) *Constestable markets and the theory of industry structure*. New York: Harcourt Brace Jovanovich.

Banister, D. and K. Button (1991) "Introduction," in: D. Banister and K. Button, eds, *Transport in a free market economy*. Houndmills: Macmillan.

Beuthe, M., B. Jourquin and J. Charlier (2002) "Report on the competitiveness of intermodal freight transport networks in Europe," in: *European Conference of Ministers of Transport, Key issues for Transport Beyond 2000, 15th International Symposium on Theory and Practice in Transport Economics, Thessaloniki, Greece, 7–9 June 2000*. Paris: ECMT/OECD.

Blauwens, G., S. Janssens, B. Vernimmen and F. Witlox (2003) "A logistical analysis of transport frequency," in: W. Dullaert, B. Jourquin, T. Notteboom and J.B. Polak, eds, *Across the border: building upon a quarter century of transport research in the Benelux*. Antwerp: De Boeck.

Button, K.J. (2000) "Transport policy in the European Union," in: J.B. Polak and A. Heertje, eds, *Analytical transport economics, an international perspective*. Cheltenham: Elgar.

degli Abbati, C. (1987) *Transport and European integration*. Brussels/Luxemburg: Office for Official Publications of the European Communities.

Dullaert, W., H. Meersman, F. Moglia and E. Van de Voorde (1999) "Regulation and deregulation in inland navigation," in: H. Meersman, E. Van de Voorde and W. Winkelmans, eds, *Selected proceedings from the 8th World Conference on Transportation Research*, Vol. 1. Antwerp.

European Commission, Directorate-General for Energy and Transport (2003) *Energy and transport in figures*. Brussels: EC (http://europa.eu.int/comm/energy_transport/etif/lists/transport.html#Goods%20Transport).

European Conference of Ministers of Transport (1999) *What markets are there for transport by inland waterways?* Round Table 108. Paris: OECD.

European Council (1996) "Directive 96/75/EC of 19 November 1996 on the systems of chartering and pricing in national and international inland waterway transport in the Community," *Official Journal L*, 304:12–14.

Fruin, J. (1997) *Liberalisation in the transportation sector in North America*. Paris: OECD (http://www1.oecd.org/ech/pub/TRANSP2.PDF).

Guasch, J.L. and P. Spiller (1999) *Managing the regulatory process: design, concepts, issues, and the Latin American and Caribbean story*. Washington, DC: World Bank.

Hensher, D.A. and A.M. Brewer (2001) *Transport. An economics and management perspective*. Oxford: Oxford University Press.

Laffont, J.-J. and J. Tirole (1991) "The politics of government decision-making: a theory of regulatory capture," *Quarterly Journal of Economics*, 106:1089–1127.

Locklin, D.Ph. (1972) *Economics of transportation*, 7th edn. Homewood: Irwin.

Morrison, S.A. and C. Winston (1999) "Regulatory reform of US intercity transportation," in: J.A. Gómez-Ibáñez and C. Winston, eds, *Essays in transportation economics and policy*. Washington, DC: Brookings Institution.

Pegrum, D.F. (1973) *Transportation economics and public policy*, 3rd edn. Homewood: Irwin.

Quinet, E. (1998) *Principes d'économie des transports*. Paris: Economica.

Teske, P., S. Best and M. Minstrom (1995) *Deregulating freight transportation. Delivering the goods*. Washington, DC: AEI.

This Page Intentionally Left Blank

Chapter 26

MAKING PRICING WORK IN PUBLIC TRANSPORT PROVISION[a]

SERGIO R. JARA-DÍAZ
Universidad de Chile, Santiago

ANTONIO GSCHWENDER
Chilean Secretariat for Transport Planning (SECTRA), Santiago

1. Introduction

It is not only the operators of public transport who have to use resources to produce trips: the users themselves have a crucial input as well – their time. Users have to walk, wait, and travel inside the vehicle and, by doing so, they spend time, which is a resource consumed in the production of the trips. The microeconomic analysis of public transport must include this fact in order to find the optimal provision of services and the corresponding optimal prices. This is not new in transport analysis: car congestion pricing is the result of the inclusion of users' travel time in the pricing analysis. It is based upon the increasing relation between car users' travel time and patronage after a certain demand level due to a negative externality in the use of road space. As a consequence, the price that induces the best possible use of the road space is the difference between the value of the marginal time for all users and the value of the average travel time for each user. This chapter examines optimal public transport fares from the same viewpoint.

Section 2 discusses the cost structure of public transport. Then, optimal fares and their financial results are examined. As the existence of scale economies is shown to be crucial for the financial result of optimal fares, returns to scale in public transport are discussed in Section 4. Following that, a brief discussion about the impact of substitute modes and second best fares is presented. In Section 6 the relation between optimal fare and distance is analyzed. The chapter closes with a summary and discussion.

[a]The work in this chapter was partially funded by Fondecyt, Chile, Grant 1010687, and the Millenium Nucleus "Complex Engineering Systems."

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

2. Costs in public transport

2.1. General aspects

In the microeconomic analysis of public transport, two types of resources have to be taken into account: those provided by the operators, as vehicles, fuel, terminals, or labor, and those provided by the users, namely their time, usually divided into waiting, access, and in-vehicle time. The total cost (C_T) can be written as

$$C_T = C_{op} + C_U, \tag{1}$$

where C_{op} and C_U are the operators' and users' costs, respectively.

There are two elements that help identify the main variables for a cost analysis of the operation of public transport: route and fleet. The relevant route-related variables that affect either operators' or users' cost, or both, are the length, density, and number of stops. The corresponding fleet variables are the number of vehicles (or frequency) and vehicle size. In order to analyze the relation between costs and patronage (Y), it is necessary to establish which of these variables are fixed and which can be adapted to the demand level. In some cases, such as urban railways, the three route dimensions are fixed owing to physical reasons; for public transport, frequency and vehicle size are the operational attributes that are the easiest to adapt to demand. In the case of buses, institutional factors tend to define relatively stable routes, but length, density, and stops are easy to change in physical terms. Bus frequency is easy to adapt as well, as opposed to vehicle size, which can only be changed when buses are renewed. The flexibility of both the route and fleet decreases with the size (cost) of the type of vehicle (bus < tram < metro).

2.2. Cost structure

Given the elements discussed above, the relation between costs and patronage will now be examined. It is better to analyze the total cost structure in public transport, separating it into its two components: operators' cost and users' cost.

Considering the inputs supplied by the operators, there are operational and capital costs. The former include energy, crew, maintenance, and administration, and the latter are infrastructure and rolling stock costs. Engineering cost studies find that, in the absence of vehicle congestion, the average operators' cost decreases with demand (Meyer et al., 1965; Boyd et al., 1978; Allport, 1981).

Regarding the inputs supplied by the users, these costs are associated with travel times (waiting, access, and in-vehicle) and their money values. Each component has a clear relation to the different route and fleet elements. First note that frequency adaptation is the simplest input adjustment when the demand changes in the case of vehicles with fixed capacity. In scheduled transit systems, frequency increases with patronage (something that will be examined later), which will make in-vehicle time increase due to vehicle interactions, except in the case of rigid modes (e.g. metro) where a maximum feasible frequency exists. Nevertheless, not only vehicle congestion increases in-vehicle time, passenger boarding and alighting do as well. On the other hand, the waiting time decreases with frequency (demand), and, if the routes design is adaptable, demand expansions will induce a densification of the system, yielding reductions in access time as well.

In summary, when the demand increases, users' in-vehicle time also increases owing to vehicle congestion and passengers boarding/alighting, waiting time always diminishes, and access time decreases if routes can change. Despite an apparently mode-specific conclusion from the preceding discussion, this qualitative analysis actually yields a common scheme for the relation between users' cost and patronage. On one hand, the most flexible mode (bus) has a higher probability of congestion, but at the same time it is the easiest to densify when demand increases. On the other hand, the rigid rail-based modes have little (tram) or no (metro) congestion likelihood. So, in the first case a demand growth makes in-vehicle and access times vary with opposite signs, and in the second case both changes are low or nil. It is the waiting time, therefore, that prevails, generating a decreasing average users' cost function ($\mathrm{AC_U}$) in all cases, which means that $\partial \mathrm{AC_U}/\partial Y$ is negative and reflects the positive externality that an extra user generates on total travel time in scheduled transit systems.

The main conclusion is that the sum of the operators' and users' costs yields a total cost that increases less than proportionally with the demand, as found by Boyd et al. (1978) and Allport (1981) in their engineering cost studies. This means that the total average cost decreases with demand, which implies that there are scale economies. As a decreasing total average cost implies that the total marginal cost is lower than the total average cost, the result is exactly the opposite to the private car case. In the next section this will be shown to have a crucial impact on the optimal fares in public transport.

3. Optimal fares in public transport

3.1. The optimal fare

Given that both the operators' and users' costs have to be considered, the optimal fare for a public transport service can be found by maximizing the social benefit (SB), which is the difference between the total "willingness to pay" (in terms of both money and time) and the total cost (of both operators and users). Formally,

$$\underset{Y}{\mathrm{Max}}\,\mathrm{SB} = \int_0^Y \mathrm{GC}(u)\mathrm{d}u - [C_{\mathrm{op}}(Y) + C_{\mathrm{U}}(Y)], \tag{2}$$

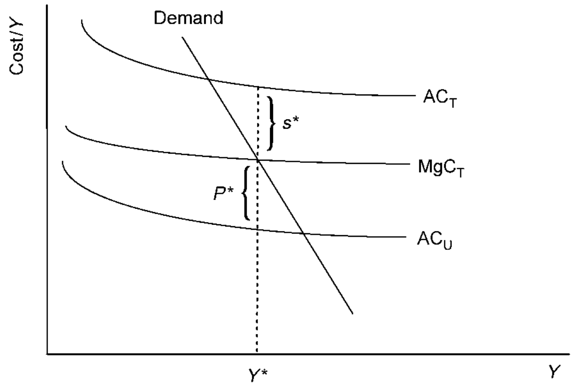


Figure 1. The optimal fare and subsidy in public transport.

where Y is the demand level, and the generalized cost (GC) is calculated as

$$\mathrm{GC} = P + \mathrm{AC_U}, \tag{3}$$

i.e. the fare (P) plus the users' average cost ($\mathrm{AC_U}$), which is the value of in-vehicle, waiting, and access times that the user experiences, given by C_U/Y. Note that GC in equation (2) is the inverse demand function.

The first-order condition for equation (2) is

$$\frac{\partial \mathrm{SB}}{\partial Y} = \mathrm{GC}^* - \mathrm{MgC_{op}}\,|_{Y^*} - \mathrm{MgC_U}\,|_{Y^*} = 0, \tag{4}$$

where $\mathrm{MgC_{op}}$ and $\mathrm{MgC_U}$ are the operators' and users' marginal costs, respectively. Using equation (3), and given that the total marginal cost ($\mathrm{MgC_T}$) is the sum of $\mathrm{MgC_{op}}$ and $\mathrm{MgC_U}$, the optimal fare is

$$P^* = \mathrm{MgC_T}|_{Y^*} - AC_U|_{Y^*}, \tag{5}$$

as shown in Figure 1. Just as in the case of the private transport optimal fare (congestion price), the user has to pay (in money) the difference between the total marginal cost and the users' average cost. By doing so, he perceives that the cost of his trip is the total marginal cost, as he already "paid" the users' average cost (the own time value).

By writing users' marginal cost as

$$\mathrm{MgC_U} = \frac{\mathrm{d}(Y\,\mathrm{AC_U})}{\mathrm{d}Y}, \tag{6}$$

the total marginal cost can be written as

$$\mathrm{MgC_T} = \mathrm{MgC_{op}} + \mathrm{AC_U} + Y\frac{\mathrm{dAC_U}}{\mathrm{d}Y}. \tag{7}$$

Replacing in equation (5) yields an alternative expression for the optimal fare (Jansson, 1979):

$$P^* = \mathrm{MgC}_{\mathrm{op}}\big|_{Y^*} + Y^* \left.\frac{\mathrm{dAC_U}}{\mathrm{d}Y}\right|_{Y^*}. \tag{8}$$

This expression shows that the passenger has to pay (in money) for the effect of his or her trip on the operators' cost ($\mathrm{MgC_{op}}$) and for the change that he or she produces in the users' average cost multiplied by the demand level, a negative figure that represents the total positive externality on the rest of the users, as explained above.

3.2. *The financial result in the presence of scale economies*

If scale economies exist, the total average cost is higher than the total marginal cost. This property yields a relevant financial result for optimal fares. Subtracting the operators' average cost from both sides of equation (5) yields

$$P^* - \mathrm{AC_{op}}\big|_{Y^*} = \mathrm{MgC_T}\big|_{Y^*} - \mathrm{AC_T}\big|_{Y^*}. \tag{9}$$

As deduced in the previous section, $\mathrm{AC_T}$ is larger than $\mathrm{MgC_T}$, which makes the operators' average cost larger than P^*. Therefore, the optimal fare cannot cover the operators' expenses, and an optimal subsidy per passenger (s^*) equal to the difference between $\mathrm{AC_{op}}$ and P^* is necessary. Expression (9) indicates that this optimal subsidy is equal to the difference between $\mathrm{AC_T}$ and $\mathrm{MgC_T}$ as well, as shown in Figure 1. Consequently, scale economies are a necessary and sufficient condition for optimal fares not being enough to cover operators' expenses. As said before, scale economies are equivalent to both decreasing total average cost and to $\mathrm{AC_T}$ being larger than $\mathrm{MgC_T}$.

4. Returns to scale in public transport

4.1. *A microeconomic model*

Due to the crucial impact of scale economies on the financial result of optimal fares, in this section returns to scale in public transport are explored in greater depth. A relevant issue for this discussion is the time period, i.e. which production factors are fixed and which are variable when fares are optimized. Jansson (1979, 1984) states that fare optimization does not make sense over a strictly defined short-term period where all inputs are fixed. In fact, the number of vehicles (fleet size) seems to be easier to change than fares. Therefore, it is reasonable to optimize fleet size first and then calculate optimal fares. Consequently, Jansson concludes that fares should be optimized in the "medium term," in which the number of vehicles is variable, implying a variable frequency as well.

The microeconomic public transport operation models available in the literature optimize different characteristics of the service, such as fleet (frequency), vehicle size, and routes spacing. If an expenditure function is considered that accounts for both operators' and users' inputs, the optimization of the variable factors yields a total cost function $C_T(Y)$, from which information regarding returns to scale can be obtained analytically. In what follows a microeconomic model is presented, in which frequency is optimized consistently with a medium-term definition that is adequate for public transport pricing. As known, deriving a cost function requires Y to be treated parametrically.

Following Jansson (1980, 1984), let us consider an isolated corridor served by one circular bus line L kilometers long, operating at a frequency f with a fleet of B vehicles. This service is used by a total of Y passengers per hour, homogeneously distributed along the corridor where each individual travels a distance l. If T denotes the time in motion of a vehicle within a cycle, and t is the boarding and alighting time per passenger, then the cycle time t_c is

$$t_c = T + t\frac{Y}{f}. \tag{10}$$

On the other hand, the frequency is given by the ratio between the fleet size and cycle time (B/t_c), which combined with equation (10) yields

$$B = fT + tY. \tag{11}$$

If c is the cost per bus-hour for the operator, and P_w and P_v are the values of the waiting and travel time, respectively, then the total value of the resources consumed (VRC) per hour is

$$\text{VRC} = Bc + P_w \frac{Y}{2f} + P_v \frac{l}{L} t_c Y, \tag{12}$$

where the first term on the right-hand side of equation (12) corresponds to the operator expenses, and the second and third terms are the users' waiting and travel time values, respectively. Note that the access time is not included in VRC because the route design is not a variable and, therefore, access cost is a constant that is not relevant to the optimization of the service.

Using equations (10) and (11), we can write expression (12) as a function of B, i.e.

$$\text{VRC} = Bc + P_w \frac{T}{2(B - tY)} Y + P_v \frac{l}{L}\left(T + \frac{tTY}{B - tY}\right) Y. \tag{13}$$

This expression shows that, *ceteris paribus*, increasing the number of vehicles diminishes the users' costs but increases the operators' costs. The users' cost reduction occurs because increasing frequency diminishes both the waiting and in-vehicle travel times, the latter because fewer individuals board and alight per bus.

Minimizing VRC with respect to B yields the optimal fleet size B^*, given by

$$B^* = tY + \sqrt{\frac{TY}{c}\left(\frac{1}{2}P_{\mathrm{w}} + P_{\mathrm{v}}tY\frac{l}{L}\right)}, \tag{14}$$

which from equation (11) yields the optimal frequency,

$$f^* = \sqrt{\frac{Y}{cT}\left(\frac{1}{2}P_{\mathrm{w}} + P_{\mathrm{v}}tY\frac{l}{L}\right)}, \tag{15}$$

known as the "square root formula" (Jansson, 1980, 1984). According to this result, the optimal frequency increases proportionally to the square root of total demand if the second term in parentheses is negligible relative to the first, but it can vary proportionally to the demand if the opposite occurs. A similar "square root formula" was found for the first time by Mohring (1972), and similar results are found in other single-line microeconomic models, as reviewed by Jara-Diaz and Gschwender (2003a).

Finally, on substitution of the optimal fleet size from equation (14) into equation (13) the minimum of VRC is obtained, i.e. the total cost function C_{T}:

$$C_{\mathrm{T}} = ctY + 2\sqrt{cTY\left(\frac{P_{\mathrm{w}}}{2} + P_{\mathrm{v}}tY\frac{l}{L}\right)} + P_{\mathrm{v}}TY\frac{l}{L}. \tag{16}$$

It is interesting to note that the operators' cost corresponds to the first term plus half of the second (i.e. the square root), and that the users' cost includes the square root plus the third term. The total average cost is

$$AC_{\mathrm{T}} = ct + 2\sqrt{cT\left(\frac{P_{\mathrm{w}}}{2Y} + P_{\mathrm{v}}tY\frac{l}{L}\right)} + P_{\mathrm{v}}T\frac{l}{L}, \tag{17}$$

which decreases with the demand level (Y), implying that scale economies exist. This reduction is due to the inclusion of the users' cost in the microeconomic formulation. This single-line result has been extended to various simple network cases (Jara-Diaz and Gschwender, 2003b).

4.2. Other relevant aspects

The model presented above does not include some important aspects that could change the cost structure. In what follows, these issues are discussed in order to ascertain their effects on scale economies.

Evans and Morrison (1997) constructed a model incorporating accident risk and non-scheduled delay. Both are included as part of the users' cost, but they also affect the operators' cost, as this has to be increased in order to reduce both risk and delays. Evans and Morrison found that the inclusion of these variables slightly increased scale economies.

On the other hand, Kocur and Hendrickson (1982) and Chang and Schonfeld (1991) included the spatial dimension optimizing the distance between parallel routes. By doing so, both optimal vehicle frequency and optimal line density are proportional to the cube root of the demand. This implies that scale economies are due not only to waiting time, which decreases at a slower rate than in the single-line case, but to access time as well.

As stated by Kerin (1992), none of these models incorporate externalities such as congestion or pollution, which would, apparently, tend to decrease the level of scale economies. If congestion is analyzed in an isolated public transport corridor in the medium term (frequency adjustment), there are two forms in which it can emerge. On one hand, as frequency grows with demand, interactions between vehicles will occur, affecting their speed. On the other hand, as frequency reaches an upper limit, growing occupancy rates will increase waiting times because of passengers not being able to board the first vehicle that arrives. The former effect is more important in the case of flexible modes (buses), and the latter is more relevant in rail modes, where a maximal frequency cannot be exceeded because of operational and safety reasons. Therefore, in the isolated corridor analysis there could be some demand level after which congestion becomes important enough to make the average total cost increase, as in the private car case.

However, the isolated corridor analysis becomes limited for high demand levels, as a possible way to adapt a public transport service to high demands is to increase the route density. The cubic root rule found by Kocur and Hendrickson and by Chang and Schonfeld for the optimal frequency shows that when route density is optimized, the optimal frequency increases slower than in the isolated corridor analysis (square root rule), slowing down congestion effects. On the other hand, the relationship between public transport and the car has to be considered when dealing with externalities, as they are usually substitute modes. This means that an increase in the total number of motorized trips has to be distributed between public and private transport. In their single public transport line optimization model, Oldfield and Bly (1988) considered that an increase in public transport demand implies a reduction in car use. Therefore, the final effect of the growth in public transport demand can be a reduction in the congestion level, due to a lower number of cars circulating. A similar analysis can be done in the case of pollution, as normally the emissions per trip are less for public transport than for cars. For synthesis, both the density analysis and the car – the transit substitution effect – make the effect of congestion on scale economies less important.

5. The impact of substitute modes and second-best fares

So far, we have implicitly assumed that all inputs and substitute or complementary modes are optimally priced (at marginal cost). Nevertheless, the car is normally underpriced. If it is a substitute for public transport, this will imply a more intensive use of the car than under the optimal modal split (congestion pricing). To correct this distortion, it is possible to underprice public transport as well, i.e. increase the amount of the optimal subsidy. Including this effect, Mohring (1979) found a negative optimal fare for a bus system. However, Kraft and Domencich (1970) state that for a given subsidy amount it is possible to attract more car users to public transport by diminishing the access, waiting, and in-vehicle times than by reducing the fare.

Nevertheless, underpricing public transport in any form (reducing fares or passenger times) to deal with the underpriced car will have an impact on the use of other modes such as the bicycle and walking (Kerin, 1992). They are substitute modes, especially for short trips, and their use will be reduced in spite of the fact that they are efficient modes in terms of congestion and pollution. Therefore, trying to optimize prices for every motorized mode would appear to be a better pricing policy than underpricing all of them.

If for any reason the optimal subsidy cannot be implemented, the second best fares are those that maximize the social benefit subject to covering the operators' cost with the fare box. In this case, those markets with larger price elasticities should have their fares closer to the optimal ones. On the other hand, other markets with lower price elasticities should have larger increases in their fares. This Ramsey pricing rule (Ramsey, 1927) allows fare revenues to be increased with the minimum modal split shift and, therefore, with the lowest loss in social benefit.

6. The optimal fare and distance

A public transport system allows trips between many origin–destination pairs. Thus, trips with different lengths will usually occur. This is not considered in models such as the one presented in Section 4, where all passengers travel the same distance and, therefore, a single marginal total cost and a single average users' cost exists, yielding a single optimal fare. Now, if passengers travel different distances, the question of the effect of trip distance on fares arises.

Cervero (1981) argued that the optimal fare should be positively related to the trip distance, but his conclusion was obtained considering only the operators' cost. When the users' cost is included in the analysis as well, the optimal fare will depend on the impact that a marginal passenger produces on the operators' cost plus the change that he or she makes to the average users' cost multiplied by the total number of passengers, as shown in equation (8). Nevertheless, it must be pointed out that this single-product analysis is not adequate when trips with different distances are considered. Thus, equation (8) only gives partial insight, as a multiproduct view is necessary for a formal analysis (e.g. Kraus, 1991).

Analyzing a feeder route and considering both the operators' and users' costs, Mohring (1972) argued that the optimal fare should be inversely related to distance, i.e. exactly the opposite to Cervero's conclusion. In the feeder route, passengers are boarding all along the line, but all of them alight at the end of the route. Thus, each passenger that boards makes those users who boarded before wait inside the vehicle for the time that he or she needs to embark. The closest to the end of the service a passenger boards (shorter trips), the fuller the vehicle will be and, therefore, more passengers will have to wait. As this makes the second term on the right-hand side of equation (8) larger, the optimal fare should be larger as well.

In a later work, Turvey and Mohring (1975) emphasized that the distance of the trip is not the relevant issue. They identify two reasons related to the occupancy of the vehicle that makes the fare higher: the first is the "Mohring effect" mentioned above, and the second (Turvey and Mohring effect) is that the higher the occupancy of the vehicle when the passenger travels in it, the higher the probability of other users not being able to board because of a full vehicle, increasing their waiting times. The latter effect is somehow positively related with trip distance.

Kraus (1991) constructed a multiproduct model where the passenger flow on each origin–destination pair is a different product. Besides including the impact of a marginal passenger on the in-vehicle time of the other users because of boarding and alighting (the Mohring effect), Kraus was the first to take into account crowding inside the vehicle. This "Kraus effect" is considered through a variable value of in-vehicle time, which increases with the occupancy factor. During the time that a passenger is inside the vehicle, he or she makes other users travel more uncomfortable. As in the case of the "Turvey and Mohring effect," this is partly positively related with trip length.

In summary, it is not very clear how trip length affects the optimal fare. Both Kraus and Turvey and Mohring found that the optimal fare should be in part positively related and in part inversely related with trip distance. Regarding the users' cost, what seems to be important for the optimal fare is the occupancy of the vehicle both when the passenger boards/alights and during the time that he or she travels in it. The first effect is that the users that are inside the vehicle have to wait when a passenger boards or alights (the Mohring effect); the higher the occupancy of the vehicle, the more time is "lost" due to the boarding/alighting of the marginal passenger. The second effect is that the space occupied inside the vehicle by a passenger increases crowding and discomfort for the other users (the Kraus effect) and increases the probability of other users not being able to board the vehicle (the Turvey and Mohring effect). Again, the fuller the vehicle, the higher the impact of the marginal passenger on the other users.

As there is no clear relation between optimal fare and trip length for a given route, flat fares seem to be a good option. Moreover, a flat fare is easier and cheaper to implement. Nevertheless, if routes with different lengths exist, it may be reasonable to have higher flat fares on the longer routes because of the operators' costs.

7. Summary and discussion

Just as in the case of private transport, public transport pricing analysis should consider both the operators' and users' costs. The former includes operational and capital costs, and the latter encompasses all types of travel time. The time frame for optimal pricing is relevant only if fixed inputs are set at a level that is different to the optimal, which will normally be the case for those services that are in operation at the time of the analysis. In general, frequency is a variable that can be adjusted easily, and, therefore, optimal pricing should always consider this as a variable. Adapting route length and route density is easier for some modes (bus) than for others (urban trains and trams). Optimal frequencies increase less than proportionally with demand, following a square root rule if everything else is fixed, and a cubic root if lines density is a variable.

Operators' cost models show that average costs diminish with patronage in general, and the same happens with those components of users' costs related with waiting and access time. In-vehicle time tends to increase because of boarding/alighting and congestion. In general, average costs diminish for both operators and users as patronage increases (economies of scale). As the optimal fare corresponds to the difference between the total marginal cost and the average users' cost, the preceding technical property translates into optimal fares that fall short of the operators' cost, making a subsidy necessary. In the absence of congestion pricing, the car is sub-priced, which makes the optimal subsidy even larger under a social optimum scheme because of the substitute nature of the car–public transport relationship. Public transport congestion would tend to diminish that subsidy.

Intuition based only on the operators' costs suggests that optimal fares should increase with distance (trip length). However, the inclusion of the users' costs in the analysis shows that the relevant variable is the occupancy rate when the (marginal) passenger boards, travels, and alights, besides the effect on the operators' costs. These two effects have opposite signs that seem to make a case for flat fares (larger for longer routes, though). Note that in this analysis of the relation between optimal fares and distance the effect on urban space has not been taken into account. Fares that increase with trip length would push toward urban concentration.

Optimal subsidies in public transport as presented here are conceived to make prices equal to the total marginal cost for the optimal operation of the system, and should not be mistaken with a capital cost subsidy. These latter subsidies may produce some distortions in the evolution of public transport systems. For example, Armour (1980) and Frankena (1987) found that capital cost subsidies caused buses to be replaced earlier. Furthermore, this type of subsidy increases the probability of constructing capital-intensive projects (e.g. metro lines) in corridors where low demands do not justify them (Small, 1992). On the other hand, when the operation is subsidized, various authors (Bly et al., 1980; Anderson, 1983; Pucher and Markstedt, 1983) found that the operators' cost increases, because of higher salaries and productivity reductions.

References

Allport, R.J. (1981) "The costing of bus, light rail transit and metro public transport systems," *Traffic Engineering and Control*, 22:633–639.

Anderson, S.C. (1983) "The effect of government ownership and subsidy on performance: evidence from the bus transit industry," *Transportation Research A*, 17:191–200.

Armour, R.F. (1980) "An economic analysis of transit bus replacement," *Transit Journal*, 6:41–54.

Bly, P.H., F.V. Webster and S. Pounds (1980) "Effects of subsidies on urban public transport," *Transportation*, 9:311–331.

Boyd, J.H., N.J. Asher and E.S. Wetzler (1978) "Nontechnological innovation in urban transit: a comparison of some alternative," *Journal of Urban Economics*, 5:1–20.

Cervero, R. (1981) "Flat versus differentiated transit pricing: what's a fair fare?" *Transportation*, 10:211–232.

Chang, S.K. and P.M. Schonfeld (1991) "Multiple period optimization of bus transit systems," *Transportation Research B*, 25:453–478.

Evans, A.W. and A.D. Morrison (1997) "Incorporating accident risk and disruption in economic models of public transport," *Journal of Transport Economics and Policy*, 31:117–146.

Frankena, M.W. (1987) "Capital-biased subsidies, bureaucratic monitoring, and bus scrapping," *Journal of Urban Economics*, 21:180–193.

Jansson, J.O. (1979) "Marginal cost pricing of scheduled transport services," *Journal of Transport Economics and Policy*, 13:268–294.

Jansson, J.O. (1980) "A simple bus line model for optimisation of service frequency and bus size," *Journal of Transport Economics and Policy*, 14:53–80.

Jansson, J.O. (1984) *Transport system optimization and pricing*. New York: Wiley.

Jara-Diaz, S. and A. Gschwender (2003a) "Towards a general microeconomic model for the operation of public transport," *Transport Reviews*, 23:453–469.

Jara-Diaz, S. and A. Gschwender (2003b) "From the single line model to the spatial structure of transit services: corridors of direct?" *Journal of Transport Economics and Policy*, 37:259–275.

Kerin, P.D. (1992) "Efficient bus fares," *Transport Reviews*, 12:33–48.

Kocur, G. and C. Hendrickson (1982) "Design of local bus service with demand equilibration," *Transportation Science*, 16:149–170.

Kraft, G. and T. Domencich (1970) *Free transit*. Lexington: Lexington Books.

Kraus, M. (1991) "Discomfort externalities and marginal cost transit fares," *Journal of Urban Economics*, 29:249–259.

Meyer, J.R., J.F. Kain and M. Wohl (1965) *The urban transportation problem*. Cambridge: Harvard University Press.

Mohring, H. (1972) "Optimization and scale economies in urban bus transportation," *American Economic Review*, 62:591–604.

Mohring, H. (1979) "The benefits of reserved bus lanes, mass transit subsidies and marginal cost pricing in alleviating traffic congestion," in: P. Mieszkowski and M. Straszheim, eds, *Current issues in urban economics*. Baltimore: Johns Hopkins University Press.
Oldfield, R.H. and P.H. Bly (1988) "An analytic investigation of optimal bus size," *Transportation Research B*, 22:319–337.
Pucher, J. and A. Markstedt (1983) "Consequences of public ownership and subsidies for mass transit: evidence from case studies and regression analysis," *Transportation*, 11:323–345.
Ramsey, R.P. (1927) "A contribution to the theory of taxation," *Economic Journal*, 37:47–61.
Small, K.A. (1992) *Urban transportation economics*. Chur: Harwood.
Turvey, R and H. Mohring (1975) "Optimal bus fares," *Journal of Transport Economics and Policy*, 9:319–327.

This Page Intentionally Left Blank

Chapter 27

TRANSPORTATION MERGERS: THE CASE OF THE US RAILROADS

WILLIAM B. TYE and JOHN HORN
The Brattle Group, Cambridge, MA

1. Introduction

Mergers among transportation carriers may raise unique issues of competition and public policy. Indeed, because of a tradition of regulation, such mergers are often handled administratively outside of the traditional merger review process for other industries. For example, the US Surface Transportation Board (STB) is the sole government agency with the power to approve or deny a railroad merger in the USA. The US Department of Justice (DOJ) can only play an advisory role to the STB, and the board does not need to abide by the DOJ recommendations. Transportation mergers may be judged under a more permissive "public interest" standard rather than on whether there would be a tendency to have a substantial adverse impact on competition (Wilner, 1997). Indeed, in some circumstances, reductions of excess capacity and "ruinous competition" have been viewed as a benefit of transportation mergers, not a cost, as would be more likely in the case of other industries.

More recently, the transportation industries have been undergoing a transition around the world from state-owned monopoly or regulated common carriers to deregulated competition (see Chapter 3). These developments naturally raise questions as to whether transportation industries still deserve special treatment insofar as merger policy is concerned. The US railroad industry provides both an excellent case study of the issues that are likely to arise, as well as lessons that can generally be applied to other transportation industries around the world (Gori, 2002).

The consequences of the rash of mergers that has occurred among the large US rail carriers has precipitated a debate over rail merger policy. Since deregulation was formalized with the passage of the Staggers Act in 1980, the number of Class I (largest) railroads decreased from 36 to (effectively) six today in all of North America (Chapin and Schmidt, 1999). Implementation of the mergers, especially the Union Pacific–Southern Pacific (UP/SP) merger in 1997 (Kwoka and White,

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

1999) and the absorption of Conrail by CSX and Norfolk Southern in 1998, has caused severe service problems, fewer choices among many shippers, and the creation of more bottlenecks within the rail network .

Responding to the merger application of Burlington Northern Sante Fe and Canadian National (BNSF/CN), the STB (2000a,b,c) placed a moratorium on new mergers among large railroads in order to determine whether it needed new rail merger policies. If approved, the merger would have created the first true transcontinental railroad in North America. Many in the rail industry believed that a chain reaction would have resulted: Union Pacific and Canadian Pacific would merge in the West, and CSX and Norfolk Southern, the two remaining Eastern Class I railroads, would be merged with the remaining two Western carriers, thereby producing two transcontinental railroads in North America. The STB decided that its prior merger policies were not adequate to determine whether such a configuration of the industry was beneficial to both carriers and shippers.

BNSF and CN challenged the STB's moratorium in the DC Court, claiming that the board did not have the authority to deny hearing their merger proposal at that time. However, the court ruled that the board was within its rights to institute the moratorium. Subsequent to the ruling, BNSF and CN decided not to pursue the merger. The STB's proposed changes were nevertheless largely incorporated later into its revised merger guidelines applicable to future transactions (STB, 2001).

There appear to be three competing visions of what is wrong with the US rail industry that help put this dispute over rail merger policy in perspective:

- Past mergers have permitted the remaining carriers to reduce competition, capacity, and service, and have failed to achieve the promised efficiency gains. Indeed, the mergers have increasingly led to service "melt downs" and elimination of the competition that was the rationale for deregulation.
- The industry continues to suffer from "Balkanization" and an inability to coordinate interline and other services. More mergers are needed to increase the length of haul, rationalize the network, and achieve efficiency gains.
- Enhancing or maintaining competition would be counterproductive, because it would lead to lower prices and difficulty in making the investments needed for improved service.

These three visions are not necessarily mutually exclusive. However, the first is generally espoused by those who are most concerned about the consequences of past rail mergers and are against additional railroad mergers. The last two are views of those who believe that past mergers will eventually realize their promise and support continued rail mergers as the long-run answer to the industry's problems. In fact, the second view is similar to those espoused about the state of the rail industry at the beginning of deregulation (Harris and Keeler, 1981).

It is probably no surprise that shippers and railroads have differing views of the effects of deregulation and mergers on the industry. This chapter will describe the effects of the increase in mergers, and evaluate the claims of railroads that they will not be able to compete effectively unless they are allowed to merge into transcontinental networks.

2. Current state of the rail industry

2.1. *The effects of rail mergers in the USA*

The STB, in its decision to change the merger standards, recognized the service problems accompanying prior mergers, the downward trend in railroad stock market equities, the lack of sufficient infrastructure (not excess capacity) that is plaguing the industry, and the STB's need to focus on and protect competition in the railroad industry. The proposed BNSF/CN transaction, in an atmosphere of widely recognized negative effects of recent mergers in other industries, prompted the STB to initiate the proceedings. The STB's initial decision comments often on the need to protect competition in the industry for the benefit of all, something that could be threatened by a "business as usual" approach to future rail mergers.

The inability of the railroads to digest the last round of mergers was no secret. Initially, many concerns related to the UP/SP merger. Subsequently, Norfolk Southern temporarily cut service to some cities and delayed the implementation of the Conrail breakup in July 1998. CSX's chairman, John Snow, acknowledged that "our terminals are congested and clogged, and we aren't getting the cars out in a timely way." The STB criticized CSX for not achieving a "sustainable level" of service improvement since the takeover of Conrail. Part of the problem was CSX cutting prices and attracting large increases in traffic . Presumably, the industry is learning a lesson here that we will return to below.

The evidence supports the conclusion that railroads have cut capacity on their lines, while at the same time increasing the flow of traffic over the remaining routes. "Major railroads have cut their workforces by about 60% and track networks by around 35% since 1980, while freight traffic has grown more than 45%." The US General Accounting Office (1999) found that from 1990 to 1997, Class I railroad mileage was reduced by about 15%, while employment fell approximately 18%. It found that about 60% of shippers responding to the service survey felt that their service was worse in 1997 than it was in 1990. Many of these shippers believed that their service had deteriorated due to mergers in the rail industry. According to the railroads, some of "the reasons cited for service problems were increased rail traffic and industry downsizing, which have created capacity constraints in the rail system" (US General Accounting Office, 1999). The study also found that while rail rates have generally decreased since 1990, they are at the same levels or higher for some routes and commodities. Those route/commodity combinations that faced the least competition (either from other railroads or other methods of transportation) had the highest rates.

Several studies have also tried to determine whether the efficiency gains in the post-Staggers Act era were due to deregulation or the mergers that occurred. In general, these researchers have found that most of the benefits that railroads have enjoyed in the past two decades are due to deregulation and competition, and not to the mergers that have characterized the industry since the passage of the Staggers Act.

Vellturo et al. (1992) and Berndt et al. (1993) analyzed the impact of mergers and deregulation on railroad costs. They found that costs decreased in the period after deregulation, but that the vast majority of the benefits were due to deregulation, not mergers, among the carriers. In analyzing the costs–benefits of four specific mergers in the late 1970s/early 1980s, the authors found that there were some cost savings. However, the savings were approximately the same as for the control group of railroads that did not merge, implying that the competitive environment due to deregulation was the cause of the cost improvements, not the mergers themselves. Chapin and Schmidt (1999) also found that efficiency has improved for US Class I rail firms since deregulation, but that it has not been due to mergers in the industry. The merged firms are more efficient at track maintenance and repair, but no more efficient in their shipping operations than their non-merged counterparts. The results also indicate that nearly half of the merged firms in the rail industry are operating above their efficient size, implying that the firms should not merge further, unless they experience continued decreasing returns to scale. Bereskin (1996) estimated cost function shifts for the railroad industry over the 1978–1993 period. He found an average productivity growth of 2.11%, although most of the increase occurred in the 1981–1982 period, with productivity slowdowns after. This later period had most of the big consolidation mergers. Schmidt (1995) found that duopoly competition reduced the price/cost margin mark-ups from the monopolistic levels, and that mergers between these duopoly competitors could have anticompetitive effects. Park et al. (1999) found that the UP/SP and BN/SF mergers reduced the price of shipping Kansas wheat to Houston, mainly due to more direct routing options and cost reductions realized from adopting the practices of the lower-cost merger partner. But in the cases examined, the merged parties still faced intramodal competition in the market after the merger, and the study concluded that "competitive rail prices will occur if shippers have access to at least two railroads" (Park et al., 1999). MacDonald (1987, 1989) also investigated the impact of deregulation on rail rates for specific commodities shipped from specific regions.

2.2. Consequences of the Staggers Act

Numerous studies have attempted to determine the effects of the Staggers Act (and, thus, deregulation) on economic efficiency, rates and service, and competition in the railroad industry (Gómez-Ibañez et al., 1999; McFarland, 1989). Many of these have found relatively positive benefits of enhanced efficiency (e.g. Burton, 1993), although there is an increasing element of skepticism. Part of the skepticism lies in the fact that most of the excess capacity in the rail industry has been removed. The lack of sufficient capacity is more of a problem in the industry, so efficiencies from elimination of excess track capacity will not be available in the future. This does not imply that there are no efficiencies left to achieve, merely that they will require more effort and innovation on the part of railroads.

Some studies have examined whether rates in the rail industry are impacted by the level of competition in general, without specific reference to the Staggers Act (Oum et al., 1999). Winston et al. (1990) found that adding one competitor to a shipper's route decreased the rates that the shipper paid. The US General Accounting Office (1999) also noted that shippers that had competitive alternatives paid lower rates. Fitzsimmons and Knudsen (1991) found that railroad competition did increase initially after the passage of the Staggers Act in 1980. However, over time, railroads learned to avoid competition with each other. Bitzan (1994) found that rail rates had increased after deregulation, as well as railroad profits. Wilson (1997) found that productivity gains were higher immediately after regulatory reform, but had returned to "pre-Staggers levels" by 1989. He also found that costs for railroads were 40% lower in 1989 under deregulation than if the railroads had remained regulated. These authors' results seem to indicate that there was a large reaction to the new deregulated environment, but once that was achieved, the productivity and cost improvements slowed down considerably.

These studies of the effects of the Staggers Act on competition and rates in the rail industry reveal three cautionary warnings about the impact of deregulation on rates. First, the more recent evidence reveals a threat to the benefits of deregulation, as firms have consolidated their market power and experienced service problems. Second, the benefits of deregulation have not been equally shared among shippers. Those beholden to one railroad have typically experienced higher rates than shippers with competitive choices. Finally, if these and other rate studies rely on revenue per ton-mile without controlling for the characteristics of the shipments, then the resulting conclusions on price will be influenced by these other factors. In general, the studies cited here do control for the other factors, but as Banks and Associates and Fieldston Company (1998) demonstrated, any results that do not explicitly control for these factors should be viewed with a critical eye toward exactly what they are measuring.

2.3. "Three-to-two" issues

It is a generally accepted principle of economic theory (Weiss, 1989; Scherer and Ross, 1990) that the more competitors there are in an industry, the closer the price will approach the competitive level. If there is only one (unregulated) firm, the monopolistic price will prevail in the industry. If there are hundreds of competitors, then the price will be at the competitive level. As the industry moves from one to many firms, the price continues to decrease in most of the commonly applied models of oligopoly that assume firms are pricing in excess of marginal costs.

What is true throughout the spectrum is also true as one moves from three to two competitors. The Cournot model of oligopolistic competition is one of standard models of economics that deals with competition in concentrated markets. As Carlton and Perloff (1994) demonstrate, adding a third competitor to the Cournot model of duopoly results in lower prices and greater output.

One of the few empirical studies (Kwoka, 1979) to examine the impact of a third competitor on competition in US industry generally found that a stronger second competitor in a market increased industry margins, while an increase in the size of a third competitor reduced industry margins. This implies that a third competitor is more of a rival than the second competitor, and generates more competition than an industry with two dominant players. The results also imply that two competitors are more likely to be able to (tacitly or explicitly) collude to raise prices than three competitors. The existence of a third competitor increases the chances of a "maverick" emerging.

Turning to the rail industry specifically, Grimm (1985) also found that:

> transformations of markets with three firms, not equally sized, to two firms appear to produce the greatest harm to competition. For increases in concentration above this range (4500–6500 HHI), effects on prices are much smaller, suggesting that, once a certain concentration is achieved, further reduction from two firms to one (or further differences in relative sizes of the two existing firms) is less significant.

> The results show that the reduction of direct competition in relevant markets will, in general, have a moderate effect on prices. However, the magnitude of the effect depends crucially on the market environment, in particular: existing levels of concentration, degree of intermodal competition, and concentration in vertically linked markets.

Grimm and Plaistow (1999) found that the UP/SP merger, which had substantial parallel components, had large competitive harms.

Levin (1981) and MacDonald (1987, 1989) also discovered that increases in rail competition lead to lower rail rates. MacDonald (1989) found that "competition among railroads has a statistically significant, fairly strong effect on rates. More competitors, as measured [in the statistical analysis] are associated with lower rates." He also noted that "railroad mergers that increase concentration will lead to rate increases. The analysis shows a statistically significant effect of concentration on prices in an industry with high barriers to entry and large capital commitments" (MacDonald, 1987). Levin found that as the degree of competition increased after deregulation, the rates that the railroads could charge would decrease. The converse of this is that reducing the level of competition in the railroad industry leads to higher rates.

Regulatory policy toward competition and mergers in the rail industry historically has been based on two related assumptions:

- the rail industry is plagued with excess capacity and must price well in excess of marginal (i.e. variable or traffic-sensitive) costs to be revenue adequate;
- excessive competition motivated by excess capacity would ruinously drive prices to marginal cost and prevent revenue adequacy.

However, all the evidence in the USA today supports the exact opposite conclusion: the rail industry suffers from excess traffic relative to capacity; past concerns over excessive competition are therefore no longer compelling. In any event, a vast number of former carriers in the industry have been eliminated through mergers. Concerns that railroads would engage in ruinous price wars under competition that needs to be eliminated through mergers are not based on the reality of the industry today.

The evidence therefore supports the STB's decision to revise its merger standards to raise the hurdle for merger applicants. The empirical evidence demonstrates that the removal of crucial third competitors has far-reaching consequences for the competitiveness of the railroad industry. The pricing and service problems characterizing the current state of the railroad industry, with fewer larger competitors, will only be exacerbated by the elimination of another competitor.

Before the UP/SP merger and the Conrail break-up, the elimination of one competitor still meant that other segments of the rail network in the USA would face competition from more than one other railroad. However, if the industry consolidates into two transcontinental firms, there will no longer be any routes where the railroads face more than one competitor. This lack of competition can be expected to lead to higher rates and greater exercise of monopoly/duopoly power throughout the system. Only two railroads in a market is problematic enough. Two railroads throughout the entire network is even worse, since competition by more than two carriers will not be present in any market in the country.

Up to this point, we have focused on parallel mergers that involve overlapping service provided by the merger applicants (i.e. the potential of a merger to create or exacerbate horizontal market power). Rail mergers can also raise market power issues when the merger carriers connect end-to-end on certain routes ("vertical market power"), to which we now turn.

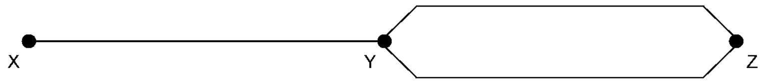


Figure 1. The "rat tail" problem.

2.4. *"One-lump" theory*

Figure 1 represents the issues that arise when a carrier controlling the "bottleneck" portion of a rail route proposes to merge with one of two competitors on a route where it interchanges traffic (the "rat tail" problem). In the diagram, the segment XY is a bottleneck, since only one carrier has tracks that lead between those two points. The portion of the route YZ is potentially competitive because two railroads have tracks that lead between these two points. "Vertical foreclosure" after a merger can occur if the bottleneck carrier proposes to merge with one of the two carriers serving the route YZ (a so-called "end-to-end" or "vertical" merger). The merged carrier either simply refuses to deal, or prices the access to its segment along XY (called a "revenue division") so high that the other carrier on the YZ segment cannot compete for traffic originating or terminating at X (i.e. it suffers a "price squeeze"). The unaffiliated connecting carrier, having lost a "friendly connection" at Y, would request regulators to impose "protective conditions" to prevent this foreclosure of competition. Under regulation, these requests were usually granted under the so-called "DT&I conditions."

The "Chicago school" of vertical merger analysis challenged the idea that a monopoly over a bottleneck portion of a network could create more monopoly power by vertically integrating and foreclosing a competitor in an upstream or downstream market (Bork, 1969). The idea was that any monopoly profits that could have been extracted (the "one-lump") would have been appropriated before the merger by the bottleneck firm in its price for that service. The one-lump theory (sometimes also referred to as the "neutrality" theory) asserts that there could be no negative effects of mergers with an upstream supplier, since there were no additional monopolistic rents to be earned. Merger policy at the STB (and its predecessor, the Interstate Commerce Commission) has espoused this view of vertical foreclosure as a rationale for refusing to impose "protective conditions" on end-to-end rail mergers in virtually every case since the Staggers Act deregulated much of the industry.

However, analysis of the potential for monopolistic behavior after a vertical or end-to-end merger has shown that this simple view of post-merger competition is not necessarily true. An emerging literature has shown that this "one-lump" theory is only a special case that requires a number of specific assumptions to be satisfied. This more recent analysis has focused on oligopolistic models of industrial organization to show that vertical mergers may in certain circumstances lead to increased monopoly power after the acquisition.

Scherer and Ross (1990) show that the necessary conditions are quite complex. Vertical integration by a bottleneck monopolist will result in an increase in monopoly power depending on at least four factors: "the elasticity of substitution between [the] monopolized input ... and other inputs" (the bottleneck portion and the other routes), "the extent to which [the bottleneck route] is important as an input into the downstream production process" (the importance of the route), "the elasticity of demand for the downstream product," and "the structure of the downstream market before integration" (number of competitors, type of competition, etc.). While not denying the possibility of the Chicago hypothesis, Scherer and Ross show that it is an empirical matter whether the necessary conditions will exist in the marketplace such that extension of monopoly power will not result from vertical integration.

Riordan and Salop (1995) also discuss the different conditions that must exist in the industry for vertical integration not to lead to decreases in competition. Some of the conditions that have to be examined include the availability of substitutes in both the input and output markets, competition in the upstream and downstream markets, elasticity of input demand, the possibility of price discrimination when selling the input, whether the acquired firm is a maverick or not, output market product differentiation, and whether the input costs are predominantly fixed or variable. While each industry has differing degrees of these various factors, the impact of a vertical acquisition needs to be determined empirically on a case-by-case basis.

Riordan (1998) shows that "backward vertical integration by a dominant firm into an upstream competitive industry causes both input and output prices to rise." Whether the cost advantage offsets "the negative effect of higher prices on social welfare" depends on the relative "input and output market shares" and "the degree of prior vertical integration."

The necessary conditions needed to validate the "one-lump" result can be rather restrictive. First of all, the theory is a static model of the industry, and does not allow for dynamic interactions between the firms. This is generally unrealistic in the context of the real economy. These dynamic considerations include investment decisions (whether the decision to incur fixed costs that the potentially foreclosed firm needs to make will be taken advantage of by the monopolist after they have been sunk), entry and exit of other competitors, and changing of rates following the introduction of competition, to name a few. Second, the theory assumes that there is perfect information about each firm's costs before the merger. Third, the theory assumes that there are no transaction costs involved in contracting for the traffic interchange. These transaction costs are partly included in the perfect information requirement, but also include the cost of contracting for all possible contingencies. These are also unrealistic characterizations of railroad firms in the economy. Indeed, many of the necessary conditions to validate the "one-lump" theory are denied by the applicants when making their claims of market imperfections (and thus efficiency gains) that can only be cured by the proposed merger (Tye, 1992).

These various studies conclude that one cannot say *a priori* that vertical integration definitely will not lead to increased monopoly power. These studies demonstrate that the particulars of each individual case need to be examined to determine how the specific industry structure and the status of the individual firms within that industry will be impacted by the vertical integration. In some cases it will result in no increase in market power, but in others it may. Empirical investigation is the only way to determine which outcome will result.

The key study of the effects of vertical integration in the railroad industry found that the "one-lump" theory was not applicable. Grimm et al. (1992), found that interline competitors did have a positive effect on competition in the rail markets served by a carrier with a bottleneck portion of the route. The authors also include a number of reasons why bottleneck carriers in the railroad industry do not extract monopoly rents prior to vertical foreclosure.

While it is by no means a necessity that vertical integration will increase monopoly power, the empirical question of where and when this applies to the railroad industry needs to be answered in future research. Uncritically accepting "one-lump" theories without examining evidence to determine whether the necessary conditions are satisfied forecloses the evidence that could show that monopoly power might increase through a vertical merger.

2.5. *Post-merger competitive access*

The arguments in support of the "one-lump" theory are very closely connected to those advanced in support of the efficient component-pricing rule (ECPR, also referred to as the "parity principle") and the theory that "voluntary negotiations" should govern traffic interchange among carriers (Baumol and Sidak, 1994; Baumol et al., 1997). These theories state that the ECPR produces the only efficient price of access to a bottleneck facility (see Chapter 3, and Kessides and Willig, 1998) and that this price will be adopted via "voluntary negotiations" among the carriers. The corollary is that granting access rights to other competitors over a bottleneck portion of a route or in any way interfering with the price of access set by the bottleneck carrier would lead to inefficient producers supplying the market. Hence, proponents argue that regulators should do nothing to prevent potential future market foreclosure in the case of end-to-end rail mergers.

Chapter 3 shows it is simply not true that voluntary negotiations leading to ECPR pricing are a necessary condition for economically efficient competitive access pricing (Tye, 1998). The alleged uniqueness of the ECPR as an efficient price is dependent on an implicit assumption in the proof (Tye, 1994; Tye and Lapuerta, 1996; Lapuerta and Tye, 1999). As long as price competition prevails, almost any method of pricing access leads to the most efficient firm providing the service over the competitive segment of a route containing a bottleneck. Furthermore, it is not the case that regulatory interference with voluntary negotiations in the post-merger competitive environment will lead to inefficiency. Indeed, constraints on post-merger route foreclosures may be necessary to preserve competition when end-to-end carriers merge (Tye, 1994).

2.6. Merger benefits

The STB was mainly concerned that overblown promises of substantial benefits from rail mergers had actually turned into adverse impacts on both carriers and shippers. The board therefore promised to heighten its scrutiny of such promised benefits during the merger review process and its vigilance in insuring that shippers were at least not harmed. The STB will undoubtedly no longer take projections of merger benefits by the applicants for granted.

Rail carriers initially proposed that regulators undertake what has become known as the "Williamsonian welfare trade-off" (Williamson, 1968). According to the trade-off, reductions in competition would be acceptable if they were more than offset by merger benefits. However, Tye (1991) points out that no trade-off is necessary. If the carriers agree to "protective conditions" to maintain the competition that would be otherwise eliminated, the merged carriers could still achieve the projected benefits.

Another tension between competitive effects and competition arises when the merger applicants make claims of merger benefits that conflict with the assertions about competitive consequences. For example, merger applicants often claimed that a merger between two end-to-end rail carriers would eliminate what has become known as the "double-marginalization" or "compounding of monopoly" problem (Spengler, 1950). According to the theory, a vertical merger between two rail carriers would actually cause the rate for the through movement to decline (Baumol, 1983). Such benefits are unlikely because rail carriers figured the problem out long before the economics profession and ordinarily charge a "joint rate" to the shipper for the through traffic. The revenues are then shared by a "divisions" agreement (Tye, 1984a,b).

Furthermore, the same carriers claiming benefits of lower prices on interline traffic also asserted that the antitrust concerns over the parallel aspects of the merger were unwarranted, because the merger candidates would still face significant competition because the markets were "contestable." If so, then there is no market power to compound and, thus, no benefits from eliminating the "compounding of monopoly" through merger. As another example, merger applicants have argued that one of the benefits of a proposed merger is to get rid of market imperfections arising from barriers to efficient contracting caused by the "gaming" behavior of railroads not under common ownership. These same applicants then deny that they would have any incentive to act similarly toward the other non-merging carriers opposing the transaction, who claimed that the merger would create new incentives for opportunistic behavior.

These contradictions are part of a more general problem (Tye, 1992). Merger benefits can only result from the elimination of a pre-existing market imperfection, which may be in conflict with the arguments by applicants that the market is highly competitive before the merger. Merger applicants will have to craft their efficiency claims much more carefully to ensure that they do not rely on market imperfections that are in turn assumed away in their analysis of competitive effects.

3. Conclusions

Since passage of the Staggers Act in 1980, the US rail industry has experienced a long series of mergers. While these mergers have had some beneficial aspects, they have also led to problems for some shippers, especially those who are captive to a bottleneck carrier. The future of the US rail industry is at a critical juncture, and the radical steps taken by the STB to revise its merger procedure guidelines demonstrate that it recognizes that special treatment of railroads is no longer justified, and that rail merger guidelines should more closely align with those from other industries (Pittman, 1990; US Department of Justice and Federal Trade Commission, 1992, 1997).

The STB recognized that the instability in the rail industry, the potential for further problems with new mergers among Class I railroads, and the "merger fatigue" that accompanied the previous round of mergers necessitated a reexamination of its merger policies. The service and financial market problems and the competitive concerns from the last round of mergers, along with the realization that the next round of mergers would be the final round and result in a permanent structure for the industry, led the STB to revise several areas of its merger policy. These include whether the merger is in the public interest, the effects on competition, the impacts on service, and the vertical effects of the mergers – both in other markets and for the future structure of the industry – and what safeguards need to be in place to ensure the benefits of the merger.

The US rail industry is stronger than it was in the 1970s, when many railroads were either bankrupt or on the verge of becoming so. Although there are some problems remaining, the application of traditional economic analysis for mergers should be sufficient to ensure that only mergers with a strong rationale will be consummated.

References

Banks and Associates and Fieldston Company (1998) *Rail freight rates in the post-Staggers era.* Washington, DC: Banks and Associates and Fieldston Company.

Baumol, W.J. (1983) "Some subtle pricing issues in railroad rate regulation," *International Journal of Transport Economics*, 10:341–355.

Baumol, W.J. and J.G. Sidak (1994) "The pricing of inputs sold to competitors," *Yale Journal on Regulation*, 11:171–202.

Baumol, W.J, J.A. Ordover and R.D. Willig (1997) "Parity pricing and its critics: a necessary condition for efficiency in the provision of bottleneck services to competitors," *Yale Journal on Regulation*, 14:145–163.

Bereskin, C.G. (1996) "Econometric estimation of the effects of deregulation on railway productivity growth," *Transportation Journal*, 35:40–41.

Berndt, E.R., A.F. Friedlaender, J.S.W. Chiang and C.A. Vellturo (1993) "Cost effects of mergers and deregulation in the US rail industry," *Journal of Productivity Analysis*, 4:127–144.

Bitzan, J. (1994) "Railroad deregulation: impacts on rates and profitability," *Upper Great Plains Transportation Institute staff paper.* Fargo: North Dakota state University.

Bork, R.H. (1969) "Vertical integration and competitive process," in: J.F. Weston and S. Peltzman, eds, *Public policy toward mergers*. Pacific Palisades: Goodyear.

Burton, M. (1993) "Railroad deregulation, carrier behavior and shipper response: a disaggregated analysis," *Journal of Regulation and Economics*, 5:417–434.

Carlton, D.W. and J.M. Perloff (1994) *Modern industrial organization*, 2nd edn. New York: Harper Collins.

Chapin, A. and S. Schmidt (1999) "Do mergers improve efficiency," *Journal of Transportation Economics and Policy*, 33:148.

Fitzsimmons, E. and J. Knudsen (1991) "Market share instability among Class I railroads and the impact of deregulation," *Quarterly Review of Economics and Business*, 31:66–75.

Gómez-Ibañez, J.A., W.B. Tye and C. Winston, eds (1999) *Essays in transportation economics and policy: a handbook in honor of John R. Meyer.* Washington, DC: Brookings Institution Press.

Gori, G. (2002) "Vote set today on Mexican rail merger: the No. 2 and No. 4 lines become one nationwide company," *New York Times*, May 16:W1, W7.

Grimm, C. M. (1985) "Horizontal competitive effects in railroad mergers," *Research in Transportation Economics*, 2:27–53.

Grimm, C.M. and J.J. Plaistow (1999) "Competitive effects of railroad mergers," *Journal of Transportation Research F*, 38:64–75.

Grimm, C.M., C. Winston and C.A. Evans (1992) "Foreclosure of railroad markets: a test of Chicago leverage theory," *Journal of Law and Economics*, 35:295–310.

Harris, R.G. and T.E. Keeler (1981) "Determinants of railroad profitability: an econometric study," in: K.D. Boyer and W.G. Shepherd, eds, *Economic regulation: essays in honor of James R. Nelson*. East Lansing: Michigan State University.

Kessides, I.N. and R.D. Willig (1998) "Restructuring regulation of the rail industry for the public interest," in: *Railways: structure, regulation and competition policy.* Paris: OECD.

Kwoka, J.E., Jr (1979) "The effect of market share distribution on industry performance," *Review of Economics and Statistics*, 61:101–109.

Kwoka, J.E. and L.J. White (1999) "Manifest destiny? The Union Pacific and Southern Pacific merger (1996)," in: J.E. Kwoka and L.J. White, eds, *The antitrust revolution: economics, competition, and policy.* New York: Oxford University Press.

Lapuerta, C. and W.B. Tye (1999) "Promoting effective competition through interconnection policy," *Telecommunications Policy*, 23:129–145.

Levin, R.C. (1981) "Railroad rates, profitability, and welfare under deregulation," *Bell Journal of Economics*, 12:1–26.

MacDonald, J.M. (1987) "Competition and rail rates for the shipment of corn, soybeans, and wheat," *RAND Journal of Economics*, 18:151–163.
MacDonald, J.M. (1989) "Railroad deregulation, innovation, and competition: effects of the Staggers Act on grain transportation," *Journal of Law and Economics*, 32:63–95.
McFarland, H. (1989) "The effects of United States railroad deregulation on shippers, labor and capital," *Journal of Regulation and Economics*, 1:259–270.
Oum, T.H., W.G. Waters II and C. Yu (1999) "A survey of productivity and efficiency measurement in rail transport," *Journal of Transportation Economics and Policy*, 33:9–42.
Park, J.J., M.W. Babcock and K. Lemke (1999) "The impact of railroad mergers on grain transportation markets: a Kansas case study *Transportation Research E*, 35:269–290.
Pittman, R.W. (1990) "Railroads and competition: the Santa Fe/Southern Pacific merger proposal," *Journal of Industrial Economics*, 39:25–46.
Riordan, M.H. (1998) "Anticompetitive vertical integration by a dominant firm," *American Economic Revue*, 88:1232–1248.
Riordan, M.H. and S.C. Salop (1995) "Evaluating vertical mergers: a post-Chicago approach," *Antitrust Law Journal*, 63:513–568.
Scherer, F.M. and D. Ross (1990) *Industrial market structure and economic performance*, 3rd edn. Boston: Houghton Mifflin.
Schmidt, S. (1995) *Oligopoly competition and market power in rail freight markets*. Working paper in progress. Schenectady: Department of Economics, Union College.
Spengler, J. (1950) "Vertical integration and antitrust policy," *Journal of Policy and Economics*, 58:347–354.
Surface Transportation Board (2000a) *Decision: public views on major rail consolidations*, Ex Parte No. 582, Mar. 17. Washington, DC: STB.
Surface Transportation Board (2000b) *Decision: major rail consolidation procedures*, Ex Parte No. 582 (Sub-No. 1), Mar. 31. Washington, DC: STB.
Surface Transportation Board (2000c) *Corrected decision: public views on major rail consolidations*, Ex Parte No. 582, Apr. 7. Washington, DC: STB.
Surface Transportation Board (2001) *Major rail consolidation procedures*, Ex Parte No. 582 (Sub-No. 1) Jun. 11. Washington, DC: STB.
Tye, W.B. (1984a) "Some subtle pricing issues in railroad rate regulation: comment," *International Journal of Transport Economics*, 11:207–216.
Tye, W.B. (1984b) "Some subtle pricing issues in railroad rate regulation: rejoinder," *International Journal of Transport Economics*, 11:219–220.
Tye, W.B. (1991) *The transition to deregulation*. New York: Quorum Books.
Tye, W.B. (1992) "Market imperfections, equity, and efficiency in antitrust," *Antitrust Bulletin*, 37:1–34.
Tye, W.B. (1994) "The pricing of inputs sold to competitors: a response," *Yale Journal on Regulation*, 11:203–224.
Tye, W.B. (1998) "Achieving competitive neutrality in rail competitive access disputes," *Journal of Transportation Law, Logistics and Policy*, 65:343–358.
Tye, W.B. and C. Lapuerta (1996) The economics of pricing network interconnection: theory and application to the market for telecommunications in New Zealand," *Yale Journal on Regulation*, 13:419–500.
US Department of Justice and Federal Trade Commission (1992) *Horizontal merger guidelines*. Washington, DC: Department of Justice and Federal Trade Commission.
US Department of Justice and Federal Trade Commission (1997) *Revision to the horizontal merger guidelines (on efficiencies)*. Washington, DC: Department of Justice and Federal Trade Commission.
US General Accounting Office (1999) *Railroad regulation: changes in railroad rates and service quality since 1990*, GAO/RCED-99-93. Washington, DC: GAO.
Vellturo, C.A., E.R. Berndt, A.F. Friedlaender, J.S.W. Chiang and M.H. Showalter (1992) "Deregulation, mergers, and cost savings in Class I US railroads, 1974–1986," *Journal of Economics and Management Strategy*, 1:367–368.
Weiss, L.W., ed. (1989) *Concentration and price*. Cambridge: MIT Press.
Williamson, O.E. (1968) "Economics as an antitrust defense: the welfare tradeoffs," *American Economic Review*, 58:18–36.

Wilner, F.N. (1997) *Railroad mergers: history, analysis, insight*. Omaha: Simmons-Boardman Books.
Wilson, W.W. (1997) "Cost savings and productivity in the railroad industry," *Journal of Regulation and Economics*, 11:21–44.
Winston, C., T.M. Corsi, C.M. Grimm and C.A. Evans (1990) *The economic effects of surface freight deregulation*. Washington, DC: Brookings Institution Press.

This Page Intentionally Left Blank

Chapter 28

PROTECTING THE CAPTIVE RAILROAD SHIPPER

KIMBERLY VACHAL and JOHN BITZAN
North Dakota State University, Fargo, ND

1. Introduction

The term "captive" is frequently used to describe shippers that have few transportation options. However, it has been used in a variety of contexts with varying meanings ranging from shippers having only one viable transportation option to shippers with a relatively inelastic demand for a particular transportation alternative. This chapter will attempt to define the captive shipper, describe the extent of shipper captivity in the USA, discuss the economic implications of captivity, and present some policy alternatives.

According to the *Oxford English Dictionary*, "captive" is defined as "having no freedom to choose an alternative." When defined in this strict sense, there are virtually no cases where the term "captive" would actually apply. A more commonly used definition of captivity refers to the number of transportation options available to the shipper and their influence on the rates and service levels realized by the shipper. For example, the closest reference to the captive shipper in US law is the definition of market dominance used in rail rate proceedings. Market dominance is defined in US law as "an absence of effective competition from other rail carriers or modes of transportation for the transportation to which a rate applies."[a] In this type of definition, captivity not only refers to a lack of transportation options available to the shipper, but also higher rates and poorer service resulting from the lack of transportation options.

In the USA, the most frequent application of the term "captive shipper" has been in the rail industry. A primary reason why this definition has been applied to the rail industry is that for many types of shipments – particularly long-distance, low-valued, bulky commodity shipments – rail has a major cost advantage. This advantage limits the abilities of other transport modes to serve as effective

[a]Section 49 USC 10707(a) of the US legal code, "Determination of market dominance in rail rate proceedings."

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

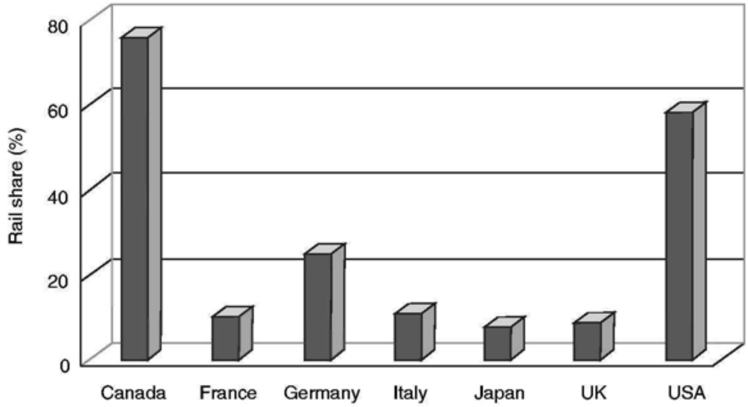


Figure 1. Rail share of domestic overland freight flows, tonne-kilometers. (Source: Cox, 2001.)

alternatives to rail, in many cases (the heavy dependence of US shippers on rail is highlighted in Figure 1). In fact, the first US transport industry regulated was the rail industry with the Interstate Commerce Act of 1887, and this regulation was pursued partially on the grounds of high rates to agricultural shippers with relatively few transportation options. The Interstate Commerce Commission (ICC) was enacted under the Act to control rail industry competition and stabilize rates in what was then viewed as a naturally monopolistic industry (Hoogenboom and Hoogenboom, 1976). Although this legislation was supported by both the rail industry and its customers as a means of stabilizing prices and controlling competition, much of the regulatory focus was on protecting the shipper with few transport options.

These price and competition controls were relaxed during the second half of the twentieth century. The financial problems of an overbuilt rail system and pressures from a more global marketplace encouraged policy-makers to substantially retract government interference in not only the rail industry but in the motor carrier, barge, and airline industries, as well. Deregulation in the USA has spurred tremendous gains in productivity, overall reductions in transport prices, widespread system rationalization, and a number of technological advancements, labor reforms, and industry consolidations. While the overall US experience with deregulation has been positive, shippers with few transportation options continue to appeal for increased government involvement in the transport system. Among the reasons such shippers continue to call for increased government involvement are:

- an increase in differential pricing resulting from deregulation, meaning higher prices charged to "captive" shippers relative to "competitive" shippers;
- an increase in national and regional transportation concentration resulting from rail mergers and consolidations, meaning increased levels of "captivity" and more transport firm pricing power;
- elimination of transportation options resulting from the rationalization of the US rail system due to an increased focus on long-distance, high density shipments.

As the deregulatory wave also moves across Canada and Europe – where rail systems have been partially or completely nationalized entities – policy-makers struggle with establishing the bounds of government involvement. In particular, such policy-makers struggle with the sometimes conflicting goals of promoting efficient rail systems and ensuring equity among shippers.

The importance of rail in the economies of nations, in terms of freight movement, varies substantially among nations (Figure 1). Past investments, comparative advantages, natural geographies, distances, industrial policies, institutional philosophies, and cultures have each contributed to current national and international freight transport networks. Cultural differences, high population densities, and smaller national land areas have contributed to a high dependence on roadways in the movement of freight in the EU and Japan. In 1970, rail was attributed with 30 and 40% of Japan and EU rail and road freight traffic, respectively (Cox, 2001). The rail freight market in these countries continues to decline in the scope of total overland domestic freight transport. Over three decades, rail freight shares in both regions have been significantly reduced. Rail is attributed with 16% of rail and road freight tonne-kilometers in the EU and 7% of the combined Japanese rail and road tonne-kilometers in recent statistics.

Although rail market shares in Canada and the USA have also exhibited declining trends, these markets are characterized by much greater dependence on rail in their domestic freight flows. This is expected given the larger land areas, large amounts of low-valued natural resource-based commodities, widespread geographic production of commodities, separation of population areas from production areas, and large rail investments of these nations. Rail is attributed with 76% of the overland freight traffic flows in Canada, and 58% in the USA (US Bureau of Transportation Statistics, 1999). While references made to captive shippers can be found in policy and industry discussions regarding Canadian and US rail industries, similar references have not been identified for other nations. This is not unexpected given the deregulatory wave in North America and the relatively prominent role of railroads in each nation's freight flows.

2. Identifying the captive shipper

Captive shippers are generally identified as a group that has relatively few alternatives in the transport services it utilizes to reposition products or services. Because of the lack of transportation alternatives, the captive shipper may not realize the benefits of attractive transport rates and service associated with transport competition. Competition in the transport market can originate from several sources, including intramodal, intermodal, geographic, and product competition. Each of these forms of competition may act to discipline railroads to provide better service at lower rates. Intramodal competition is generated from within a transport sector. To attain benefits associated with intramodal rail competition, a shipper or receiver has direct or nearby access to at least two railroads. The effectiveness of intramodal competition may depend on several factors including the scope of each railroad's network, the size of the shipper, and the ability of each railroad to serve the destination market in question. Intermodal competition refers to competition among modes. Intermodal competition may come in the form of modal alternatives, such as rail versus truck transport, or it may come in the form of modal combinations such as rail versus truck-barge movement. The extent of this type of competition depends heavily on product characteristics and their compatibility with various modal alternatives, and on the proximity of alternate modes to the route in question.

Geographic and product sources refer to competitive pressures associated with a shipper's or receiver's ability to substitute other commodities for that being shipped or to ship/receive the goods to/from alternate locations. Geographic competition can act as a limit on rail rates when a particular receiver (shipper) of a commodity can receive (ship) the same good from (to) several sources (destinations), and each of those sources (destinations) is served by a different railroad. In these situations, railroads are competing with each other across space, even though they do not directly serve the same markets. Similarly, product competition can act as a limit on rail rates when a particular receiver (shipper) of a commodity can substitute a different good for that being shipped, where that good is transported by a different railroad.

These four measures of competition have played an important role in US transportation regulatory policy, and continue to provide insight into the captivity of various shippers. According to US rail regulatory policy, the reasonableness of a rail rate cannot be challenged unless it can be established that the railroad has "market dominance" over the movement in question. Historically, there have been two hurdles towards proving "market dominance" in US rail rate cases: the rate must exceed 180% of variable costs,[a] and the rate must not be constrained by

[a]US legal code, Section 49 USC 10707(d)).

effective intramodal, intermodal, geographic, or product competition (Interstate Commerce Commission, 1981).[a] The logic behind using the market dominance test is that maximum rate guidelines are only necessary for shippers that have few transport options – that is, those shippers that are captive.

Because a lack of transport competition translates into a lack of transportation alternatives for the shipper, a numerical measure of captivity to a particular carrier or mode may be obtained by examining the price elasticity of demand for a particular carrier's or mode's service. Price elasticity of demand measures the percentage change in the quantity demanded of a particular service resulting from a 1% change in price. Shippers that have few transport options will have a more inelastic demand for rail service than those with many transport options – that is, the quantity they ship will be less sensitive to changes in price. As stated by the US Interstate Commerce Commission in its market dominance guidelines: "Effective competition from other carriers or modes of transportation, for the traffic to which a rate applies means that, if a carrier raises the rate for such traffic, then some or all of that traffic will be lost to other carriers or modes" (Interstate Commerce Commission, 1981).

Oum et al. (1992) provide a review of rail price elasticity estimates by previous authors (Table 1). As the elasticity estimates show, there is a wide variation in estimated price elasticities of demand among commodities. These may suggest different levels of captivity to rail among different commodities. However, as noted by the authors, these elasticities also vary widely depending on functional forms used. Moreover, to obtain an assessment of captivity of specific shippers, elasticity estimates would be needed at very disaggregate levels. With few exceptions, such elasticity estimates have not been made, and doing so would be a daunting task.

However, because the profit-maximizing, price-discriminating firm will price inversely to the market elasticity of demand, captivity might be measured using a more direct approach. As stated previously, a carrier that faces effective competition on a particular route cannot raise prices above a certain level otherwise they will lose the traffic to competing modes or firms. Thus, on routes where effective competition is realized, the price–cost margins should be low. Alternatively on routes that face little effective competition – captive routes – the price–cost margins would be expected to be high.

The US Surface Transportation Board (STB; the agency that replaced the Interstate Commerce Commission in 1995) estimates the revenue-to-variable cost ratio for every US rail shipment for every year. To the extent that the variable cost estimate is a reasonable proxy for marginal cost, we therefore have estimates of

[a]However, recently the market dominance guidelines were revised to exclude the consideration of geographic and product competition (Surface Transportation Board, 2001).

Table 1
Elasticity of demand estimates for rail freight (all are negative)

	Type of specification			
Commodity	Log linear	Aggregate logit	Translog	Discrete choice model
Aggregate commodities	1.52	0.25–0.35, 0.83, 0.34–1.06	0.09–0.29, 0.60	NA
Chemicals	NA	0.66	0.69	2.25
Fabricated metal products	NA	1.57	2.16	NA
Food products	0.02, 1.18	1.36	2.58, 1.04	NA
Iron and steel products	NA	NA	2.54, 1.20	0.02
Machinery	NA	0.16–1.73	2.27–3.50	0.61
Paper, plastic, and rubber products	0.67	0.87	1.85	0.17–1.09
Petroleum products	NA	NA	0.99	0.53
Stone, clay, and glass products	NA	0.69	1.68	0.82
Textiles	NA	2.03	NA	0.56
Transport equipment	NA	NA	0.92–1.08	2.68
Wood and wood products	0.05	0.76	1.97, 0.58	0.08

Source: Oum et al. (1992) – estimates are from 11 other studies by various authors.

price–cost markups for all US rail movements. While captivity is a relative term, revenues in excess of 180% of variable costs have traditionally been used in US rail regulation in determining the "captive/competitive" cutoff. As stated in Section 49 USC 10707 of the US legal code, the STB, in making a market dominance determination, "shall find that the rail carrier establishing the challenged rate does not have market dominance over the transportation to which the rate applies if such rail carrier proves that the rate charged results in a revenue-variable cost percentage for such transportation that is less than 180 percent." Using 1993 data, the STB estimated that 33% of US rail traffic moves at rates exceeding variable cost by more than 180% (STB, 1996). Based on this definition, we might infer that approximately one-third of US rail traffic is captive. However, not all of this traffic is subject to US rate regulation, as 28% of it moves under contract and 17% involves commodities that are exempt from regulation (STB, 1996).

Moreover, there is some evidence to suggest that this 33% measure may be increasing. Since 1992, the number of Class I freight railroads in the USA has declined from 13 to seven as a result of rail mergers. More strikingly, the total railroad industry revenue accounted for by the top four firms has increased from 57% in 1992 to 81% in 1999 (Association of American Railroads, 1993, 2000).

Evidence regarding changes in regional concentration in the USA is more limited, but some evidence at this level also suggests a possible increasing level of captivity. It is well known that recent mergers have reduced the number of rail competitors in grain-originating regions. Klindworth (1998) reports that in 1992, 58 crop-reporting districts in the top 20 grain-producing states were served by less than three railroads, while 87 crop-reporting districts in those same states were served by less than three in 1996.

Captivity in the Canadian rail system is also an important issue, as rail is the dominant mode in connecting spatially its dispersed economy. In its move to deregulate, the Canadian government established revenue caps for rail carriers and offers pricing freedoms within these revenue caps. A report published by Alberta Transportation in 2002 uses a benchmark similar to the STB's revenue-to-variable cost (RVC) ratio to assess the level of "captivity" of specific shipments. In this investigation, a benchmark level of variable costs plus a 20% contribution to fixed costs is used, based on levels established by the Western Grain Transportation Act – enacted in 1983 – as the contribution that would be expected under effective competition. Captivity for various commodities is assessed by comparing the total revenues for various commodities to the total estimated variable costs. They find that in Western Canada, commodities that seem to be most captive are sulfur, coal, fertilizers, and grains. Collectively these commodities represent 68% of Western Canada's originating rail volumes (Alberta Transportation, 2002). Moreover, Western Canada's rail volumes account for two-thirds of Canada's entire rail volumes. Although it is not possible to estimate the total portion of traffic that is captive from these estimates, it appears that the percentage of Canadian traffic that is captive could be substantially larger than that for the USA. Unfortunately, similar estimates of captivity are not available for other countries.

3. Implications of captivity for pricing and service

The cost structure of the rail industry offers little certainty for guaranteeing competitive and socially optimal rail rate structures. Socially optimal pricing requires price to equal marginal cost. However, in the railroad industry, there are substantial economies of scale. In order to provide service to a single shipper, a number of facilities are needed, including track, locomotive power, railcars, the right-of-way, etc. As more traffic is generated, the railroad needs less of these facilities per shipment, and is able to produce such traffic at lower costs. Because average costs are falling with increased traffic, marginal costs lie below average costs. This suggests that the social-welfare-optimizing rule of setting price equal to marginal cost would result in losses for the railroad firm.

Further complicating any attempts to set prices in a socially optimal way is the fact that many railroad costs are common or unattributable to specific shipments. Rail yards, rights of way, rails, and locomotives are among some of the costs that are necessary to make any shipment, but cannot be attributed to any particular shipment. As pointed out by Baumol et al. (1987), there is no non-arbitrary way to allocate such common costs.

As a result, regulation of the US rail industry has relied on second-best pricing scheme alternatives to the first-best socially optimal pricing scenario – in particular it has used constrained market pricing to regulate rail rates. Constrained market pricing consists of four constraints on rail rates, and is based on the principal of Ramsey second-best pricing. In its Coal Rate Guidelines, the Interstate Commerce Commission (1985) noted:

> "Ramsey pricing" is a widely recognized method of differential pricing, that is, pricing in accordance with demand. Under Ramsey pricing, each price or rate contains a mark-up above long-run marginal cost of the product or service to cover a portion of the unattributable costs. The unattributable costs are allocated among the purchasers or users in inverse relation to their demand elasticity. Thus, in a market where shippers are very sensitive to price changes (a highly elastic market), the mark-up would be smaller than in a market where shippers are less price sensitive. The sum of the markups equals the unattributable costs of an efficient producer.

Because of the cost structure of the rail industry and the use of Ramsey-type pricing, "captive" shippers are more likely to pay higher rates for rail service than "competitive" shippers. However, this does not suggest that such shippers are worse off because of the Ramsey type of pricing. Conversely, if railroads were forced to charge the same rail rates to all shippers, many "competitive" shipments would likely shift to other modes leaving a larger portion of common or unattributable costs to be borne by captive traffic – a point also made by Baumol et al. (1988) and others.

A number of academic studies in the USA have found evidence of a positive relationship between the level of "captivity" and railroad rates paid by shippers. A landmark study in this area was MacDonald's 1987 study of rail grain export grain rates. Specifically, MacDonald estimates the impact of intramodal and intermodal competition on rail rates for export corn, soybean, and wheat shipments. He finds that rates are lower in regions with more railroad-to-railroad competition and in regions that are closer to competing water transportation. Specifically, regarding railroad-to-railroad competition, he finds that a switch from a railroad monopoly in a crop reporting district to a duopoly (two railroads hauling equal amounts of grain from the crop reporting district) would lead to an 18% reduction in corn rates, on average. Further, a switch from two railroads originating equal amounts of grain to three railroads originating equal amounts of grain in a crop reporting district would lead to an 11% reduction in corn rates, on average. These findings provide strong empirical evidence for the suggestion that "captive" shippers experience higher transport prices than their "competitive" counterparts.

At the same time that "captive" shippers are likely to pay higher rail rates, they may also receive lower quality service. Although the evidence is very limited, due to a lack of quality of service data, there is reason to believe that shippers with few transportation options are likely to receive a lower quality of service because of a more inelastic demand for rail transportation with respect to service quality. Shippers that have a lot of transport options are likely to switch transport modes or carriers in the event of a service quality decline, while those with few options are likely to be less sensitive to such service quality changes. As a result, when transport capacity is limited, rail carriers have an incentive to serve shippers that are sensitive to service quality first, knowing that they will still be able to handle the "captive" traffic at a later time. In a survey of shippers in the Northern Plains of the USA, by the US General Accounting Office, shippers attributed rail service problems to a lack of competitive alternatives to rail transportation (US General Accounting Office, 1999).

While rates and service are likely to be affected by the number of transportation options available to the shipper, the extent of such differences may be affected by the regulatory scheme employed. For example, the bounds for differential pricing in Canada are somewhat different from those in the USA. Revenue caps serve as stronger limits on the amount of differential pricing in Canada as compared to the US differentials in Western grain movements, which are also governed by a set of distance based baseline rates established under the Western Grain Transportation Act. Although such rate caps may appear to limit the amount of differential pricing in a workable way, it is likely that differential pricing is limited as much by the relative homogeneity of Canadian traffic as it is by the regulatory scheme employed. That is, a large portion of Canadian traffic does not have nearby access to transportation alternatives such as barge, making the traffic more similar in terms of "captivity." A similar level of "captivity" among shippers suggests similar demand elasticities, which in turn suggests less differential pricing. Moreover, the US experience with regulation and deregulation has shown vast improvements of the current "market based" approach to pricing over the previous "cost based" approach that is more similar to the current Canadian system.

4. Institutional alternatives

The STB has set a benchmark that is commonly used in the US discussion of captive shipper issues: the RVC ratio exceeding 180%.[a] In addition to this quantitative requirement, the STB requires that qualitative analysis confirms market dominance through the absence of actual or potential competition for the traffic in question. If the market dominance requirement is met, the STB (2002) considers the reasonableness of the rail rate "while affording railroads the opportunity to cover all their costs and earn a reasonable profit," using demand-based pricing principles. In its interpretation of the governing statue, the STB provides that the law ensures rail customers access to rates that are not burdensome, considering the contribution to common costs of the rail carrier.

It has been suggested that the market will arrive at an efficient and equitable market solution if allowed to work freely. The USA has largely adopted this philosophy. One exception, however, is in the recognition of captive shippers. The constrained market pricing approach, adopted in 1985 by the Interstate Commerce Commission for determining rate reasonableness, consists of four constraints on rail rates:

- stand alone costs;
- checks on inefficient management;
- revenue adequacy;
- phasing of substantial rate increases.

When the railroad is not revenue-adequate, however, the primary constraint is the stand alone cost constraint. Stand alone cost is a theoretical concept that represents the cost that a competitor railroad operating the rail network comprised of that needed to make the shipment in question alone would realize. The stand alone cost concept is designed to act as a theoretical competitor would by limiting the rates charged by the railroad in question. It is used since the barriers to entry in the rail industry are high, and the actual entry of such a competitor is unlikely.

While the constrained market pricing approach has considerable theoretical appeal, its practical application is problematic for the small shipper. One of the major problems in applying the constrained market pricing approach to small shippers is the high costs associated with determining stand alone costs. The Interstate Commerce Commission recognized that the constrained market pricing procedures were not easy to apply or economically viable for shippers of sporadic or light volume traffic, and consequently opened a rule-making proceeding to determine separate maximum reasonable rate guidelines for non-coal traffic

[a]Note that exceeding the 180 ratio does not deem one a captive shipper, it merely establishes the initial criteria for consideration of a rail rate reasonableness by the STB.

(STB, 1996). The simplified rules for small shippers were adopted in 1986, but have yet to be used. In spite of this, complaints over high rates and poor service continue to be made by captive shippers in the USA.

Because of the continued complaints about rates and service in the USA by captive shippers, several policy proposals have been made in recent years. Policy proposals have varied in scope, and have included proposals such as:

- applying rate caps as in Canada;
- making further revisions to rail rate guidelines to ease rate complaints;
- revoking railroad antitrust immunity;
- introducing "open access" to rail lines where railroads pay for access and use of other railroads lines.

Each of these proposals will now be briefly addressed.

Theoretical and empirical evidence provide strong support for the need for differential pricing by US railroads. The large amount of common costs and the widely varying demand elasticities suggest that a viable and efficient rail system requires such pricing. Moreover, the USA experience has shown that the increase in differential pricing accompanying deregulation has enabled important innovations in the industry that have resulted in broad efficiency gains in the rail industry. Thus, applying rate caps is not likely to be beneficial for US rail transportation.

The simplified rate guidelines introduced by the STB were an important step in making rate complaints more accessible to small rail shippers. However, despite continued complaints over high rates and poor service, these new guidelines have not been used. This may suggest that the rate guidelines should be revisited.

The idea of revoking railroad antitrust immunity has been a popular idea among some US legislators in recent years. The argument suggests that the rail industry should not be treated differently from other US industries. However, the US railroad industry is characterized by a natural monopoly in localized markets (Bitzan, 1999). Such an industry should not be treated like industries not characterized by natural monopoly. In a natural monopoly rail industry it is desirable to have only one rail firm in local markets, and to prevent the abuse of such monopoly power through regulation.

Finally, the idea of "open access" has been proposed as a way to introduce competition to railroad networks. Proponents of the idea have suggested that such a system could result in substantial rate savings for shippers. However, previous research suggests that there are efficiencies associated with joint production of railroad rights of way and transport services over the rights of way, and that railroads are natural monopolies in providing transport services over their own network (Bitzan, 2003). These findings suggest that policies aimed at easing the rate complaint process would be preferred to those of introducing competition.

References

Alberta Transportation (2002) *Economic evidence of rail shipper captivity in Western Canada*. Edmonton: Alberta Transportation (http://www.tu.gov.ab.ca/Content/doctype60/production/ShipperCaptivity.pdf).

Association of American Railroads (1993) *Railroad facts, 1993*. Washington, DC: Association of American Railroads.

Association of American Railroads (2000) *Railroad facts, 2000*. Washington, DC: Association of American Railroads.

Baumol, W., M. Koehn, and R. Willig (1987) "How arbitrary is "arbitrary"? – Or toward the deserved demise of full cost allocation," *Public Utilities Fortnightly*, Sept. 3.

Bitzan, J.D. (1999) "The structure of railroad costs and the benefits/costs of mergers," *Research in Transportation Economics*, 5:1–52.

Bitzan, J.D. (2003) "Railroad costs and competition: the implications of introducing competition to railroad networks," *Journal of Transport Economics and Policy*, 37:201–225.

Cox, W. (2001) *Freight rail's potential to reduce congestion*. Austin: Texas Public Policy Foundation, (http://www.texaspolicy.com/pdf/2001-10-30-transportation-freight.pdf).

Hoogenboom, A. and O. Hoogenboom (1976) *A history of the ICC, from panacea to palliative*. New York: Norton.

Interstate Commerce Commission (1981) *Market dominance determinations and consideration of product competition*, Ex Parte 320 (Sub-No. 2). Washington, DC: ICC.

Interstate Commerce Commission (1985) *Coal rate guidelines nationwide*, Ex Parte 347 (Sub-No. 1). Washington, DC: ICC.

Klindworth, K. (1998) *Agricultural transportation challenges for the 21st century*. Washington, DC: US Department of Agriculture.

MacDonald, J. (1987) "Competition and rail rates for the shipment of corn, soybeans and wheat," *Rand Journal of Economics*, 18:151–163.

Oum, T., W. Waters II, and J. Yong (1992) "Concepts of price elasticities of transport demand and recent empirical estimates: an interpretive survey," *Journal of Transport Economics and Policy*, 26:139–154.

Surface Transportation Board (1996) *Rate guidelines – non-coal proceedings*, Ex Parte 347 (Sub-No. 2). Washington, DC: STB.

Surface Transportation Board (2001) *Market dominance determinations – product and geographic competition*, Ex Parte No. 627. Washington, DC: STB.

Surface Transportation Board (2002) *Testimony of Linda J. Morgan at a hearing of the Senate Committee on Commerce, Science, and Transportation on rail freight transportation in North Dakota, Field Hearing*. Washington, DC: STB.

US Bureau of Transportation Statistics (1999) *G-7 countries: transportation highlights*, Publication No. BTS99-01. Washington, DC: US Department of Transportation.

US General Accounting Office (1999) *Railroad regulation: changes in freight railroad rates from 1997 through 2000*, GAO-02-524. Washington, DC: GAO.

Part 5

EVALUATION FRAMEWORKS

This Page Intentionally Left Blank

Chapter 29

COST–BENEFIT ANALYSIS

KENNETH G. WILLIS
University of Newcastle upon Tyne

1. Introduction

Approval for funding any public transport or highway scheme normally requires that the benefits of the scheme exceed its costs. The appraisal of such projects usually entails undertaking a cost–benefit analysis (CBA), based upon the expected costs and benefits of the scheme. The purpose of a CBA is to inform decision-making; and also to maximize the welfare of society by selecting transport schemes where benefits are maximized relative to costs. In project appraisal, no method other than CBA has such a sound theoretical economic base and widespread application.

2. Financial versus social costs and benefits

Financial accounting records "total economic value" as revenue or sales from a transport operation; and total costs as the financial cost of acquiring resources to build and operate a transport project. However, financial values differ from total welfare. Individuals derive utility from intra-marginal journeys over and above the price they pay for the journey. In CBA this "consumer surplus" is counted as a benefit in addition to the price they are willing to pay for a journey. For financial costs, CBA substitutes the concept of opportunity cost, or the social value foregone when resources are moved away from another economic activity into the transport project. Resources are valued in terms of their opportunity cost: what is being lost elsewhere in the economy if these resources are used in a transport project. If the opportunity cost is less than the financial market price of the resource, the difference is termed an "economic rent." Economic rent is excluded as a cost in CBA.

In appraising a new road, for example, financial revenue will be zero if no tolls are charged, so that the financial costs of building and maintaining it will always exceed financial revenue. However, the social benefits of the road might be quite

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

large if, in alleviating congestion, it shortens journey times and reduces fuel costs to users. However, the road might produce externalities, such as noise, which are suffered by households living along the road. A financial analysis would deem the road unprofitable, since revenue was less than costs. A CBA would deem the road to be economically feasible if the social benefits of, or people's willingness to pay for, journeys along the road exceed the opportunity cost of resources used to build it. CBA would also include in the appraisal the externality or disutility from noise suffered by households living along the road (e.g. measured by way of depreciation in property values), in judging whether the road was socially justified.

Thus, CBA considers the costs and benefits to all sectors of society from a transport project, not just the benefits to people using the transport facility. Indeed, the need for CBA in transport arises precisely because it is affected by so many distortions and non-marketed outcomes: the "public good" nature of many roads; environmental externalities such as noise and air pollution; the fact that items such as non-working journey time and the value of preventable fatalities and injuries are not explicitly priced in the market; and the fact that private transport (e.g. cars) are highly taxed (through fuel taxes) while public transport (typically railways) is subsidized by the state.

3. Identification of costs and benefits

A CBA of a transport scheme must be undertaken relative to some alternative, which is usually the "do nothing" scenario. The "do nothing" scenario is a realistic view of what is likely to happen in the absence of the scheme. Estimating traffic flows with and without the scheme involves some form of transport modeling. For example, a model to estimate future traffic growth along a route without any increase in route capacity; an assignment model to estimate the volume of traffic induced by the expansion of road capacity; a modal choice model to estimate the numbers of people who would switch between public and private transport, etc. Only when these changes in physical quantities have been documented and estimated, over the life of the scheme, can we begin to place monetary values on them, and assess whether the benefits of the scheme are likely to exceed its costs.

4. Relevant costs and benefits

In CBA, all relevant costs and benefits associated with the project are enumerated over the life of the project. For example, in evaluating a motorway scheme, the relevant costs would include land, construction, and maintenance costs of the road. The benefits would include savings in journey time, fuel, accidents, noise, and any other technological externalities, i.e. any externality that affects production or satisfaction (utility) of consumers.

5. Market prices and "shadow prices"

Optimality in an economy requires marginal rates of substitution (MRS) (marginal values in the demand for goods) to be equal between consumers; if this is not the case, then the exchange of products between consumers will make one or more consumers better off and leave no consumer worse off. Optimality also requires marginal rates of transformation (MRT) (marginal costs of factors of production) to be equal in the production of goods; if this is not the case, then by substituting factors of production between industries it would be possible to reduce aggregate inputs with no change in aggregate output. Finally, optimality requires MRS = MRT, to ensure consumers' preferences and welfare are maximized (Winch, 1971; Layard and Walters, 1978). Under these conditions, market prices reflect appropriate MRS and MRT, and are used to measure costs and benefits in CBA.

However, market prices do not reflect appropriate MRS and MRT where there are monopolistic prices, unemployed resources, subsidies, taxes, constraints on use, public goods, and externalities. In such situations, implicit or "shadow prices" are used to derive appropriate MRS and MRT in CBA; and these shadow prices normally replace observed "market prices" in a CBA.

5.1. Monopolistic prices

In constructing a road or rail line, all landowners along the line of the road are in a monopoly position, since other parcels of land, not on the line of the road, are not substitutes in production. Compulsory purchase powers exist to avoid this monopoly situation, or "hold-out" problem, allowing transport authorities to compulsorily purchase the land "at a fair market price," usually the current use value. Both compulsory purchase and absence of compulsory purchase can lead to inefficiency.

Ultimately, without compulsory purchase powers, the hold-out problem might result in the road not being built at all, if one landowner tries to capture more than the entire economic rent of the road scheme. The absence of compulsory purchase usually results in less land than is optimal being purchased for the road, resulting in a resource misallocation cost. Moreover, the higher price of land under the hold-out problem creates an incentive to substitute other inputs for some of the land, e.g. capital, creating decks of roads. Unfortunately, compulsory purchase can also lead to inefficiency and too much land being used for the road development, with compensation for the acquisition of the land being less than the utility loss to its owner. This arises because, under compulsory purchase, land is acquired at a "fair market price" that would be accepted by a willing seller and a willing buyer. However, compulsory purchase invariably involves purchasing land from a non-willing seller, whose "true" reservation price for the land is above its fair market price (otherwise it would have already been sold!).

Optimal land acquisition for the road occurs where the marginal value product (MVP) curve for the road – initially sloping upwards, to reflect the fact that the value of the new road increases as more land is acquired, e.g. to make a dual carriage way, but then declining as a third, fourth, and fifth lane add increasingly less to time and petrol cost savings – equals the marginal cost of acquiring contiguous parcels of land (MCc).

The market price of land paid under land compensation legislation for compulsory acquisition is the marginal cost of acquiring randomly scattered (MCr) houses: this is a horizontal cost line (i.e. the supply is perfectly elastic) since each additional house can be purchased at the same market price. The supply curve for contiguous properties, MCc, is less elastic, because of the rising probability of encountering sellers whose reservation price exceeds the average market price of the property (Munch, 1976; Willis et al., 1998). It is a function of the number of properties purchased in a given period. The slower the rate of land acquisition the closer MCc approaches MCr.

Monopoly problems may also pertain to other inputs. Suppose a cement cartel fixes the market price of cement well above its marginal cost (MC). Should the market price or the marginal production cost of cement be used to value this input? If the production of cement increases by the amount used in the road scheme then it should be valued at its MC. However, if no extra cement is produced then cement should be valued at its alternative use value, which is the monopolistic market price for cement.

5.2. *Unemployed resources*

The guiding rule for estimating the "shadow price" of a resource, whether land, labor, or capital, is the value it produces in the use from which it is transferred, plus any additional monetary amount above its existing use value required to induce the transfer of the resource to the new transport project.

Where labor is drawn from the unemployment pool, the "true" cost of unemployed labor to society on the project is zero, apart from any loss to the "black economy" (e.g. as a house-husband) and any loss of leisure time. Unemployed labor is not valued in terms of unemployment payments received from the state, since unemployment pay is simply a transfer payment from the government to the unemployed person. Thus, if $400 per week has to be paid to induce an unemployed person into a job, and he or she no longer receives $200 in unemployment payments, which now revert to the rest of the community, and his or her contribution to the "black economy" is $75 per week, then the cost to society of engaging his or her labor is only $75 per week, since this is all that is being lost. Increasing levels of unemployment increases the probability that the worker will be drawn from the unemployment pool (Epp, 1979). However, an unemployed worker will not remain unemployed forever, so that the social opportunity of labor and other resources needs to be estimated over time (Willis and Saunders, 1988), which is not any easy task. Thus projects that are infeasible, in CBA terms, in conditions of full employment can become feasible in conditions of high unemployment.

A transport improvement typically reduces journey times. Where this impinges on working time, the valuation is based on the employer's valuation of time saved (gross wage paid plus employer's employment tax plus pension contributions) rather than the employee's value. However, many reduced journeys times are for non-working trips: leisure, shopping, and journey to work trips. The value of non-working time depends upon whether there is unrestricted substitution between work and leisure. Unrestricted substitution implies non-working time can be valued in terms of earnings foregone. Constrained substitution (e.g. where hours of work are fixed) implies an opportunity cost of time based on foregone leisure (e.g. recreational and social experience). This value of non-working travel time can vary according to context (Wardman, 1998).

5.3. *Subsidies*

The principle adopted in dealing with costs where a resource is subject to subsidy is to estimate the opportunity cost of the subsidized resource. Roads are often built across agricultural land. Subsidies under the Common Agricultural Policy (CAP) in the EU comprise direct subsidies (e.g. hill sheep subsidy), price guarantees, and tariffs at the frontier. Where agricultural land is used in road construction, its opportunity cost is not the lost agricultural output minus agricultural inputs, nor its market price (since this is artificially raised by the price guarantees and tariffs at the EU frontier), but what it would cost to replace this lost agricultural output on world markets (Willis et al., 1988). Resources subject to subsidies should be measured not at their market price but at their resource or factor cost (market price minus subsidy element). In urban areas the market value of any housing land taken for road construction should be adjusted to take account of subsidies to housing (e.g. income tax relief on mortgage interest payments).

5.4. *Public goods and externalities*

Non-marketed goods can be valued in a number of ways. Revealed-preference models in transport can be used to derive the value of time-savings in transport. Methods of measuring the utility from public goods and externalities such as green belts, nature conservation and other environmental sites, designated land, landscape impacts, air pollution, noise, etc., can be measured by a production function approach (e.g. expenditure to mitigate or avert the externality), revealed-preference approaches (e.g. how much more people are prepared to spend on a house to acquire public goods and positive externalities), and stated-preference approaches (e.g. how much people say they would be willing to pay to acquire the good or would be willing to pay to avoid its loss) (for an overview of all these techniques, see Garrod and Willis, 1999).

5.5. *Taxation*

Taxed inputs are measured at their factor cost and not their market value. Again, the principal concern is with measuring the cost of resources corresponding to their real use excluding transfer payments (taxes). Thus, in calculating the operational benefits of a new road, fuel savings are valued at the market price of fuel minus the tax on fuel. If the market price of fuel is $1 per liter which includes tax of 60 cents, and the new road saves 1 liter of fuel per journey, then the value of fuel saved is not $1 but 40 cents. By saving a liter of fuel, the savings to the economy is the 40 cents it costs to extract, refine, and distribute the fuel. The 60 cents tax on the liter of gasoline now remains with the car owner instead of the government. An exception to this treatment of fuel taxation would be in an economy subject to constraints in the import or production of energy. If fuel was diverted from an alternative use in the economy, then its resource value would be what this alternative use was willing to pay for it, and this price would be inclusive of tax.

5.6. *Marginal cost of public funds*

Although resources should be valued at their factor cost (market price minus tax), the public funds injected into the transport scheme should be valued at the capital amount multiplied by the marginal cost of public funds (MCF). The MCF occurs because of the deadweight loss from taxation to raise public revenue. All taxes, except those on "economic rent," create distortions in the economy. Income tax deters work effort, so the welfare cost of taxation can be substantial (Browning, 1987). The marginal excess burden of raising extra revenue under the tax system varies between taxes (Fullerton and Kodrzycki Henderson, 1989). Government spending on a project should not proceed until

$$\sum \text{MRS} \geq \text{MCF} \times \text{MRT}.$$

Some public sector transport projects are financed by borrowing rather than taxation. But this raises the prospect of public investment displacing private investment, which also implies that there is a marginal opportunity cost for raising public funds by borrowing.

5.7. *Constraints on use*

Land-use planning controls can create constraints on land-use, which give rise to a divergence between MRS and MRT. Resources subject to constraints on use should be valued at their opportunity cost rather than their current use, since the former is what society is losing by transferring the land to transport rather than using it for, say, housing. Care must be taken in specifying the alternative use. For example, the SACTRA Report (Standing Advisory Committee on Trunk Road Assessment, 1992) argued that the opportunity cost of land (alternative use precluded by planning control) equated with the value of amenity loss from a road development. However, land use zoning is instituted through the political market place on the basis of voting or pressure group mechanisms. Public choice theory points out that preference revelation under such mechanisms is an imperfect representation of individuals' utility and that too much or too little "public good" provision will result, compared with an economic optimum, depending upon whether those with environmental preferences are in a majority or in a minority.

5.8. *Summary*

There are clearly many things wrong with market prices, where market prices are defective representations of MRS. However, market prices may be better than the alternatives: prices that are derived rather than observed. Market prices represent millions of people providing information about MRS, and induce millions of people to adjust their purchases and sales to prices so that those prices reflect approximately what an extra unit is worth to all users. It can be extremely costly to acquire improved information on "shadow prices." Thus, the existence of defects in market prices does not automatically mean that some derived price or "shadow price" is automatically better.

6. Social discount rate: opportunity cost versus time preference

What social discount rate (SDR) to use in a CBA is a perennial problem. There are two theories on the issue:

- Social time preference (STP) theory sees the discount rate as society's marginal rate of substitution of consumption between two points in time: it is the price that has to be paid to induce people to delay their consumption and make available the resources for investment.
- Social opportunity cost (SOC) theory argues that the discount rate should reflect society's next best alternative use to which the funds employed on the project might been put.

In theory, in equilibrium and in the absence of taxation in an economy, the STP rate equals the SOC rate. However, in practice the SOC rate is usually higher than a STP rate.

7. Decision rules

7.1. The net present value, internal rate of return, and benefit/cost ratio

Through discounting it is possible to compare different transport projects that have different time patterns of costs and benefits, and different lengths of life. Discounting is necessary because people have time preferences for benefits (a preference for benefits now rather than at some point in the future), or because there is an opportunity cost of investing in transport – the loss of benefits elsewhere in the economy from this investment in transport. The net present value (NPV) of a project compares the discounted stream of benefits (B) with the discounted stream of costs (C) over every year of the life of the project (t):

$$\text{NPV} = \sum B_t(1+i)^{-t} - \sum C_t(1+i)^{-t},$$

where i is the discount rate. Any project is socially profitable if NPV > 0, or if projects are mutually exclusive the project with the highest NPV is implemented.

Decisions can also be based upon the internal rate of return (IRR) earned by the project. The IRR is a discount rate (i) that equates the discounted value of the benefit and cost streams, i.e. a rate that gives an NPV of zero. The advantage of the IRR is that the rate of return on the investment in the project can be compared with the rate of return at which the money for the investment is borrowed. If the IRR is greater than or equal to the interest rate at which the money is borrowed, then the project is deemed to be feasible, but not otherwise. The disadvantage with the IRR is that rates of interest can change (this can be included in the NPV approach); there may be more than one IRR if the investment profile is –, +, –; and where two projects are mutually exclusive, the NPV may approve project A, while the IRR suggests project B should be implemented. When two projects are mutually exclusive, to determine which project should be implemented requires an assessment of a hypothetical project that is the difference between projects A and B.

Where government finance is constrained in any time period, the *B/C* ratio can help identify which set of projects produces most benefits relative to costs. The *B/C* decision rule sanctions projects sequentially and lexicographically from those with the highest *B/C* ratio until all investment funds are exhausted. However, the *B/C* ratio rule is not infallible. For example, if the budget constraint was 100, and there were three projects (X, Y, Z) with gross present values of costs, gross present value of benefits, and *B/C* ratios respectively of (X: 100, 200, 2.0), (Y: 50, 110, 2.2), and (Z: 60, 130, 2.17), the *B/C* ratio criterion would select Y, after which the budget constraint would prevent any other project being undertaken. However, net benefits are maximized by selecting project X, even though it has the lowest *B/C* ratio (Pearce and Nash, 1981).

B/C ratios higher than 1.0 are also justified as the appropriate decision rule where there are "irreversible effects" (Henry, 1974). An irreversible decision is one that reduces the long-term variety of options or choices in the future. For example, destroying historic buildings for a road expansion, or an archaeological site to build a parking lot. Where the amenity value for historical features is growing over time, but this growth rate is not known with certainty, then a *B/C* ratio decision criterion >1.0 is economically justified. Hodge (1984) provides some examples.

7.2. Social welfare

If the benefits of a transport project exceed its costs, then the project may be deemed desirable. Whether it is sanctioned depends upon the distribution of the utility changes. The Pareto criterion sanctions a project that makes at least one person better off and leaves no one worse off.

However, any transport project invariably entails an increase in utility for some people and a decrease in utility for others. Sanctioning a project then involves weighing-up the positive and negative utility changes. There are several methods of weighing up positive and negative welfare changes. First, the Kaldor–Hicks criterion states that if the gainers could compensate (bribe) the losers (to accept the change) while continuing to gain themselves, then the project should proceed. But if compensation is not actually paid, it can only be claimed that the project offers a potential Pareto improvement. Second, weights could be attached to the positive and negative utility changes to reflect the distributional decisions of society, e.g. marginal rates of income tax. Hence, if money or utility changes were:

rich +200 and poor –100, and the distributional weights were 0.5 and 1.5, respectively, the project would be deemed undesirable since[a]

$$\Delta W = 0.5(200) + 1.5(-100) = -50 < 0.$$

Unfortunately, the problem with tax rates is that they are neither a continuous function nor are different tax rates consistent with each other in an equity sense. Rather, tax rates are political expedients. In addition, the marginal cost of redistribution can be high (Browning, 1996).

Third, a social welfare function (of which the Pareto criterion is a special case) could be employed. A simple welfare function might be

$$W = \frac{1}{\alpha} u_{\mathrm{R}}^{\alpha} + \frac{1}{\alpha} u_{\mathrm{P}}^{\alpha},$$

where α is a distributional weight. The utilitarian position is $\alpha = 1$, i.e. the benefits (or utility u) are simply summed across the rich (R) and poor (P). The smaller α is, the more egalitarian the criterion; that is, if $\alpha < 1$, say –1, then social welfare depends more on the benefits to the sector (or group) with the lowest utility or income. The Rawlsian position is $\alpha = -\infty$, where social welfare comes to depend solely on the income of the poorest person or group.

8. Timing of investments

Decisions may have to be made not only about whether a transport project should be undertaken but also when it should be undertaken. Where demand or costs are changing over time there is an optimal construction date. This is determined by calculating the NPV at year 0, and the NPV at years 1 to n. For example, if benefits are growing over time, and there are some savings by delaying construction costs, then the NPV of the project will increase in the future. The optimal timing of investment in the transport project is the year in which the NPV is maximized. Unfortunately, the analysis of the optimal timing of public investments is by no means routine: in fact it is extremely rare. But it is extremely relevant to public investment in transport where demand or costs are changing over time.

9. Multiple period investments and budget constraints

Typically in transport planning, a number of projects are implemented over more than one time period, with each time period being subject to a budget constraint.

[a]The anomaly here is that if the project went ahead, the rich could compensate the poor and leave them no worse off, with the rich still gaining 100. Indeed, if the entire 200 were redistributed from the rich to the poor (e.g. by a fare increase) the poor would gain by $1.5 \times 100 = 150$, which would be better than not undertaking the project at all!

The problem now becomes one of optimizing NPV over more than one time period: this is the realm of dynamic linear programming. The problem is not solved by ranking projects in descending order of their NPVs or *B*/*C* ratios, and then going down the list constructing as many projects as the present period's budget permits. This may not maximize the overall NPV across time periods, because this concentrates on the absolute advantage amongst projects within each period instead of looking at the comparative advantage among projects between periods. Appendix 1 provides a simple example illustrating the problem.

10. Risk and uncertainty

Transport projects are subject to risk and uncertainty about the outcomes, particularly those based on technological advances. Where knowledge exists about the probabilities of various outcomes, then risk can be assessed. Which project(s) will be chosen, from a set of projects, depends on the decision-maker's attitude to risk. If two transport projects are mutually exclusive and have the same expected positive NPV, but one project has a greater dispersion of probable outcomes (from negative to positive) than the other, a risk-neutral decision-maker would be indifferent to the two projects. If the decision-maker is risk-averse then he or she will be averse to the probability of losses, particularly any prospective large losses, and hence will give greater weight to this in making a decision between projects. Indeed, while risk neutrality means that a decision-maker will accept any project with a positive expected NPV, this is by no means certain for a risk-averse decision-maker. Various techniques exist to assess risk (Moore, 1983).

Decision-makers maximize utility by adopting the expected utility criterion: choosing those projects that maximize probability and outcomes over a large number of projects. No other decision criteria will outperform this expected utility strategy, and it is the strategy that governments, in inaugurating hundreds of transport projects each year, should adopt. However, if the decision-maker is only confronted with one or two projects, then the choice of project may be more heavily dependent upon his or her attitude to risk and uncertainty.

Where probabilities of outcomes (future costs and benefits) are unknown, the decision-maker has to adopt some rule to select a project in the face of this uncertainty. Attitudes to uncertainty might mean that the decision is based upon (from most to least risk averse):

- Maxi-min: choosing the policy or project that is least likely to make a loss.
- Laplace criterion: placing equal probabilities on each outcome (on the grounds that we do not know which one is likely to occur) and selecting the highest NPV project (a modification of the expected value).
- Regret theory: selecting the project that minimizes regret (difference between outcomes of the different policy or project options) (Sugden, 1993).
- Maxi-max: choosing the policy or project giving the highest net social benefit.

11. Issues in CBA

Many of the principles of CBA are well established (Brent, 1996). However, there are a number of issues surrounding the application of CBA. These include:

- the framing and context in the measurement of benefits;
- the estimation of future costs, e.g. the future environmental costs of a road scheme;
- the estimation of environmental costs and benefits associated with transport.

11.1. Framing and context

The Hicksian compensating surplus (CS) measure for a change in transportation provision is the maximum (minimum) amount of money that can be taken from (or given to) an individual in order to leave him or her indifferent between his or her *ex ante* and *ex post* transport program situations. Conventional contingent valuation (CV) procedures tend to value the environmental aspect of each transport improvement (q) of a program as if it was a single independent element, maintaining all other elements at their initial level. Hoehn and Randall (1989) have shown that such independently valued elements of a program cannot be used directly in a CBA of individual schemes without biased results. This occurs because the independent valuation and summation (IVS) procedure is equal to the total CS only if the K program schemes are independent, which is unlikely to be the case in a transport program that comprises a large number of schemes. IVS will over-estimate the true valuation of transport benefits if program schemes are substitutes. As the number of transport schemes increases so substitution effects increase. Consequently, when IVS-derived values are adopted from a series of transport schemes, the attributes derived from the different "independent valuation" studies will result in an overestimate of CS. Too many of the non-statutory schemes will pass the cost–benefit test.

A program of transport scheme improvements can be correctly valued by either a sequential (SEQ) approach or a simultaneous (SIM) approach. When elements of a program are measured sequentially, the order in which the subcategory benefits are valued influences the values ascribed to these elements. Schemes that are valued first are valued more highly than schemes valued later. The value ascribed to each program element is not unique, but depends on the sequence selected: the path dependency issue. The sequential (SEQ) path approach derives separate valuations for the K components or schemes by a valuation path beginning at $(q10, q20, \ldots, qk0)$ and ending at $(q11, q21, \ldots, qk1)$. A sequential path shifts program schemes one at a time from their initial *ex ante* position to their *ex post* program prescription, in a sequence of valuation changes, $q10$ to $q11$, then $q20$ to $q21$, and so on until all the program schemes are complete with $qk0$ to $qk1$.

In the simultaneous approach, respondents value all the transport schemes in the program as a whole, and also value each scheme or attribute within the program as a part-worth. Choice modeling is a simultaneous approach, since each transport attribute or scheme is simultaneously valued and traded off against all others in the set.

Economic theory predicts that IVS will lead to higher estimates of benefits across a transport investment program than either the SEQ or SIM approach. Thus, the adoption of an IVS approach may over-estimate the benefits of each transport scheme, with too many schemes passing the CBA test, and too much investment in transport schemes. A sequential valuation strategy is required to estimate transport benefits if these are to be used in a CBA to optimize a project investment program over time in different capital budgeting periods. If a transport planning organization considers improving any or all of its transport operations within a budget period to identify an optimal investment strategy, a simultaneous approach needs to be adopted.

11.2. Future costs

Environmental costs from transport arising in the future might be difficult to estimate; and, because they arise in the future, they may be omitted from an *ex ante* project appraisal. In the UK, the Land Compensation Act 1973, Part 1, states, for land not taken for the road scheme but affected by it, that where the value of an interest in land is depreciated by the physical factors arising from the use of a new road (noise, vibration, smell, fumes, smoke, artificial lighting, and the discharge onto the land of solid and liquid substances), the Secretary of State shall pay compensation. However, the UK Valuation Office Agency only evaluates such claims 1 year after the road has opened to traffic, by assessing how much depreciation in value has been caused in total by the new road, deciding what proportion of the reduction in value has been caused solely by the above physical factors, and deducting any enhancement in value arising from the new road.

11.3. Environmental costs

Some environmental costs, such as noise, can be estimated relatively using hedonic price models (HPMs), although if noise is confounded with traffic fumes it might be difficult to disentangle these two effects using HPMs. Confounding can be eliminated by using stated-preference methods to derive people's willingness-to-pay not to experience traffic noise. Other externalities such as air pollution are more difficult to value. PM_{10} (particulate matter of 10 μm or less) and sulfur dioxide from vehicle exhausts can cause respiratory problems to some people.

However, health effects and fatalities from PM_{10} and sulfur dioxide occur many years in the future, and typically reduce life expectancy by only a few months. There is some debate over how to estimate this value (Rabl, 2003), but it is clear that the value of statistical life (VOSL) applied to road accident fatalities is not applicable to deaths due to air pollution from transport. Carbon dioxide emissions give rise to global warming, and there is considerable debate over the social value of carbon, with estimates varying from US \$6 to US \$197 per tonne of carbon rising over time (Pearce, 2003). A more accurate and robust estimate will require further research, but whatever value is adopted it will have implications for the future of transport policy.

12. Conclusions

The institutional structure within which transport proposals are formulated and appraised can have an important influence on CBA outcomes. This is especially likely to be the case if the cost–benefit analyst is not neutral with respect to the outcome of the CBA. Politics may favor public over private transport. In such a case the application of CBA may not result in greater efficiency in the economy. CBA is both a science (firmly based on economic theory) and an art (requiring practical skill in deciding what is worth including, and deriving values for these variables). Unfortunately, the art element leaves plenty of scope for the manipulation of the results.

Appendix. The sequence of investments in the presence of budget constraints

Assume the group of projects consists of just two: a road improvement scheme and a rapid transit improvement scheme, and both cost the same amount, \$100 million each. The budget constraint permits only one to be built now (year 0), the other being forced to wait until the next budget period (in 5 years time).

Assume the discount rate is 5% and that the road improvement yields benefits from year 1 to year 20 of

$$\$7\text{m} \times (1-1.05^{-20})/0.05 = \$87.235\text{m}$$

plus benefits into perpetuity from year 21 onwards of

$$\$20\text{m} \times 0.05^{-1} \times 1.05^{-20} = \$150.756\text{m}$$

giving an NPV of

$$\$124.662\text{m} + \$150.755\text{m} - \$100\text{m} = \mathbf{\$137.991m}.$$

The rapid transit improvement if built in year 0 would produce benefits in perpetuity of

$$\$10\text{m} \times 0.05^{-1} = \$200.000\text{m}$$

giving an NPV of

$$\$200\text{m} - \$100\text{m} = \mathbf{\$100.000m}.$$

A simple decision rule would therefore be to undertake the road improvement in year 0, and the rapid transit improvement at the start of the next budget period at year 5.

However, if the NPV for each project is calculated as if both developments started in year 5, then the road improvement yields benefits year 5 through to year 20 of

$$\$7\text{m} \times 1.05^{-5} \times (1 - 1.05^{-15})/0.05 = \$56.929\text{m}$$

and PV of benefits from year 21 of

$$\$20\text{m} \times 0.05^{-1} \times 1.05^{-20} = \$150.756\text{m}$$

and PV of construction costs in year 5 of

$$\$100\text{m} \times 1.05^{-5} = \$78.353\text{m}.$$

Hence, the NPV is

$$\$56.929 + \$150.756\text{m} - \$78.353\text{m} = \mathbf{\$129.332m}.$$

The rapid transit commencing in year 5 produces benefits of

$$\$10\text{m} \times 1.05^{-5} \times 0.05^{-1} = \$156.705\text{m}$$

and PV of construction costs in year 5 of

$$\$100\text{m} \times 1.05^{-5} = \$78.353\text{m}.$$

Hence, the NPV of the rapid transit project commencing in 2005 is

$$\$156.705\text{m} - \$78.353\text{m} = \mathbf{\$78.353m}.$$

The optimal sequence of projects is therefore rapid transit improvement in year 0, followed by the road improvement in year 5 (Table 1). This produces higher overall net benefits, compared with a rule of maximizing the NPV in the first time period. In the presence of budget constraints, choosing the project with the

Table 1
Maximization of NPV across budget periods ($ millions)

	Commencement date		
Transport improvement	Year 0	Year 5	Overall net benefits
Road improvement	137.991	129.332	229.332
Rapid transit scheme	100.000	78.352	216.344

highest NPV today may result in choosing the inferior course of action. In the presence of budget constraints it is not the absolute advantage of one project over another, but rather the comparative advantage between periods that determines the optimal sequence of development. The objective is to maximize the combined NPV of the chosen projects over time.

References

Brent, R.J. (1996) *Applied cost–benefit analysis*. Cheltenham: Edward Elgar.

Browning, E.K. (1987) "On the marginal cost of taxation," *American Economic Review*, 77:11–23.

Browning, E.K. (1996) "The marginal cost of redistribution: a reply," *Public Finance Quarterly*, 24:63–74.

Epp, D.J. (1979) "Unemployment and benefit-cost analysis: a case study test of a Haveman–Krutilla hypothesis," *Land Economics*, 55:397–404.

Fullerton, D. and Y. Kodrzycki Henderson (1989) "The marginal excess burden of different capital tax instruments," *Review of Economics and Statistics*, 71:435–442.

Garrod, G.D. and K.G. Willis (1999) *Economic valuation of the environment*. Cheltenham: Edward Elgar.

Henry, C. (1974) "Investment decisions under uncertainty: the "irreversibility effect," *American Economic Review*, 64:1006–1012.

Hodge, I.D. (1984) "Uncertainty, irreversibility and the loss of agricultural land," *Journal of Agricultural Economics*, 35:191–202.

Hoehn, J.P. and A. Randall (1989) "Too many proposals pass the benefit cost test," *American Economic Review*, 79:544–551.

Layard, P.R.G and A.A. Walters (1978) *Micro-economic theory*. New York: McGraw-Hill.

Moore, P.J. (1983) *The business of risk*. Cambridge: Cambridge University Press.

Munch, P. (1976) "An economic analysis of eminent domain," *Journal of Political Economy*, 84:473–497.

Pearce, D.W. and C.A. Nash (1981) *The social appraisal of projects*. London: Macmillan.

Pearce, D.W. (2003) "The social cost of carbon and its policy implications," *Oxford Review of Economic Policy*, 19:362–384.

Rabl, A. (2003) "Interpretation of air pollution mortality: number of deaths or years of life lost?" *Journal of Air and Waste Management Association*, 53:41–50.

Standing Advisory Committee on Trunk Road Assessment (1992) *Assessing the environmental impact of road schemes*. London: HMSO.

Sugden, R. (1993) "An axiomatic foundation for regret theory," *Journal of Economic Theory*, 60:159–180.

Wardman, M. (1998) "The value of travel time," *Journal of Transport Economics and Policy*, 32:285–316.

Willis, K.G. and C.M. Saunders (1988) "The impact of a development agency on employment: resurrection discounted?" *Applied Economics*, 20:81–96.

Willis, K.G., J.F. Benson and C.M. Saunders (1988) "The impact of agricultural policy on the costs of nature conservation," *Land Economics*, 64:147–157.

Willis, K.G., G.D. Garrod and D.R. Harvey (1998) "A review of cost–benefit analysis as applied to the evaluation of new road proposals in the UK," *Transportation Research D*, 3:141–156.

Winch, D.M. (1971) *Analytical welfare economics*. Harmondsworth: Penguin.

Chapter 30

MULTICRITERIA EVALUATION OF TRANSPORT POLICIES

RON VREEKER and PETER NIJKAMP
Free University Amsterdam

1. Introduction

Transport has been at the origin of the economic development of many countries. Transport is to a large extent a derived demand, but its evolution also determines the welfare of cities, regions, and nations. This two-faceted role means that it is necessary to evaluate the role that transportation plays in society with care.

In general, transportation – and spatial interaction – mirrors the socio-economic, spatial, and political dynamics of societies. In the 1960s, a period with unprecedented economic growth in many Western countries, transportation policies were strongly oriented toward network and capacity expansion. From the 1970s onward, however, the "limits to growth discussion" marked a more modest role for infrastructure policy, in which the efficient use of existing networks received more attention than the straightforward expansion of the physical network. The 1980s saw the emergence of new views, reflected in the environmentalist movement (e.g. green parties), with its strong concern about the negative impact of transport on the general quality of life (Banister et al., 1997). From the 1990s onward, there has also been considerable interest in the potential of modern technologies (e.g. telecommunication) for network improvement, notably in the context of the "missing networks discussion" and of the evolving network economies (e.g. Handy and Mokhtarian, 1995; European Industrial Bank, 1996).

Predicting the spatial mobility and transportation effects of the great variety of structural changes in our dynamic economies is complex, especially in the long run. The societal value of mobility – from a long-term point of view – can only be assessed if sufficient insight is obtained into the endogenous and exogenous dynamics of the spatial systems.

Transport planning is therefore addressing complicated questions, which are partly of a methodological or theoretical nature, and partly a matter of a practical policy. Among policy-makers and researchers, awareness has grown that

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

transportation research and planning should extend its scope by focusing on those systems or domains that interact with transport.

One of the basic policy questions in a changing planning situation concerns the relationship between transport infrastructure and land use patterns. For example, modern settlement systems (at both the urban and metropolitan levels) may have mutually contrasting impacts on the transportation sector. Transportation appears to be a major stimulus for the development of new urban centers, while at the same time, it is endangering balanced urban development. Positive and negative externalities appear to play an ambiguous role in the spatial and urban evolution of our modern society. In this context, the relationship between the transportation sector and land use is becoming more interwoven. In particular, the following objectives appear to occupy a prominent place in policies:

- economic efficiency, reflected in the increased competitiveness of regions through an improvement in connectivity;
- social equity, reflected in more equal opportunities for better access to transport facilities (for different socio-economic groups; for less central areas);
- environmental sustainability, reflected in more emphasis on coping with the negative externalities of the transport sector, such as pollution, noise, landscape decay, congestion, and lack of safety.

It goes without saying that, in many cases, these objectives are not immediately compatible. This also implies that transport impact analysis and evaluation is fraught with multiple problems, as the assessment of the spatial–economic consequences of new transport systems is a far from easy. The multi-dimensional nature of these transport policy objectives also makes the application of conventional financial evaluation tools problematic.

We may conclude that, to a large extent, new socio-economic and political developments are being projected onto the field of transportation planning. This also implies that transportation planning cannot be undertaken in isolation from other fields of planning and policy-making (e.g. economic, environmental, or technological policies), so that currently transportation planning is, by definition, a multidimensional activity focusing on multiple public and private interests with a particular emphasis on conflict resolution and on sustainable development. Transportation is therefore a multi-faceted activity, which cannot easily be judged with a simple yardstick.

The aim of this chapter is to review the use of multicriteria analysis (MCA) in transportation planning. The chapter offers various historical and methodological reflections on evaluation theory with a particular view on the potential of multiple criteria decision tools. A classification of various methods as well as a concise overview of the use of these methods in transportation planning will be offered as well.

2. Evaluating transport policies

Policy analysis or evaluation, in general, aims at rationalizing planning and decision problems by systematically structuring all relevant aspects of policy choices (e.g. the assessment of alternative choices). Evaluation is usually not a one-shot activity, but takes place in all phases of decision-making. In addition, a systematic support to complex planning and decision problems presupposes a balanced treatment of too many details and too little information. Additionally, the results of an evaluation procedure have to be transferred to policy-makers in a manageable and communicable form, particularly because the items of an evaluation are usually multidimensional in nature and often include incommensurable or even intangible aspects of transport.

Finally, the "planning environment" is usually highly dynamic, so that judgements regarding the political relevance of items, alternatives, or impacts may exhibit sudden changes, hence requiring policy analysis to be flexible and adaptive in nature. Rigid evaluation techniques run the risk that an evaluation does not cover all planning issues in a satisfactory way, so that biased decisions may be taken.

Every planning or policy-making process presupposes choices to be made, regardless of the evaluation methods that have been adopted. The difficult task of choosing among alternative courses of action (choice possibilities, options, etc.) cannot be circumnavigated and is common to all public planning problems. Besides the structure of a decision problem, the specific evaluation method used will also determine which data are requested for the policy analysis at hand. For instance, checklist approaches, cost–benefit studies, planning balance sheet techniques, goals achievement methods, MCA, and multiple objective programming models all have their own specific data needs. Clearly, assuming a cyclical model of planning implies also that the relevance of an evaluation technique has to be judged in the light of available data. In this regard, it is worth noting that monitoring is a necessary ingredient of an adaptive evaluation methodology, so that in each phase of the planning problem both the data and the evaluation method are critically judged.

Policy evaluation has become a major component of modern public planning. The history of plan and project evaluation, however, shows that various periods in evaluation can be distinguished.

The early history of plan and project evaluation before the Second World War was marked by a strong tendency toward a financial trade-off analysis. A good example can be found in the USA, where a specific emphasis was placed on the regional growth aspects of an integrated regional hydrological project in the Tennessee Valley.

Later on, much attention was focused on the judgement of alternative US military defense systems, based on costs-effectiveness principles. After the Second World War, social cost–benefit analysis (CBA) gained increasingly popularity in public policy evaluation. Cost–benefit methodology became a leading evaluation instrument until the 1970s, especially in countries with a market economy or a mixed economy. Several limitations inherent in CBA exist, including:

- *Accuracy of information*: in general, it is very difficult to assess in an accurate manner all direct and indirect impacts of a plan over a large number of periods, e.g. the effect of infrastructure on biodiversity.
- *Distributional equity*: in general, distributional effects are largely omitted in many evaluations, because the assessment of these effects is considered to be a political activity. However, distributional effects can be taken into account by employing the notion of compensatory payments to attain a Pareto equilibrium.
- *Compensatory payments*: a decline in the utility of certain individuals or groups can be assessed by means of a monetary amount that compensates this decline. The notion of compensation is a direct application of welfare theory, but its precise amount is in general very hard to operationalize (due to lack of information on preference curves). Furthermore, compensatory payments are frequently only introduced as fictitious payments to assess social costs.
- *Discount rate:* the time preference of transport plan impacts is reflected by the social rate of discount, but an operational determination of this parameter is very difficult. The value of a social discount rate is not an unambiguous parameter, but is essentially the result of a socio-political decision.
- *Lifetime of the project*: the lifetime of a transportation project is sometimes hard to assess.

The limitations of CBA, however, were relaxed by the introduction of amendments to the method such as the planning-balance sheet method, goals-achievement method, or shadow project methods.

From the 1970s, a new class of evaluation methods arose, called MCA, much of it stemming from work undertaken in France (e.g. by Roy, Guigou, and Jacquet-Lagrèze). There are a number of reasons for the increasing influence of multicriteria evaluation techniques vis-à-vis conventional evaluation methods:

- the impossibility of including intangible and incommensurable effects in conventional, unidimensional evaluation techniques;
- the conflict nature of modern planning problems, so that – instead of a single decision-maker – various multi-level formal and informal decision agencies determine the final choice;
- the shift from conventional "one-shot" decision-taking to institutional and procedural decision-making where many political aspects play a major role;
- the desire in modern public decision analysis not to be confronted with a single and "forced" solution but with a spectrum of feasible solutions.

Over the course of the development of evaluation methods a closer orientation toward actual decision-making processes can be observed, which means that, besides a monetary CBA-based approach, complementary, in particular decision-making-based and institutionally-based, approaches have also been developed. These methods can be used for different purposes and in various contexts. In fact, one may distinguish at least four types of evaluation styles in the transport planning literature:

- a monetary decision approach, based, for example, on cost–benefit or cost-effectiveness principles;
- a utility theory approach, based on *a priori* ranking of the decision-maker's preferences using MCA;
- a learning approach, based on a sequential (interactive or cyclical) articulation of the decision-maker's view;
- a collective decision approach, based on multi-person bargaining, negotiation, or voting procedures.

In general, the efficiency check on transport decisions can most properly be carried out by means of CBA methods, while equity and sustainability checks need broader approaches based on MCA approaches. In addition, in the case of soft information (e.g. ordinal), the latter class of methods is also very suitable. Thus, one is essentially seeking the most appropriate decision support method in the case of a complex transport evaluation problem. To see how this can be embedded in a real-world context, one can consider the structure seen in Figure 1.

In the context of transportation planning, the systematic provision of information is extremely important, since many investment decisions have a process character. For example, a project to build a multi-modal terminal near a river will require the careful examination of all alternative locations and of all relevant judgement criteria (benefit–cost ratio, environmental impacts, risk, accessibility, and so forth). After a first deliberation, alternative solutions and proposals may be put forward by policy-makers, experts, or environmental action groups. Consequently, the evaluation of alternative plans has to be repeated. This iterative process continues, until a final decision has been reached. During the process of evaluation, generating alternatives, judging the efficiency of the projects, comparing the intangible effects and so forth, the expert has to inform responsible policy-makers, interest groups, etc., about the trade-offs involved in alternative plans, so that people know precisely how much they gain or lose in a certain respect, when the first plan is substituted for the second, and so forth. So the expert or planner is essentially a communication agent. Thus, the main task of the expert or planner is to rationalize decisions by offering a coherent, systematic, and surveyable frame of reference during the entire evaluation process of plans or projects.

Rather than employing evaluation or planning methods in a technocratic manner, the expert's responsibility can be seen as that of confronting decision-makers with

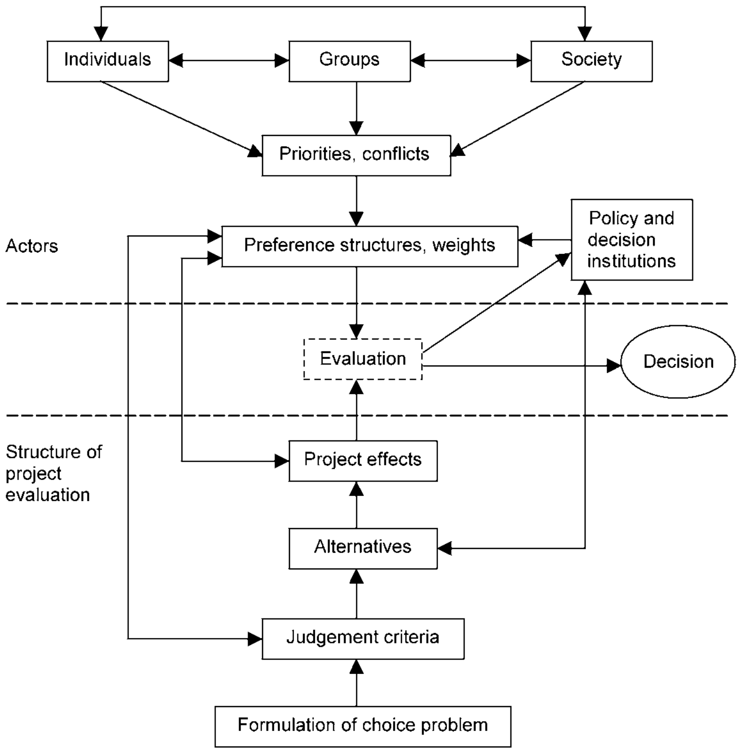


Figure 1. General structure of an evaluation problem.

the consequences of their priorities and choices. Such methods thus serve as an aid to improve the quality of decisions. Decision support techniques have a learning character, based on the interplay between suppliers and users of information regarding the broad spectrum of consequences and scarce resources used. It is necessary to have some understanding of the uncertainties, non-measurable effects, unforeseeable consequences, etc., so that policy-makers may also take account of risks, uncertainties, and stochastic information before taking a final decision on a plan concerning transportation or related land use.

3. Multicriteria evaluation methods

The presence of externalities, risks, long-term effects, spatial spillovers, unreconcilable interests, and qualitative information have been seen to generally preclude a meaningful application of unidimensional evaluation methods such as

CBA. Consequently, over recent decades attention has been devoted to the development of alternative evaluation methods, such as multicriteria evaluation methods. In the 1970s and 1980s an avalanche of multicriteria methods were developed. A review of various types of multicriteria evaluation methods is given by Zeleny (1982) and Vincke (1992), among others.

An essential difference between MCA and CBA is that CBA takes consumer preferences as the starting point and tries to achieve market conformity. In MCA, the preferences of the main actor (often the government) are not modeled as the sum of individual preferences. Instead, the government is assumed to have its own preferences and responsibilities.

Rather than a specific appraisal method, MCA is a family of methods. This family comprises a collection of around a hundred techniques that share some basic principles, but differ in other, mainly technical, aspects.

Janssen and Munda (1999) offer a typology of evaluation methods that helps to classify methods according to various characteristics (Figure 2). They base their typology on the following distinctions:

(1) The set of alternatives: discrete versus continuous problems. In evaluation practices, a distinction is often made between discrete and continuous problems. Discrete decision problems involve a finite set of alternatives. Continuous problems are characterized by an infinite number of feasible alternatives.
(2) The measurement scale: quantitative versus qualitative attribute scale. Some problems include a mixture of qualitative and quantitative information. Qualitative and mixed evaluation methods can handle this type of information to analyze the alternatives. If the information concerned is not exact, fuzzy evaluation methods can be applied.
(3) The decision rule: prices or priorities. The decision rule is unique for each method. Priorities used in MCA reflect the relative importance of the criteria considered in the analysis. In CBA, prices are used to calculate benefit–costs ratios. These prices are derived directly or indirectly from market prices or are assessed by means of various valuation methods.
(4) The valuation functions: standardization versus valuation. In order to make scores comparable, they must be transformed into a common dimension or into a common dimensionless unit. This can be done by transforming the scores into standardized scores by means of a linear standardization function, or by using value or utility functions. Utility and value functions transform information measured on physical measurement scales into a utility or value index.

In the last 25 years, developments in multicriteria techniques have been reflected in the progress of the three main theoretical schools: (1) utility or value system approaches; (2) programming methods; and (3) outranking techniques.

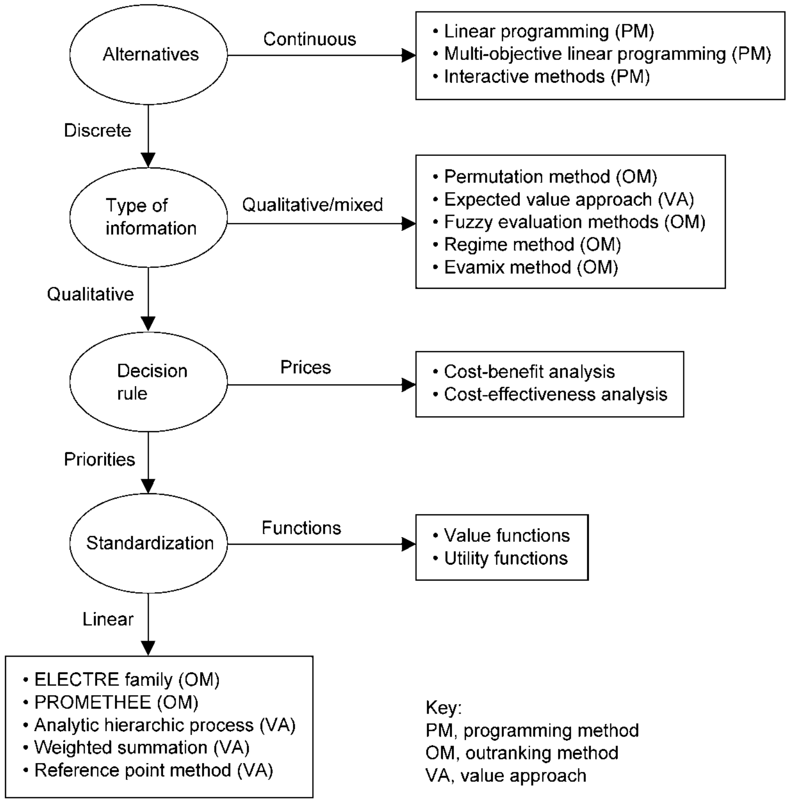


Figure 2. A typology of evaluation methods (Janssen and Munda, 1999).

3.1. *Utility or value system approaches*

The first school is the American school or value system approach. This approach uses multiattribute utility theory (MAUT) and multiattribute value theory (MAVT).

These approaches are based on the hypothesis that, for any decision problem, a value or utility function can be defined for the considered alternatives that the decision-maker wishes to maximize (Vincke, 1992). The role of the researcher is to determine this function.

MAUT and MAVT differ in that MAVT uses a value function to represent the outcomes for the alternatives considered, while MAUT relies upon the definition of a utility function, which allows the computation of an expected utility derived from each alternative. Although a utility function is also a value function, a value function is not necessarily a utility function. By constructing a value or utility function, a multicriteria problem is reduced to a unicriterion optimization problem.

In MAUT and MAVT, decision-makers are twice involved in the decision-making process. Their first involvement concerns the development of the utility function for each criterion. Their second involvement is related to the calculation of the expected utility by means of an aggregated utility function. In this phase of the decision-making process, decision-makers are asked to assign probabilities to certain outcomes. Preferences are assigned directly, aggregated, and used to develop the utility function.

3.2. *Programming methods*

Multi-objective programming (MOP) is an MCA method that simultaneously maximizes or minimizes several objectives, subject to a set of constraints. In contrast to the value system approach, the value or utility function is unknown. Although it is assumed that a value function exists, it is regarded as unknown, and no assumptions are made about this value function.

However, due to conflicts between the various objectives included in the analysis, a simultaneous optimization of these objectives is impossible. Therefore, instead of optimization, the aim of the decision-maker is to achieve a set of predefined goals as far as possible (satisficing behavior). MOP aims to find the set of efficient solutions, and divides the feasible set of solutions into a set of efficient solutions and a subset of inefficient solutions (Ballestero and Romero, 1998). "A set of solutions in a MOP problem is efficient or non-dominated, if their elements are feasible solutions such that no other feasible solution can achieve the same or better performance for all the criteria being strictly better for at least one criterion. This is a necessary condition to guarantee the rationality of any solution to a MOP problem."

As long as there is no additional information regarding the preference structure of the decision-maker, any choice from the set of efficient solutions is acceptable and reasonable. The final or best solution is the solution preferred by the decision-maker above all other efficient solutions.

3.3. *Outranking methods*

The third school of thought in MCA is the French school (e.g. Roy, 1968, 1972). The French school in MCA is based on pairwise comparisons between alternatives, and tries to build an outranking relationship between them. The associated techniques are not based on utility theory, and do not try to construct a mathematical model in order to represent the decision-maker's preference structure.

The foundation of the French school is that it is better to accept a result less strong than one yielded by MAUT or MAVT by avoiding introducing mathematical hypotheses, which are too strong, and by not asking the decision-maker questions that are too complicated.

Outranking allows for the performances of the alternatives, according to each criterion, to be compared with respect to a margin of error or indecision. It is obvious that some differences in the scores of the alternatives are irrelevant, and that any difference needs to be of a certain magnitude before it has some meaning in the comparison of alternatives. Therefore, outranking methods often make use of thresholds. Well-known outranking methods are ELECTRE, PROMETHEE, NAIADE and REGIME.

4. Applications of multicriteria analysis in transportation planning

Transportation planning offers a rich application field for MCA, as transportation evaluation comprises a variety of different policy perspectives, such as cost minimization (either from the side of users of transport systems or from the side of governments investing in infrastructure), consideration of landscape values (e.g. segmentation or visual beauty), minimization of ecological damage (protection of biodiversity, minimum noise annoyance, or air pollution), maximization of transport safety, or minimization of traffic congestion. Many of these policy perspectives cannot be achieved to the full extent simultaneously, as they may be more or less contradictory. Hence, they offer clear examples of a multiple criteria decision problem, for which MCA techniques may be helpful.

The fields of application range widely in transportation planning. They may concern supply side areas, such as the design or construction of new infrastructure or its embeddedness in local and regional landscapes. This supply side may refer to road transport (e.g. construction of motorways), railways (e.g. the location of new railway stations), aviation (e.g. the expansion of the number of runways at a given airport), etc. But the field of application may also concern the demand side, such as the optimal distribution of transport flows on the basis of various judgement criteria (user costs, environmental damage, etc.). So there is a vast application potential for MCA methods in the field of transport planning. The number of applications is formidable, and too large to summarize. We offer by way of illustration three concise empirical examples of the use of MCA in transportation planning.

4.1. *Multi-objective programming*

In the 1970s and 1980s, the application of MOP and related methodologies, such as CBA, to the evaluation of various transport investments was not uncommon (e.g. Roy and Jacquet-Lagreze, 1977). However, the number of studies that focus on MOP in the exploitation of transport networks is small.

During the 1980s, transportation network analyses have mainly been conducted by means of single-objective problems, such as the minimization of transport costs (Nijkamp, 1975). In some cases, the conclusion was that such an approach is not satisfactory, because in transportation problems, multiple optimal solutions are likely to occur (Barr and Smillie, 1972).

Rietveld (1980) provides a good description of a MOP solution to a transportation network problem. The decision problem to be dealt with concerned the determination of the optimal pattern of home-to-work trips in a certain region. In the region, several locations can be distinguished with unbalanced local labor markets. With respect to the resulting commuting pattern, the following questions were answered:

- In which locations will the labor force working in a certain location find employment?
- By means of which modes of transport will commuting take place?
- Along which routes will commuting take place?

The alternative commuting patterns were evaluated according to three criteria:

- total private transport cost measured in monetary terms;
- total damage to the urban environment measured in appropriate units of urban environmental quality;
- total damage to the natural environment measured in appropriate units of natural environmental quality.

These objectives, which must be minimized, are of a composite nature. The first objective, for example, reflects several aspects of trips, such as costs, duration, and comfort.

Rietveld (1980) tackled the above questions by means of a linear MOP method. By formulating the problem in this manner, Rietveld was able to:

- analyze possible conflicts and compromises between goals;
- calculate the values of the objectives;
- ascertain the intensity of traffic on the various routes included in the model;
- ascertain the unemployment arising in the various locations;
- calculate the probability that the pertaining solution is an optimal compromise.

MOP is appropriate for dealing with decisions that involve several groups or persons with different interests. For example, it is possible to analyze the probability of the formation of coalitions between groups of stakeholders.

4.2. The Maastricht-Aachen Airport expansion

In this section we will illustrate the application of regime analysis by evaluating the development plans for the airport expansion in the Maastricht area in the southern part of the Netherlands (Vreeker et al., 2001).

Regime analysis is a discrete multicriteria method; in particular, it is a generalized form of concordance analysis, based in essence on a generalization of pairwise comparison methods.

In 1945, the US liberation forces constructed a military airport in the Maastricht area in the southern part of the Netherlands. In the subsequent decades this small military airport became a regional airport, which is now known as the Maastricht-Aachen Airport (MAA).

Since the landing strip at the moment is only open for a limited period during the day, aircraft movements are limited. This has led to a diminishing comparative advantage and a feeble position in the airfreight market with respect to other European regional airports.

To cope with these problems, four alternative development scenarios will be considered. The four scenarios differ very much in nature and aims. The first scenario, “business as usual” (A), is designed as a reference scenario, and refers to the situation where no changes in current trends and policies occur. The second scenario, “MAA serving as a passenger airport” (B), aims to analyze the impacts of changes in the logistics of the airport. The “MAA serving the ‘Euregio’” (C) scenario aims to evaluate the impacts of land use changes in the region, and consists of the construction of a new runway. The last scenario, “tradable permits” (D), is dominated by changes in the institutional setting of the airport. In this scenario, changes occur in the legislation regarding the emission of carbon dioxide in the Netherlands. This scenario differs in nature somewhat from the previous ones, but has been mentioned several times in recent discussions. It can also be applied in combination with the three preceding ones.

The main attention in our case study will be devoted to the attainment of the best possible solution, namely one that is functionally effective and at the same time compatible with the constraints imposed by relevant environmental and social circumstances.

The various data referring to the sub-criteria in our case study are expressed and measured on an ordinal scale, through which numerical values on a nine-point scale (where the highest value represents the best score) were assigned to the various effects. The consequences that each distinct alternative has can now summarized by an impact matrix (Table 1). In this the pre-defined criteria are linked to the alternatives by means of the values that each alternative scores on the pre-defined criteria.

The regime method allows us to analyze an impact matrix containing mixed data and a weight vector in order to calculate a rank order of alternatives. The

Table 1
The impact matrix for alternative airport expansion plans

	Criterion	A	B	C	D
Economic	Economic benefits for the region	4	8	9	6
	Employment in sector transport and logistics	4	5	8	6
	Employment in sector finance and business to business	5	9	7	6
	Employment in sector tourism and recreation	6	5	1	3
	Development and supply of industrial sites	5	3	1	6
	Infrastructure	5	8	9	6
	Business traffic	6	9	8	6
	Supply of skilled jobs	4	7	9	6
Social	Nuisance	5	8	5	3
	Safety	5	7	2	3
	Health	5	6	5	3
	Recreational traffic	6	9	9	7
	Total income	4	6	7	6
	Residential areas	5	5	2	5
Environment	Nature conservation areas	5	4	1	3
	Disturbance of fauna habitat	5	8	1	3
	Air quality	4	6	1	4
	Water quality	4	5	2	4
	Soil quality	4	3	2	4
	Biodiversity	4	5	2	4

weights may be assumed to be equal, but alternative weight compositions can also be handled by means of a sensitivity analysis.

In the study research, the regime analysis was conducted in two steps. First, a regime analysis was performed on each of the main classes (economic, social, and environment). Using the values of each alternative scores on the relevant sub-criteria, the scores for each main class were determined. These outcomes are presented as the intermediate results in Figure 3. In the second step the intermediate results formed the input, together with a uniform weight vector, for a final regime analysis.

The intermediate results show that alternatives B and C have the highest scores on the economic indicators. Alternative B scores also very well on the social and environmental indicators. It will be no surprise that scenario B is the most favored one in the final regime analysis:

(1) scenario B – passenger airport;
(2) scenario C – MAA to serve the "Euregio";

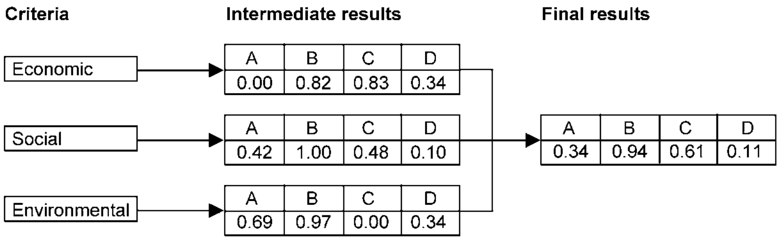


Figure 3. Rank order of alternatives (weights of criteria are equal).

(3) scenario A – business as usual;
(4) scenario D – tradable permits.

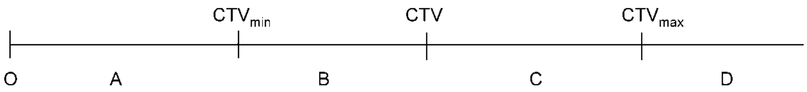
The results indicate that the MAA should concentrate on the passenger market as a source of profit. Although freight handling will still be present at the MAA, it is not recognized as the "core business." In this scenario, a new runway (the east–west strip) is constructed. To accommodate larger aircraft at the MAA. In addition, the Dutch National Aviation School will leave the airport – which gives the MAA the opportunity of extending the number of passenger flights to 55 000 aircraft movements.

4.3. The Cilento and Vallo di Diano National Park

This section discusses the application of the flag model to a "real-world" case concerning the design of a new road network in the area of the Cilento and Vallo di Diano National Park in Italy (Torrieri et al., 2002). The flag model was developed to assess the degree of sustainability of policy alternatives by means of normative limits (Nijkamp, 1995; Nijkamp and Ouwersloot, 1997; Nijkamp and Vreeker, 2000). There are three important components in the model:

(1) identifying of a set of sustainability indicators;
(2) establishing a set of normative reference values;
(3) evaluating alternatives.

The inputs into the program are an impact matrix and a set of critical threshold values defined for each relevant sustainability indicator. The sustainability indicators have two formal attributes: class and type. There are usually three classes of indicators, which correspond to the main dimensions of the sustainability analysis: biophysical, social, and economic. The two types are defined as benefit indicators and cost indicators.



Section A	Green flag	No reason for specific concern
Section B	Orange flag	Be very alert
Section C	Red flag	Reverse trend
Section D	Black flag	Stop further growth

Figure 4. Critical threshold values in the flag model.

The critical threshold value represents the reference system for judging actual states or future outcomes of policies or scenario experiments. Since in certain areas and under certain circumstances, experts and decision-makers may have conflicting views on the precise level of acceptable threshold values, we estimate a bandwidth of values of the thresholds ranging from a maximum value (CTV_{max}) to a minimum value (CTV_{min}). This can be represented as shown in Figure 4.

The third component of the model, the impact assessment, provides a number of instruments for the analysis of the sustainability problem. This analysis can be carried out in two ways. The first is an inspection of a single strategy or scenario. The second is the comparison of two strategies or scenarios. In the former procedure, one decides whether the scenario is sustainable or not. In the latter case, by comparing the scenarios, one can identify which policy or scenario scores best fits the perspective of the sustainability issue.

The Cilento and Vallo di Diano National Park, located in the south of Italy in the Campania region, was established in 1991. The park, also classified in the Mediterranean Ecosystem, was included in 1997 in the UN's "Man and Biosphere" program because of its mixed historical, socio-economic, artistic, and spiritual features. The landscape is characterized by a mountainous area, flatland zones, and coasts rich in caves and inlets. It is considered to be one of the most important territories for the preservation of biological diversity and the survival of endangered species. Moreover, its cultural heritage provides insight into rural societies and their traditions.

The main objective of the infrastructure project is to improve accessibility for the communes. It can be statistically seen that there is a close correlation between the economic, social, and cultural development and the level of mobility of people and goods. From this perspective, it is possible to identify the specific goals of the project:

- to integrate of the communes inside the park, in the main valley (valley of Sele);
- to reduce emigration;
- to preserve and increase the value of the area's natural and cultural heritage;
- to protect the environment.

The need for feasible solutions within the constraints imposed by nature, geology, and hydrology has led to the identification and design of three project alternatives (A, B, and C). These three different routes are described in Table 2.

The scales of the outcomes in this matrix are partly quantitative (for the economic and accessibility classes) and partly qualitative (for the environmental class), as described above.

The next step to be taken in the flag analysis is the definition of a set of reference values or threshold values (limits, standards, and norm) to verify the impact of the three alternatives. In this context, the notion of carrying capacity is of great importance, as it indicates the maximum environmental resource usage that is still compatible with an ecologically sustainable economic development.

Table 2
Impact matrix

	Criterion	Index	A	B	C
Economic	Investment costs (–)	Billion It. lira	107	143	127
	Maintenance costs (–)	Billion It. lira/year	1	1.3	1.3
	Transport costs by C and L (–)	Billion It. lira/year	36	39	35.4
	Transport costs by C and L+ time costs (–)	Billion It. lira/year	64	71	63
Accessibility	Time needed from C and L on entire network (–)	Minute	32.2	37	32
	Time for C to access from the communes to north-west point (–)	Average minute	30	34	29.7
	Time for L to access from the communes to north-west point (–)	Average minute	32.5	38	32.2
	Time for C to access from the communes to north-east point (–)	Average minute	38.7	42	38.4
	Time for L to access from the communes to north-east point (–)	Average minute	43.8	48	43.5
	Time for C to access from the communes to Roccadaspide (–)	Average minute	24.3	29.4	24.1
	Time for L to access from the communes to Roccadaspide (–)	Average minute	26.8	32.7	26.5
Environmental	Vicinity of population centers (+)	Qualitative	3	1	2
	Possibility of accidental fall of dangerous material (+)	Qualitative	2	3	3
	Landslide risk (+)	Qualitative	3	1	2
	Hydrology risk (+)	Qualitative	2	3	3
	Loss of vegetation (+)	Qualitative	2	1	2
	Alteration of fauna habitat (+)	Qualitative	2	1	2
	Violation of regulations on natural environment (+)	Qualitative	3	1	1
	Fitting into the landscape (+)	Qualitative	2	3	1
	Change in landscape morphology (+)	Qualitative	2	3	2

Table 3
CTV values for the selected indicators

Indicator	CTV_{min}	CTV	CTV_{max}
Investment cost (billions It. lira)	85	100	150
Maintenance cost (billions It. lira)		1.5	
Transport cost C and L (billions It. lira)		33	
Transport + time costs C and L (billions It. lira)		66	
Internal accessibility (min)	15	20	30
External accessibility (min)	30	45	60

This means that this concept refers to a threshold value that cannot be exceeded without causing unacceptably high damage and risk to the environment. Clearly, for each of the sustainable indicators, whether environmental or socio-economic, a CTV has to be specified such that an entire set of CTVs may act as a reference system for judging the actual states or future results (Table 3).

Regarding the environmental criteria, as qualitative representations of the scores, we assume that the value 2 in the predefined ordinal scale (1, 2, 3) represents the minimum allowable value a environmental indicator may attain, beyond which a disturbing development could take place.

Figure 5 shows that alternative A is the most sustainable. In fact, it has ten green flags in the total scores, determined mainly by the biophysical indicators, as was expected. Upon examining the economic criterion, all alternatives appear to be unsustainable, while for the social criterion (accessibility), alternatives A and C are more sustainable than alternative B.

5. Conclusions

Whereas CBA is linked to the principles of welfare economics, this is not the case with MCA. Although the basic principles of welfare economics and CBA are straightforward, there are several important difficulties in the application of CBA. These include second-best conditions, difficulties related to the discounting principle, equity concerns, and non-priced effects. This has led to the development of alternative evaluation techniques such as MCA.

Table 4 summarizes the differences between CBA and MCA. The first difference between the two approaches is the decision rule used in the evaluation process. CBA uses prices to make efficiency attributes compatible, whereas MCA is characterized by a weighting system, implicitly or explicitly involving relative priorities of decision-makers.

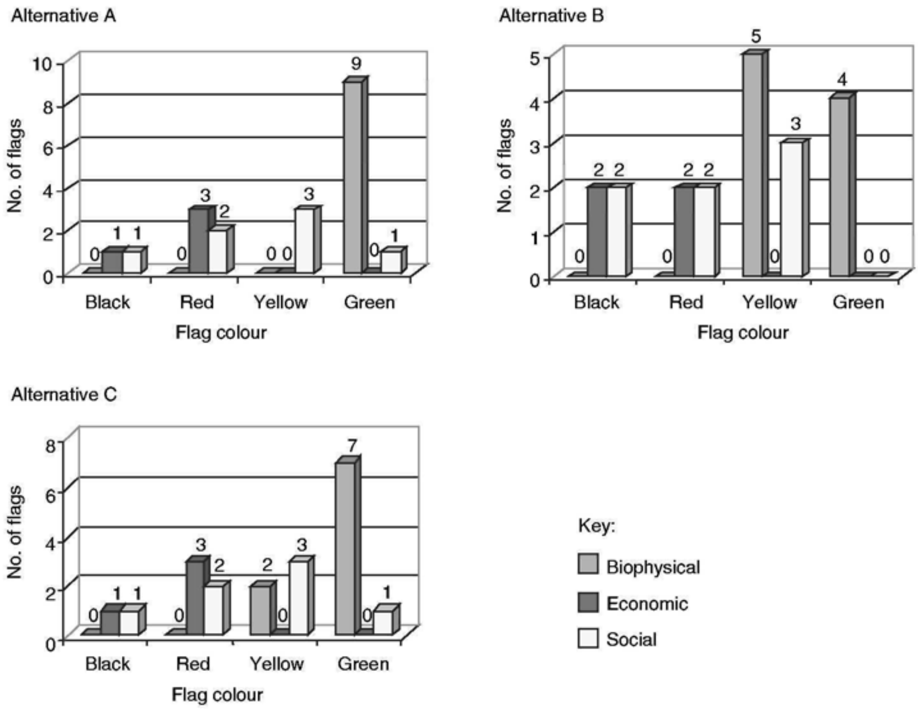


Figure 5. Alternatives A, B, and C.

Table 4
Differences between CBA and MCA

Criterion	CBA	MCA
Systematic comparison of alternatives	Yes	Yes
Explicit use of weights in the analysis	Yes	Yes
Evaluation of effects	Market based	Political
Political abuse of the technique	By manipulating inputs	By manipulating inputs and weights
Degree of compensation of effects	Every negative effect can be compensated by means of a larger positive effect	Depends on the chosen multicriteria method
Risk of "double counting"	Limited, applies to indirect effects	Yes
Inclusion of weights to reflect stakeholder preferences	Not in the basic form of CBA	Yes

The second difference concerns the criteria used in the evaluation. In contrast to CBA, which is focused on efficiency and therefore on internal rates of return, benefit–cost ratios or net present value, MCA does not impose any limits on the number and nature of criteria. However, this poses a serious risk of "double counting" effects.

Furthermore, to apply CBA, prices need to be known. CBA, therefore, requires that effects on efficiency attributes are measured in quantitative terms. MCA does not impose such requirements. There are three groups of MCA techniques with respect to effects: one that requires quantitative data, a second that processes only qualitative data, and a third that can deal with both types of effect simultaneously.

Although CBA and MCA differ in many aspects, and most debate has been on which method is most suitable in project evaluation, current thought regards these two approaches as complementary analytical tools rather than competitive methods. A combined application might, for example, include a CBA to calculate the efficiency effects of alternatives. The results of this exercise (e.g. benefit–cost ratios or net present values) can form an input in an MCA, in which, in addition to the efficiency effects, both the equity consequences as well as unpriced effects are also taken into consideration. Seen from this perspective, a mixture of monetary and non-monetary methods may be a fruitful way to proceed in evaluation research and processes.

It should be stressed that both types of method can be useful as tools in decision-making processes, as long as policy-makers and evaluators are aware of the limitations of the methods.

References

Ballestero, E. and C. Romero (1998) "Uncertainty and the evaluation of public investment decisions," *American Economic Review*, 60:364–378.

Banister, D., S. Watson, and C. Wood (1997) "Sustainable cities: transport, energy and urban form," *Environment and Planning B*, 24:125–143.

Barr, B. and K. Smillie (1972) "Some spatial interpretations of alternative optimal and sub-optimal solutions for the transportation problem," *Canadian Geographer*, 16:356–364.

European Investment Bank (1996) *The trans European networks for transport and energy infrastructure in the 21st century.* Luxembourg: EID/EC.

Handy, S. and P.L. Mokhtarian (1995) "Planning for telecommuting: measurement and policy issues," *Journal of the American Planning Association*, 61:99–111.

Janssen, R. and G. Munda (1999) "Multi-criteria methods for quantitative, qualitative and fuzzy evaluation problems," in: J.C.J.M. Van den Bergh, ed., *Handbook of environmental and resource economics*. Aldershot: Edward Elgar.

Nijkamp, P. (1975) "Reflections on gravity and entropy models," *Regional Science and Urban Economics*, 5:205–255.

Nijkamp, P. (1995) *Transport and sustainable regional development. Serie research memoranda*. Amsterdam: Vrije Universiteit.

Nijkamp, P. and H. Ouwersloot (1997) "A decision support system in regional sustainable development," in: J.C.J.M. van den Bergh and M.W. Hofkes, eds, *The flag model, theory and implementation of sustainable development modelling*. Dordrecht: Kluwer Academic.

Nijkamp, P. and R. Vreeker (2000) "Sustainability assessment of development scenarios: methodology and application to Thailand," *Ecological Economics*, 33:7–27.
Rietveld, P. (1980) *Multiple objective decision methods and regional planning*. Amsterdam: North-Holland.
Roy, B. (1968) "Classement et choix en presence de points de vue multiple (la methode ELECTRE)," *RIRO*, 2:57–75.
Roy, B. (1972) "Decision avec criteres multiple," *Metra*, 11:121–151.
Roy, B. and E. Jacquet-Lagreze (1977) "Concepts and methods used in multicriterion decision models: their applications to transportation problems," in: H. Strobel, R. Genser and M.M. Etschmaier, eds, *Optimization applied to transportation systems*, Laxenburg: International Institute for Applied Systems Analysis.
Torrieri, F., P. Nijkamp, and R. Vreeker (2002) "A decision support system for assessing alternative projects for the design of a new road network," *International Journal of Management and Decision-making*, 3:114–136.
Vincke, P. (1992) *Multicriteria decision-aid*. New York: Wiley.
Vreeker, R., P.T. Nijkamp, and C. Ter Welle (2001) "A multicriteria decision support methodology for evaluation airport expansion plans," *Transportation Research Part D*, 7:27–47.
Zeleny, M. (1982) *Multiple criteria decision-making*. New York: McGraw-Hill.

Chapter 31

NEGOTIATED AND COMPETITIVELY TENDERED PERFORMANCE-BASED CONTRACTS

ERNE HOUGHTON and DAVID A. HENSHER
University of Sydney

1. Introduction

The ongoing institutional reform of the transport sector has led to the accumulation of a great deal of experience on the performance of ownership and competitive structures implemented across different modes. A central feature of many of the reforms has been the specification of a contract regime for both procurement and payment established under competitive regulation. The most widely used instruments for achieving this have been competitive tendering for services and competitive franchising for infrastructure and operations. Preston (see Chapter 5) provides an overview of the relevant literature and some evidence (see Preston 2001, 2003). In this chapter we focus on competitive tendering with particular reference to public transport; although the experiences have implications for the entire transport sector.

With the accumulation of evidence on the success of initiatives relating to contracting of transport infrastructure and operations functions, there arises the question as to what refinements might provide better outcomes. Competitive tendering (CT) has shown itself to be a relatively popular instrument for change.[a] As time passes, however, a number of deficiencies in the existing CT processes have emerged, raising questions about where this approach is most suitable and the best ways of applying it. Some of these deficiencies are attributable in part to the inadequacy of the regulatory framework within which CT is delivered and monitored and some are due to the nature of CT. What we see in particular is that competitive tendering of a large public sector provider delivers an immediate cost saving but it is typically a once-only gain. There is also evidence of significant savings from corporatization of previous public monopolies (e.g. Toronto, Dublin, Sydney, and Melbourne – Stanley and Hensher, 2003), suggesting that CT may be a sufficient but perhaps not a necessary condition for delivering such savings. The

[a]For example, competitive tendering is proposed as an instrument to make radical change in service delivery in Santiago, Chile, to replace 4000 bus operators (with 8000 buses) with 15 operators.

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

evidence, though, would tend to support the view that larger and earlier savings generally result from CT and other forms of direct competition. Subsequent re-tendering appears to deliver very little gain in a financial sense, and in situations where a large number of small operators in the informal transport sector, as in Brazil, are replaced by a few larger operators, the costs of service delivery under CT can increase. The tendency for the numbers of bidders for re-tenders to decrease in some countries suggests that this issue of the sustainability of initial cost savings may become widespread.

Such issues suggest a reconsideration of competitive tendering as the preferred way of contracting under all circumstances and a need to consider negotiated performance-based contracts as an alternative (and/or sequenced complement)[a] to CT as a means of deciding rights to deliver public transport services. Negotiated contracts are common in public–private partnerships in the infrastructure area but are much less so in public transport operation. Berechman (1993) suggests that

> If costs of having a private firm supply the services could be reduced by means of a negotiated contract, the considerable costs of organizing a competitive bidding would be averted. Indeed ... a competitive tendering scheme might in some cases be inferior to methods of contract renewal or negotiation.

This chapter is premised on the assumption that all rights to provide public transport service expressed through contracts should be dependent on the performance of the provider, and that this should be expressed through a performance-based contract, be it competitively tendered or negotiated. We begin with a definition of the dimensions of performance-based contracts, including issues that need to be dealt with in developing contracts that most effectively meet a government's objectives in public transport service provision. In particular, discussion is focused on the relative merits of negotiated contracts, compared with competitively tendered contracts, in delivering value for money outcomes. We use examples of practice from a number of countries as illustrations, including a detailed case study using the methods developed in Hensher and Houghton (2003a,b).

2. Defining performance-based contracts

Performance-based contracts (PBCs) as a broad class of contracts offer a framework within which the government (S: strategic), the regulator (T: tactical),

[a]In South Africa, CTs are a way to attract new entrants into the market. Subsequently, based on performance, an extension is negotiated. To attract new entrants, a minimum percentage of subcontracting is stipulated, so that after 1 year of subcontracting, the subcontractor can become a "set aside" and can operate in their own right as a fully fledged operator.

the operator (O),[a] and society at large can participate as trusting partners in securing value for money in the allocation of the total subsidy budget or (2) in the delivery of non-subsidized services. Within such a contractual regime an operator provides services (be it designed at the T level or integrated at the O level) at best practice cost levels for a given level of service delivery either:

- in return for direct financial support from the government (i.e. a social subsidy that may be awarded by either CT or negotiation), or
- in return for permission to operate a negotiated/agreed level of service (without subsidy but, for example, subject to a cost-plus fare determination).

Social support can be defined in a number of ways:

- a fixed payment (e.g. a community service obligation (CSO) payment linked to a minimum service level (MSL) program determined by negotiation or CT, or a partnered service design and level), or
- a set of incentive payments above the fixed payment linked to patronage and/or service levels (vehicle-kilometers, frequency by time of day, etc.).

The incentive payments linked to patronage growth may reflect benefits derived from all sources (i.e. consumer or user surplus) and additional benefits specific to reducing negative environmental impacts. Those linked to service levels may incorporate a mechanism for supporting new entrants in developing markets. Performance incentive payments (PIPs) can be based on various criteria, e.g. passenger boarding and passenger-kilometers to account for the trip length distribution as well as the actual number of passengers, which lowers the risk to the operator of patronage factors that the operator cannot control. Figure 1 shows the contractual components that bind the STO entities. Although the maximum fare is on the *laisser-faire* side of regulatory processes, while social support presents many contract specification challenges to effectively promote goals consistent with strategic objectives, all contract components can apply to all contract types. There are many forms of contract, each specifying a particular combination of components and corresponding measurement methods (for an overview, see Wallis, 2003).

The payment rates associated with the form of contract establish the service delivery cost to the government of achieving the strategic goals, which are usually expressed as dollar benefits to passengers and to other road users. Benefit rates can then be compared with the shadow prices of government funds. Payment rates may be determined by two major processes: CT, or negotiation. CT-linked and -negotiated PBCs can contribute in various ways to both the choice of operators

[a]See Chapter 6 in this volume for a discussion of the STO framework. The STO framework recognizes that policy, planning, and operations exist within a hierarchy of objectives functionally split into three interdependent entities.

(i.e. the sourcing of operators) to provide service and to specification of contract payments (i.e. how contracts are structured). Negotiated PBCs facilitate, in particular, system-wide determination of value for money. CT and negotiated PBCs can be complementary in a temporal sequence (Figure 2). For example, one can use a service incentive payment under a negotiated PBC to assist new entrants into new markets (including a base CSO) including training/skill enhancement support. When a market is established (given sufficient elapsed time, e.g. 5–10 years) one might introduce CT to rationalize the number of "competing" operators in a corridor (as is proposed for Santiago, Chile) or select an individual operator at a route or corridor or area level, or move to a negotiated PBC system. Alternatively, a government might use CT to short-list a number of suppliers with whom it then negotiates to select the preferred supplier.

Incentive payments can be introduced through CT or negotiated PBCs. For example, one can establish a PIP of the following possible types:

- the Adelaide model (agreed non-competitive sum per additional passenger);
- the Hensher–Houghton model – with a fixed or variable PIP budget competed for amongst a predefined set of operating areas, which can be referred to as competition at the later service delivery stage, as distinct from at the tendering stage.

Given that patronage is often not strongly influenced by operator initiatives so much as by exogenous effects, a service incentive payment (SIP) may also be warranted (which in the case of the Adelaide model requires a tactical level sign off on service design).This SIP may be a marginal payment rate (as in the Adelaide model) or a be an amount competed for by operators who grow service off a base payment linked to agreed MSLs. The introduction of a service incentive payment, where one does not compete for subsidy budget between operators in different spatial settings, is an appealing model for South Africa and also Brazil (the Brazilian model is shown on the right-hand side of Figure 2 by the thicker line only).

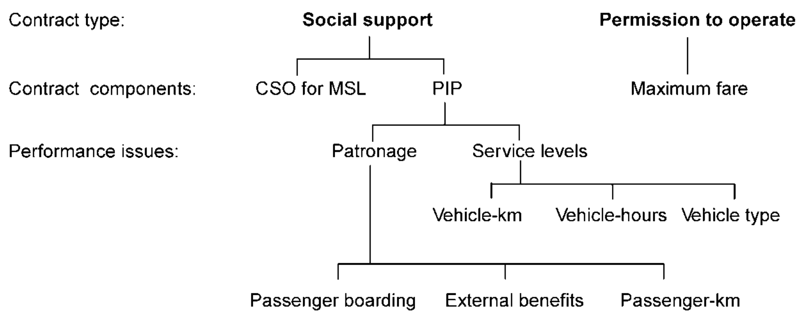
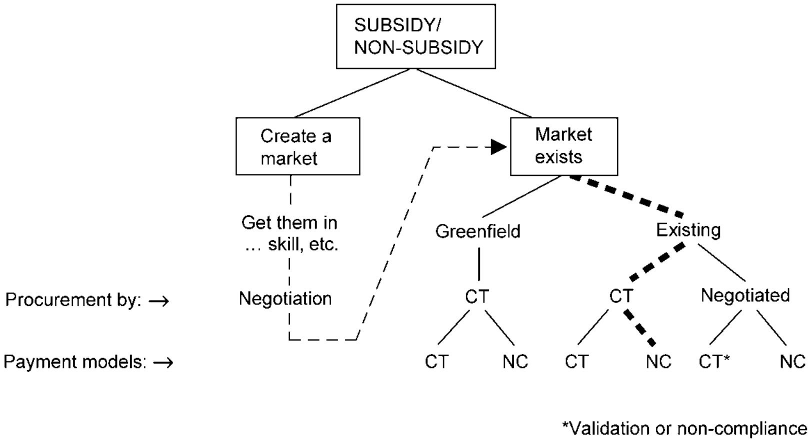


Figure 1. Contract types and components (Hensher and Houghton, 2003b).



Notes: (1) A greenfield site is different to "creating a market." The latter is more global in its national context, and refers to a general absence of expertise that can readily participate in the market, be it an area already serviced or a new development with no services (i.e. a greenfield site). (2) The block under the greenfield site could also be negotiated. For example, in South Africa (e.g. Durban) an expression of interest for new services that is not subject to CT is common

Figure 2. Processes for procurement and payment rates determination.

3. Examples of the implementation of PBCs: the Hordaland (Norway) and New Zealand models

Norway and New Zealand provide examples of how performance-based approaches to public transport delivery can be structured. The aim in both cases is to give greater effect to the economic rationale for service subsidy, bringing operations more into line with social surplus considerations. Benefits to existing and new users from service improvements are rewarded. There is also a prospective reward for reducing external costs.

3.1. The Hordaland model

Hordaland[a] is one of three areas in Norway where performance contracts (called "quality contracts") have been implemented for public transport service provision. The contracts start from the premise that the operator usually has the best

[a]Hordaland is a county in western Norway, and includes the city of Bergen. The total population is 450 000. There are three major operators delivering in 1999 about 24 million revenue-kilometers per annum and carrying 35 million passengers per annum. The total annual deficit of the operators is 170 million Norwegian kroner.

knowledge of the market and should be left to design the most appropriate route system. For this system to be designed and operated effectively from a social perspective, however, proper incentives need to be present. The contracts recognize that a profit maximizing operator, in making decisions about service changes, will normally only consider the direct marginal implications for costs and fare revenue. This misses two important elements from a social surplus perspective:

- the benefits to existing public transport users from an improved service level (these are essentially an externality to the operator);
- the benefits from reducing car use, when that use is "underpriced" in terms of its marginal social costs.

The Hordaland framework seeks to internalize these benefits within an operator remuneration framework that is related to the level of service and to passenger numbers.[a]

The key principles in the performance-based contract introduced in year 2000 are, first, that the operator is given financial incentives for product development. Secondly, the authorities define a framework comprising overall quality requirements regarding price, service, and accessibility. The county may cancel the contract if the operator fails to fulfil the predetermined criteria. Joint cooperation is required for the contract to be fulfilled, and the authorities are obliged to enforce measures to improve the effectiveness of the public transport system (e.g. with respect to matters such as bus priority treatment).

Public transport, considered as a quasi-public good, requires incentives additional to those from the market-place to avoid a level of production lower than what is economically effective. Such incentives apply in the Norwegian approach for minimum kilometers, e.g. with regard to school buses and other socially necessary services, although this may be granted as a fixed subsidy. More importantly, it applies for increasing frequency and vehicle-kilometers, which implies gains for existing passengers as well as attracting new passengers (modal shift). This is especially valid for peak hour passengers, when the marginal costs of extra departures are high.

Larsen (2001) and Johansen et al. (2001) present the modeling on which the Hordaland contract remuneration system is based. Fare levels, bus revenue-kilometers and bus capacities are chosen so as to maximize a social welfare function. Fare subsidies and revenue-kilometer subsidies are then calculated so as to induce a revenue maximizing bus operator to select the socially optimal levels of revenue-kilometers and bus capacities. Fares are regulated by the county, but the total payment per passenger received by the operator is the sum of the fare and a subsidy component.

In Hordaland, the entire subsidy amount is performance based. There are specified rates for subsidies per route-kilometer, and per vehicle-hour for peak

[a] For a formal economic treatment, see Johansen et al. (2001).

hours and off-peak.[a] An additional amount per passenger in peak hours was suggested but not implemented. These rates vary among operators, depending on the proportion of urban versus rural kilometers. In principle there is no upper boundary for any of the given subsidy components, but due to budgetary constraints in the county there is a ceiling for the total amount granted. The authorities define a framework for the minimum quality of service, with regard to fares and accessibility. This also involves a customer satisfaction survey. If customer satisfaction falls below 90% of the target level, the authority, Hordaland County Council, can cancel the contract and select another operator.

The operators are granted a substantial degree of responsibility for planning and product development. They decide on timetables and frequencies, vehicle types, and fares, i.e. elements belonging to the tactical level, not only the operational. The authorities define certain minimum criteria, and otherwise do not intervene at the tactical level. The operators are free to establish and withdraw routes except for school buses. However, they cannot reduce the number of overall network kilometers without the prior consent of the county.

A commonly voiced argument against performance contracts is that they protect the incumbent. In Norway there have been two ways to handle this. First, a "threat of competition" has been included in the contract: if service quality drops below a specified level (e.g. a customer satisfaction index), the authorities may tender the contract. Secondly, as is the case in Grenland (a city in Telemark County), the performance contract itself will be tendered. This competition will not follow the "lowest bid wins" principle. Instead, a fixed amount of subsidy will be offered, and a multi-criteria method will be developed to select the operator offering the best service (i.e. in terms of delivered quality, frequency, vehicles, etc).[b]

3.2. The New Zealand model

Transfund New Zealand has developed a patronage funding policy for public transport that provides direct incentives for patronage growth. Central government public transport funding goes directly to the New Zealand regions (not to operators

[a]Specifically, in the Hordaland model, vehicle-kilometers, vehicle-hours, and passenger trips (differentiated between peak hours and normal hours) are all part of the subsidy calculation and are not related to the MSL as such. The total level of subsidy must at least allow for the fulfillment of the MSL obligations. The vehicle-kilometer-based subsidy primarily reflects user benefits of increased frequency rather than MSL. In the Norwegian model, the MSL is given as a contractual obligation, and technically speaking the subsidy per vehicle-kilometer is not offered for the MSL as such. Rather, as in the Oslo model, additional vehicle-kilometers and additional passenger trips exceeding a base level are compensated for.

[b]Berge et al. (2003) provide an update on the performance of these contracts.

direct) is based on matching base funding levels that existed in 1999–2000, "kick-start" funding of a share of the costs of approved new services and initiatives, and a patronage incentive.

The patronage incentive is based on the same two components included in the Hordaland model, user benefits and externality benefits of improving services and gaining new passengers. Wallis and Gale (2001) report that the externality component includes estimates of benefits from reduced road congestion, plus an allowance for safety and environmental benefits. As a consequence, the payments vary by city, time period, and distance traveled. Thus, for example, payments are higher for peak period patronage increases in more congested cities than for off-peak patronage increases in cities with little traffic congestion. The approach is unambiguously intended to direct funding toward locations where public transport improvements can make a difference in reducing road congestion.

To seek some cross-sectoral parity with funding of road improvements, a shadow price (or "hurdle rate") of funds is introduced into the funding formula, such that only public transport projects that achieve a marginal benefit–cost ratio (BCR) similar to, or better than, that of marginal road projects which receive funding will be supported. The shadow price is introduced as the value of the marginal BCR for funded road projects, this being used as a divisor of the public transport benefit measure (user benefits plus external benefits from public transport improvements).

The values of the externality benefits in the New Zealand work are presented by Wallis and Gale: environmental and safety benefits are typically in the range of NZ. \$0.08–0.13 per marginal passenger-kilometer, across all centers and peak/off-peak, and congestion benefits vary by city, reflecting congestion levels, and are only significant at peak periods. Values were in the NZ. \$0.40–0.50/diverted passenger-kilometer for the peak in the largest two cities of Auckland and Wellington.

The benefits to existing public transport travelers from service improvements were expressed as a function of the generalized cost (per passenger) of travel by public transport, which was assessed as

$$G = \$2.65 + \$0.48 \times \text{trip length}.$$

Dividing this expression by the elasticity of demand with respect to this generalized cost produced the relevant benefit estimate. Elasticity values were put at –1.0 for peak periods and –1.5 for off-peak.

3.3. Assessment

The Hordaland and New Zealand approaches provide a start toward the development of a PBC remuneration system that reflects the service delivery goals of the governments that are providing service funding support. In particular, they direct attention to a support framework in which the social goals of generating benefits to existing public transport users and reducing the external costs of road use are embodied. The New Zealand approach also includes external benefits from service improvements, but it does not deal with the question of defining a minimum public transport service level that might be required in recognition of the CSO function of public transport.

The framework developed in Hensher and Houghton (2003a, 2004) for Australia integrates all the ideas from Hordaland and New Zealand (with some variations) to optimize service delivery and subsidy provision across the CSO component, the additional benefit items (user benefits plus external benefits) and operator returns.

4. A case study: PBCs in Sydney

A mathematical programming model to assess PBC design and implementation in the bus industry was presented in Hensher and Houghton (2004). The model was developed from operations data for a benchmark operator, and an abbreviated version is given below. It is then applied to a case study in Sydney, for which the data have been massaged to preserve operator confidentiality. To simplify the presentation, both the model and results refer to only a single passenger class and constant average travel demand over daily operations.

4.1. Model summary

The model comprises demand and cost functions, operational constraints, and a social surplus objective function, which are summarized below.

Annual passenger demand

The demand for bus travel (Y) is defined as one-way annual passenger trips per contract period. It is assumed to be influenced by service levels (X), where the latter is proxied by revenue vehicle-kilometers (RVKM). Fares are assumed to be fixed. Before the implementation of the proposed scheme (base case B), demand levels, Y^{B}, are based on existing service levels. After the implementation of the proposed scheme (application case A), predicted demand (Y^{A}), is given in equation (1) as a function of a base demand (Y^{B}), the direct RVKM elasticity of demand, and operator responses to the scheme through changes to RVKMs:

$$Y^{\mathrm{A}} = Y^{\mathrm{B}} \exp\left(\frac{\varepsilon_Y^X}{X^{\mathrm{B}}}(X^{\mathrm{A}} - X^{\mathrm{B}})\right). \tag{1}$$

An equivalent prediction of patronage associated with MSLs, Y^{MSL}, is based on X^{MSL} in place of X^{A}.

Annual total cost

Total predicted cost (C^{A}) is defined in line with the format of equation (1), as a function of benchmarked base cost (C^{B}), predicted responses in total vehicle-kilometers (VKM), predicted changes in total passenger demand, predicted responses in the number of buses assigned (#bus), and the corresponding set of cost elasticities:

$$C^{\mathrm{A}} = C^{\mathrm{B}} \exp\left(\frac{\varepsilon_Y^X}{\mathrm{VKM}^{\mathrm{B}}}(\mathrm{VKM}^{\mathrm{A}} - \mathrm{VKM}^{\mathrm{B}}) + \frac{\varepsilon_C^Y}{Y_{\mathrm{S}}^{\mathrm{B}}}(Y^{\mathrm{A}} - Y^{\mathrm{B}}) + \frac{\varepsilon_C^{\#\mathrm{bus}}}{\#\mathrm{bus}^{\mathrm{B}}}(\#\mathrm{bus}^{\mathrm{A}} - \#\mathrm{bus}^{\mathrm{B}})\right). \tag{2}$$

Vehicle-kilometers and MSLs

The MSL that must be provided as a condition of public transport service delivery is defined below for the purpose of the empirical study as 67% of current service VKMs:

$$X^{\mathrm{MSL}} = (0.67)\mathrm{VKM}(1 - R) \qquad X^{\mathrm{A}} \geq X^{\mathrm{MSL}}. \tag{3}$$

The inclusion of R enables us to assess the implications of various mixes of MSL and PBC service levels.

Traffic and capacity

Capacity per RVKM required in the base period is defined by

$$Z^{\mathrm{B}} = \frac{(\#\mathrm{pers} \times \#\mathrm{bus})^{\mathrm{B}}}{X^{\mathrm{B}}}, \tag{4}$$

where #pers is the bus capacity (seating plus standing). An increase in base traffic results in a reduction in X^{B} and an increase in Z, which has the effect of increasing the capacity required, $X^{\mathrm{A}}Z^{\mathrm{B}}$, for a given solution X^{A}. $Z^{\mathrm{A}} = Z^{\mathrm{B}}$ is used in

$$X^{\mathrm{A}} \leq \frac{(\#\mathrm{pers} \times \#\mathrm{bus})^{\mathrm{A}}}{Z^{\mathrm{B}}} \tag{5}$$

to impose equivalent traffic conditions in equivalent periods to the base case, and the inequality gives an upper bound on X^{A} that is fixed by the number of buses assigned. For a given #bus, the bound may be loosened by reducing service quality.

Service quality

The fundamental form of the service quality constraint is $Y^{A}/X^{A} \leq Y^{B}/X^{B}$. This is eased by a control variable, κ, initialized by

$$\kappa = \frac{Y^{B}}{X^{B}Z^{B}}, \tag{6}$$

which measures the base trip-rate per unit carrying capacity allocated, and the modified constraint is

$$Y^{A} \leq \kappa Z^{A}X^{A}. \tag{7}$$

The control variable relates to how full the buses are on average, given normal operating practices, providing a measure of service quality with respect to loading that is able to be adjusted up or down.

Total subsidy cap

PIPs are made on the value of patronage growth above Y_{MSL}, where the value is defined by the sum of the consumer surplus and external benefits from car transfer that result from patronage growth above Y_{MSL}, indicated by $(\mathrm{CS} + \mathrm{EB})^{+}$. PIPs result from a pay-out rate (P) applied to the value of patronage growth above MSL. The amount of subsidy available for PIPs cannot exceed the total allocated subsidy budget net of commitments to CSO payments on MSL operations. These requirements are expressed in

$$P(\mathrm{CS} + \mathrm{EB})^{+} \leq \mathrm{TB} - \mathrm{CSO}(1 - R) \qquad \text{for } (\mathrm{CS} + \mathrm{EB})^{+} > 0. \tag{8}$$

Objective function

The objective of the model is to maximize the value of net social surplus, given by the sum of community benefits and producer surplus net of the subsidy budget, as given by

$$\begin{aligned}\text{Max: } &(\mathrm{CS} + \mathrm{EB}) + P(\mathrm{CS} + \mathrm{EB})^{+} + qY^{A}_{\mathrm{ACC}} - C^{A} + \mathrm{CSO}(1 - R) - \\ &\qquad [\mathrm{CSO}(1 - R) + P(\mathrm{CS} + \mathrm{EB})],\end{aligned} \tag{9}$$

which reduces to

$$\text{Max: } [(\mathrm{CS} + \mathrm{EB}) + qY^{A}_{\mathrm{ACC}} - C^{A}]. \tag{10}$$

4.2. The case study

In this case study, the operator's business and community outcomes are the result of decisions on VKM, which generate patronage through the demand function (3), and, therefore, determine all fare revenues as well as costs. Fare levels are assumed fixed at current values. The operator is currently operating at a patronage level 11% above the MSL, and this existing growth is to be rewarded along with new growth under any regulator scheme. The primary business outcome is given by the producer surplus as a percentage of operating cost (%PS). Capital costs are included in the cost function (2), and the producer surplus is a measure of returns to entrepreneurship.

The intention of a regulator scheme in this case study is to provide an incentive to the operator to deliver patronage growth above MSL levels. Community valuation of patronage growth is at the value of the corresponding increase in community benefits, $(CS + EB)^+$. The operator's VKM decisions are implemented in the context of the regulator scheme, defined by a budget level, TB, and a pay-out rate, %P, paid on the value, $(CS + EB)^+$, of patronage growth generated by the operators VKM decisions. The total payment, $\%P(CS + EB)^+$, is referred to as the PIP. Operator decisions (VKM) determine PIPs as well as fare revenues, and in that sense translate a given scheme (%P, TB) into operator profitability (%PS), depending on the relative efficiency of the operator as defined in the model. We assume a benchmark operator, and a regulator who is well informed about the economics of benchmark operations. The current level of the subsidy budget is \$7 304 306, and the maximum budget level is \$7 669 521, on the assumption that a 5% subsidy budget increase is a politically feasible maximum.

The model is used to evaluate the potential effectiveness of alternative schemes in providing positive incentives to patronage growth through VKM decisions. Model simulations determine the budget requirements, and operator business outcomes that would result from optimum operator decisions at each design point, defined by a specified patronage growth rate to be achieved under a scheme with a specified pay-out rate.

The results are presented in Figure 3, where the vertical axis measures the subsidy budget level, TB, and the horizontal axis measures operator return %PS. The current budget line is shown as TB(CURR), and the maximum budget line as TB(MAX). The current, or base case, position is shown at point A, where the scheme is defined by the current budget level, TB = \$7 304 306 on the vertical axis, and %P = 14.54% as shown directly under point A. When this pay-out rate is applied to existing patronage growth, the resulting PIPs, when added to the CSO payment, exhaust the budget. The corresponding operator return is read off the horizontal axis as 5.4%. Point A involves zero patronage growth above existing growth. Point A is simulated by running the model with TB set at the current

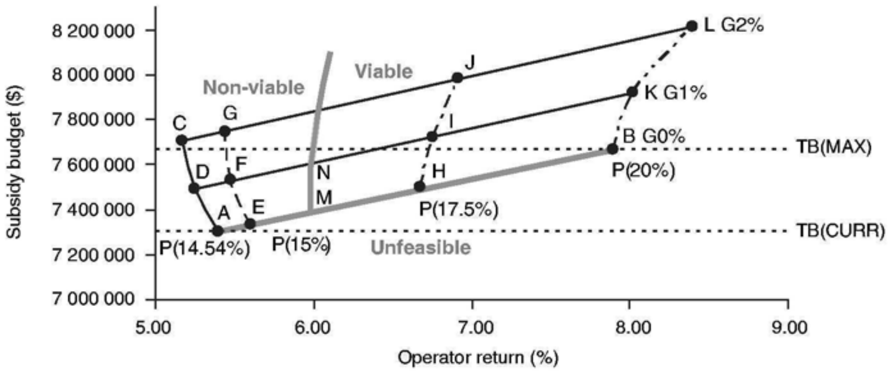


Figure 3. PBC schemes and outcomes.

budget level, and %P set at the level that exhausts the current budget when applied to existing growth.

The set of zero-growth schemes and outcomes is shown in Figure 3 as AB. For design points along AB, PIPs on existing-only growth above MSL are made to exhaust an increasing budget requirement. Pay-out rates increase from 14.54% at A, through 15% at E and 17.5% at H, up to 20% at B, where the pay-out rates are shown directly under the corresponding scheme–outcome points, as before; and we note that the maximum budget requirement is at B. With no corresponding patronage improvement, the increasing pay-out rates along AB steadily increase operator profitability as shown, from 5.40% at A, up to 7.9% at B. In the same way, the sets of 1 and 2% growth schemes and outcomes are shown as DK and CL, respectively.

The pay-out rate, %P, is constant at 14.54% along the path ADC, and constant at 15, 17.5, and 20% along the paths EFG, HIJ, and BKL, respectively. All points (%PS, TB) in the scheme–outcome space below the zero-growth line, AB, are unfeasible in the sense that no pay-out rate can give the operator return on the horizontal axis from the corresponding budget on the vertical axis. AB is seen to give a tighter lower bound on TB in (scheme, outcome) space than the current budget level, TB(CURR).

The required growth rate at each design point establishes $(CS + EB)^+$, and leaves producer surplus as the only component of social surplus to be optimized by the model. The best way to achieve the required growth in terms of social surplus outcomes is, therefore, equivalent to the best way to achieve it in terms of producer surplus outcomes. Since VKM is the single component of the operator's decision, it is clear that Figure 3 may be used to give operator returns, %PS, resulting from optimum VKM decisions made to achieve the specified growth rate, %G, under the specified scheme (%P, TB) associated with each design point.

Viable schemes

Where the pay-out rate on patronage growth and the budget, are increased, the regulator's intention is that the operator will pursue increasing PIPs by increasing VKM (through extra service runs and network extensions), thereby increasing patronage through equation (3). The effect of PIPs is to increase the effective marginal revenue associated with a decision to increase VKM. Other implications of the VKM decision are that greater fare revenue will be generated, and marginal costs will increase as operations are extended beyond currently optimum bounds. The viability of the scheme, therefore, depends on the relativities of the changes to marginal revenue and marginal cost that are associated with increased VKM. In this model, an improved revenue–cost environment created by a scheme is reflected in a higher return to entrepreneurship in the simulated model responses. The incentive offered by alternative design points is, therefore, assessed through %PS.

A viable PBC scheme is one that will induce patronage growth, which assumes that the rate of return to entrepreneurship, %PS, does not decline as an outcome of growth. The design points of Figure 3 identify corresponding schemes and also indicate viability. Point C, for example, identifies the scheme (14.54%, \$7 705 290), which is also recognized as not viable because the %PS is less at the 3% growth necessary to claim the budget, than at zero growth, etc. In this way, Figure 3 may be used to assess the viability of alternative schemes.

Clearly, if %P were to remain at its base case position of 14.54%, and budget increments were made available to support corresponding payments on any achieved growth, the scheme would be non-viable. Any further growth above the current (existing) level would require increased VKM to increase patronage through equation (3), taking costs above the current optimum benchmark position, with no revenue effects. In this way, growth would reduce operator returns below benchmark levels, from %PS = 5.4% at the zero-growth solution, to 5.25% for 1% growth, and 5.17% for 2% growth. Clearly none of the schemes (14.54%, TB) shown in Figure 3, would induce the operator to pursue growth. Nor would increasing the pay-out rate to 15% be sufficient to bring viability as decreasing %PS prevails along EFG. But the schemes of (17.5%,TB) are viable. The return on the zero-growth solution would become 6.68%; the scheme (17.5%, \$7 723 504) would induce the operator to increase VKM to achieve 1% growth, thereby increasing %PS to 6.75%; and the scheme (17.5%, \$7 979 089) would induce the operator to achieve 2% growth, thereby increasing %PS to 6.92%. The threshold pay-out rate for viability is %P = 16%, which would give a return of 5.98% to the zero-growth solution as shown at M, and the maximum budget scheme (16%, \$7 669 521) would induce the operator to achieve 1.30% growth, thereby increasing operator returns to 6.00% as shown at N. The threshold pay-out rate path is seen to divide the feasible design point region above AB into areas of viable

Table 1
VFM under viable schemes

	M	H	B	N
TB ($)	7 488 914	7 496 458	7 669 521	7 669 521
SS ($)	11 750 623	11 754 756	11 986 290	13 685 555
VFM (%)	156.91	156.80	156.28	178.44

and non-viable design points, where the viable design point region is given by MHBN.

Optimum schemes and scheme protocols

The social surplus outcomes corresponding to points MHBN are given in Table 1. The value for money invested in the subsidy scheme is given by VFM = SS/TB,[a] and it is clear that VFM will change little as design points are moved along MB from M. The zero growth associated with schemes along MB allocates any increase in TB to PS, and both the numerator and denominator of VFM are increased by the increase in TB. On the other hand, VFM increases rapidly as the scheme is moved along MN from M, since the increase in TB is predominantly used to induce growth and generate social surplus while the benchmark producer surplus is fairly constant. Scheme movements along NB from N give slowly falling VFM, as the (maximum) budget increase is slowly reallocated from growth inducement to improving operator productivity.

The preferred position of a regulator interested in cost effective generation of patronage growth is at the point of maximum social surplus, given by N, which also identifies the maximum VFM scheme, which is the communities' optimum scheme. An unthinking regulator might prefer other design points to N; for example, a solution on the %P = 14.54 path on TB(MAX) close to C, or, indeed, any point on the same pay-out path between the 1.3% growth point and TB(MAX). But in a "trusting partnership" the non-viable region is discarded and such schemes are considered non-viable.

It is seen in Figure 3 that an operator on any %P path within the viable region has a preference for the schemes at the top of that path, which allows a higher operator return to be generated through patronage growth than schemes identified by any lower points on the same path. These preferred schemes, along NB, show increasing %PS as B is approached from N. It is, therefore, clear that a profit-

[a]This may be expanded to

$$\mathrm{VFM} = [(\mathrm{CS} + \mathrm{EB}) + P(\mathrm{CS} + \mathrm{EB})^{+} + qY^{\mathrm{A}}_{\mathrm{ACC}} - C^{\mathrm{A}} + \mathrm{CSO}(1 - R)]/[\mathrm{CSO}(1 - R) + P(\mathrm{CS} + \mathrm{EB})].$$

maximizing operator will have a preference for point B, and will always prefer schemes with higher pay-out rates and an absence of any growth obligations. However, in a trusting partnership, the operator is expected to recognize that the solution at N maintains operator returns at benchmark levels.[a]

The typical conflict of a scheme definition process is demonstrated by the potential conflict between regulator and operator in this case study, who have preferences for opposite sides of the viable region. The scheme definition is assumed to result from the interaction of the STO partners, either through a process of CT or NC. Obviously a position between N and B is sought, and, as noted, the informed regulator will be reluctant to increase the pay-out rate above 16%, while the operator will bring pressure to raise the pay-out rate above 16%. The expectation of an informed regulator working with benchmark operators in a trusting partnership is that a final position will be reached close to N, under either negotiation or competitive tendering.

The expected scheme is, therefore, defined by TB = \$7 669 521, %P = 16%, and an expected growth rate of 1.3%. If the operator meets the expected growth rate, the reward accrues in the form of increased PIPs yielding a return to entrepreneurship of %PS = 5.99%. If the expected growth rate is under-achieved, PIPs will be lower, and %PS will be approximately the same rate applied to a lower base.

Competition and ex post *pay-out rates*

The information content of Figure 3 is limited to the interaction between the regulator and a single operator. Recognizing that other operators in other areas compete for the total subsidy budget, the outcomes associated with given schemes are not so definite as the figure implies. Although TB is fixed *ex ante*, the actual pay-out rate on actual patronage growth achieved is determined *ex post*, at the end of the scheme period, allowing over-performing operators across all regions to capture a greater proportion of the budget, and under-performing operators to capture less, on account of their performance. This *ex post* feature of pay-out rates converts the marginal revenue effect of the scheme that drives patronage growth, discussed above in the section "Viable schemes," to an expected marginal revenue effect based upon the operator's perception of their own competitive performance. Outside competition may cause the *ex post* pay-out rate to either rise or fall compared with the specified *ex ante* scheme rate.

The uncertainty of the competitive scheme will have two effects: it will encourage the operator to negotiate-up the *ex ante* %P (above 16%), and perform up to the

[a]The operator might also expect that a mistakenly implemented scheme that is later seen to be non-viable will be adjusted by the regulator to introduce viability.

growth expectation of 1.3%. It is, therefore, important that the regulator be well enough informed about benchmark operator performance to be able to provide a strong force toward optimum *ex ante* %P values, and the structure of the competitive scheme in which the benchmark operator is placed is expected to ensure patronage growth close to expected rates.

5. Ongoing issues to consider in contracting

Some specific issues associated with contracting will require consideration in most circumstances to get the most out of performance-based contracts. The level of financial incentives required to encourage operators to seek out ways to grow patronage has not been established in a way that offers broad guidelines on what actually works. Substantial differences are thought to exist between different countries and cultures based on prior expectation and overt experiences in what is an acceptable business proposition for investing in services to grow patronage. There is a risk that levels offered are way off the mark (either too generous or inadequate), although calibration and simulation provide initial and operational reality checks. Documenting what is known to date will be useful. In addition an improved knowledge of patronage growth potential will provide important data in establishing the extent to which passenger growth incentives are likely to be appropriate.

An area of variable success is the commitment of the regulator to adequate auditing and monitoring of operator performance. This budgeted item often gets short change as the budgetary cycle evolves, resulting in a service drop-off unless there is a major complaint from a passenger or politician. A much more serious commitment to monitoring is required, especially where there are inadequate incentives to deliver services through the life of a contract. Internalizing monitoring and reporting costs within a contract price has much merit. The costs of monitoring/auditing may be built-into baseline contract prices to ensure it happens. Such monitoring should provide a mechanism for developing key performance indicators on operating performance and service quality, giving all parties a rich data set for planning improvements in services. This approach should assist in ensuring that outcomes are checked against strategic objectives as well as contract compliance. It should also facilitate an open-book approach to checking benchmark costs (which may be more acceptable under negotiated PBCs than CT since it then implies a lower threat to the incumbent operator), and the regulator can source suitable evidence as widely as possible to establish confidence in the revision of benchmark costs over time.

A growing number of PBCs are defined in a multiple component form, incorporating a baseline (or minimum) level of service financed by a fixed payment, and above-baseline levels of service and patronage (marginal activity) funded by various incentive payments schemes. While this two-tier approach provides some form of security to the operator (which is greater as the fixed payment becomes a higher expected percentage of funding sources), there might be a case for a single-tiered approach in which all service and patronage levels are funded by incentive payments. This is worth investigation since it might help to establish the merits of multiple-tiered approaches. The first tier of a two-tiered approach, which is often referred to as the MSL tier, raises many questions of definition and complexity. In a trusted partnership (T–O) under PBC it is important to negotiate up front a desired minimum service profile (coverage, frequency, fares, other service quality attributes – i.e. an agreed performance assessment regime) and an agreed commensurate fixed payment (either a fixed total sum as in Adelaide or a sum per vehicle-kilometer). How a suitable T–O negotiation process may be introduced *ex ante* in a CT context is unclear.

A way of ensuring that T and O roles are defined through the outcome of a broad-based systems planning and design approach involving area agreements/ quality partnerships ("Who should design the services?") is needed. Under the STO system, there is a tendency to focus on contracting at the operational level. However, there is much opportunity and perhaps high appeal in improving the tactical tasks (especially the interface between the T and O levels) through PBCs. Examples might include putting the transport network design and implementation out to CT, especially where the interfaces between infrastructure and operations are critical to network integration; or collecting all fares by smartcard, where collection is undertaken by a bank on behalf of the government, as proposed for Chile.

The introduction of contract regimes for the provision of bus services is usually premised on a prior assumption that the size of the physical contract area is given and that any policies related to interactions between contract areas such as integrated ticketing and fares require agreement. Research is required to establish a position on appropriate contract area sizes before re-contracting, and on the benefits of service-quality-related issues such as an integrated fares policy that are assumed to be impacted by the number of contract areas. Given that a growing number of analysts (especially in Europe and Australia) are promoting the appeal of increasing physical contract area size to facilitate service-quality-related issues such as an integrated fare regime, it is timely to set out the pros and cons for such reform to ensure that they are not counter-productive to the desired outcomes of a reform process. Alternative ways of delivering cross-regional and broad based network benefits should be considered at the same time, to assess whether the perceived gains from a reduction in the number of contract areas is real or illusory (e.g. Cmabini and Filippini, 2003). If the gains in network economies are not sufficiently large to outweigh any likely loss of internal efficiency, then the case for amalgamating contract areas is weak. Where the major focus is on local service provision, opportunities to deliver appropriate cross-regional and cross network services might best be revealed and promoted by T–O partnerships.

Negotiated contracts require benchmark costs to be determined from diverse sources, including data that may be available from the current group of operators and other local and overseas operators. Accumulated data from around the world is highly productive provided proper control for different cost drivers at different sources is introduced. The uncertainty associated with a benchmark cost analysis for a given area is best represented by a cost band. The location of the contract costs within the band would then be determined through negotiation. Benchmark bands are also required under CT to avoid the risk of contrived cost statements. An "open-book" system to check costs is usually requested in negotiated contract processes, where it is less of a threat than under CT processes. Some CT processes do publish full details of tender awards.[a]

Regulatory capture is always raised when discussing partnerships across the STO supply chain, and this is often used as an argument (or perhaps an excuse) to throw water on the proposition that trusting partnerships can achieve a great deal in securing appropriate system-wide outcomes (in contrast to the more narrow focus on securing the least cost operator for a service that lacks innovation and network integrity benefits). At another level, the same argument is used to claim that CT leads to market concentration, although all systems incorporating T–O interaction are subjected to this claim. We need more evidence on the extent to which regulatory capture is a serious issue and the extent to which it may be the product of information asymmetry in favor of a specific operator. In particular, this investigation should be conducted with the objective of establishing how to make contracting work at the T and O level. Project alliancing, sharing risk and reward, and replacing the master–servant relationship with a trusting partnership should all be central issues. The challenge then would be to bring the regulatory component of STO to a commitment in favor of genuine partnerships that are free of corruption.

Regulatory challenges differ depending on whether there exists a well-defined and stable regulatory environment or a poorly defined and unstable environment. Both environments make CT and negotiated PBCs problematic; however, one thing we now know is that operator associations appear to have a growing importance in assisting governments in preparing operators for the new PBC environment, be it via CT or negotiation. This is especially urgent for situations where there are many small operators, many of whom lack experience in dealing with formal supply mechanisms (as seen in South Africa with the empowerment of operators using 16-people-capacity vehicles).

[a]For London see http://www.tfl.gov.uk/buses/cib_tender.shtml.

References

Berechman, J. (1993) *Public transit economics and deregulation policy*. Amsterdam: North Holland.

Berge, D.M., S. Brathen, O. Hauge and F. Ohr (2003) *Experiences with quality contracts in public transport in Norway*. Molde: Molde University College and Molde Research Institute.

Cmabini, C and M. Filippini (2003) "Competitive tendering and optimal size in the regional bus transportation industry. an example from Italy," *Annals of Public and Cooperative Economics*, 74:163–182.

Hensher, D.A. and E. Houghton (2003a) *Implementing performance-based quality contracts in the bus sector: growing patronage and post-transitional arrangements*. Sydney: Institute of Transport Studies, University of Sydney.

Hensher, D.A. and E. Houghton (2003b) "Performance-based contracts, workshop report," in: *8th International Conference on Competition and Ownership of Land Passenger Transport*. Rio de Janeiro.

Hensher, D.A. and E. Houghton (2004) "Performance-based contracts for the bus sector: delivering social and commercial value for money," *Transportation Research B*, 38:123–146.

Johansen, K.W., O. Larsen and B. Norheim (2001) "Towards economic efficiency in public transport," *Journal of Transport Economics and Policy*, 35:491–511.

Larsen, O.I. (2001) "Designing incentive schemes for public transport in Hordaland County, Norway," in: *7th International Conference on Competition and Ownership of Land Passenger Transport*. Molde.

Preston, J. (2001) "An overview of public transport reforms in Great Britain and forecasts for the future," *International Journal of Transport Economics*, 28:23–48.

Preston, J. (2003) "The road to Rio: a brief history of the International Conference on Competition and Ownership in Land Passenger Transport," in: *8th International Conference on Competition and Ownership in Land Passenger Transport*. Rio de Janeiro.

Stanley, J.A. and D.A. Hensher (2003) *Performance based contracts in public transportation: the Melbourne experience*. Sydney: Bus Association of Victoria and Institute of Transport Studies, University of Sydney.

Wallis, I. (2003) *Contract incentives in urban public transport – appraisal of practice and experience to date*. Wellington: Booz Allen Hamilton.

Wallis, I. and J. Gale (2001) "Economic incentives to increase public transport patronage – the theory and the practice," in: *7th International Conference on Competition and Ownership of Land Passenger Transport*. Molde.

Chapter 32

DISCRIMINANT, LOGIT, AND NEURAL NETWORK MODELS FOR MEASURING FINANCIAL FITNESS: APPLICATION TO THE US AIRLINE INDUSTRY

RICHARD D. GRITTA
University of Portland, OR

1. Introduction

An airline filing for bankruptcy protection in the USA was unheard of just 21 years ago, when Braniff Airways failed in 1982. It has now become an almost common everyday event. Since that year, there have been a total of over 140 filings in the USA under the Bankruptcy Act of 1938. Out of the 16 major US carriers that have existed since the 1960s – known as trunklines then and now called "majors" – eight have filed, some more than once. Two have disappeared forever. Braniff and Eastern declared bankruptcy and ceased operations. Pan American filed, and now operates as a regional carrier. Financially troubled Northeast, the smallest of the so-called trunklines, merged with Delta in the 1970s to stave off what would have been the only insolvency in the pre-deregulatory period. Continental and TWA filed twice, and US Airways and America West once. Northwest nearly filed in the early 1990s after recording record losses. Add to this list, the recent filing of United Air Lines, US Airways (for the second time), and the threat of American to file, and the magnitude of the problem in clearly driven home. These carriers represent over 50% of this segment of the US industry. It is doubtful that any other significant industry in the USA has had such a poor financial track record over such a long period of time.

While the events of September 11, 2001 certainly had an impact on carriers such as United and US Airways, the causes of the spate of bankruptcies run deeper. Many US carriers were in a weakened condition, even during the record profit years of the mid- and late 1990s. Before that date, US Airways and American West had threatened bankruptcy filings. In addition, one has only to recall the record losses incurred by the whole of the US air transportation industry in the early 1990s. The highly cyclical nature of the industry, the presence of relatively high fixed costs, and the intense competition resulting from deregulation have combined with the rather excessive use of debt finance to cripple many in this industry. And

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

the "boom and bust" cycle of air transport is hardly a new phenomenon – it dates back decades (Gritta et al., 1998).

It would be helpful for any interested party to have tools that would allow for the anticipation of such financial problems. That is the purpose of this treatise. It will present quantitative models that can allow industry analysts to gauge financial stress and to forecast the likelihood of bankruptcy in advance of the event. Given the difficulties faced by the carriers, and because of the critical nature of this industry to the US economy, an assessment of financial stress is important. This topic should be of interest to many groups, including government regulators, stockholders, creditors (bondholders, general creditors, lessors, etc.), unions, and local communities affected by the cuts in service.

This discussion will emphasize the major US air carriers, but the techniques can be applied to all segments of the air transportation industry, and indeed to any industry in any country. The treatment will be broken down into several parts. Section 2 will outline the important variables measuring financial health used by analysts, and will outline a basic but very useful model. Section 3 will apply that key model to the pre-September 11 environment. It will show that the airlines' problems predate the events of that date. Section 4 will explore the causes of the plethora of filings, while Section 5 will look at several alternative models. The final section is a conclusion.

2. Measures of financial fitness

For years, financial analysts have sought ways to more accurately measure financial condition and the probability of stress or failure. Traditional analysis has employed financial ratios derived from the income statement and the balance sheet. It is common to organize these ratios into four different groups that gauge the different aspects of financial health (Brigham et al., 1999). These four groups are liquidity, leverage, activity, and profitability ratios. Liquidity ratios assess the ability of the firm to pay its obligations as they come due. Leverage ratios measure the extent to which a firm uses debt finance to acquire assets. They also are an indirect indication of the volatility in returns. Activity ratios measure the efficiency of asset usage. Profitability ratios gauge the profitability of the firm relative to some base such as sales, total assets, and equity or net worth.

Financial analysts have tried to increase the effectiveness of these types of ratios by the use of statistical and mathematical models that combine ratios to increase their predictive ability. This chapter will summarize several different models, some of which have been developed using only air carrier data. Three of the models discussed employ a technique known as multiple discriminant analysis (MDA). Another uses a regression approach called logistics regression (logit). A final approach borrows from artificial intelligence and seeks to train networks that can learn to categorize, much like the brain functions.

2.1. The Altman Z score model

Altman (1968) was the first to design a bankruptcy-forecasting model using MDA that could specify key variables that forecasted stress in advance. MDA is a statistical technique used to classify an observation into one or several *a priori* groups depending on the observation's individual characteristics. In this case, the groups are failed or solvent firms. The characteristics, called the independent or predictor variables, are quantitative, that is, they can be measured on a scale. Here, they are financial ratios. The dependent variable, however, cannot be expressed quantitatively. The dependent variable is thus coded in as [0, 1] and [1, 0], for failed and solvent. The MDA technique attempts to derive a linear combination of the independent variables that best discriminates between the groups.

The Altman model has several different variants, the primary one of which is called the Z score model. It weights the four different types of ratios noted above. The model is

$$Z = 0.012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5,$$

where X_1 to X_5 are defined as the following financial ratios derived from the firm's income statement and balance sheet:

- X_1 is the working capital (current asset minus current liabilities)/total assets. It is a measure of liquidity.
- X_2 is the operating profits (or earnings before interest and taxes)/total assets, a measure of current profitability.
- X_3 is the retained earnings/total assets, a measure of past profitability.
- X_4 is the market value of stockholders' equity/book value of total debt. It is the inversion of the popular debt/equity ratio used by bankers, and it measures financial leverage. The market value of equity is the average stock price for the year (high plus low divided by 2) multiplied by the number of shares of common stock outstanding.
- X_5 is the operating revenues/total assets. It is an activity or turnover measure.

While the derivation of the model is complex, its use is quite simple, as will be demonstrated shortly.

Note that all four types of financial ratio appear in the model. The higher each of these ratios, the stronger is the financial condition of the firm. The intercepts terms for each of the ratios in the model were derived via the application of stepwise MDA regression. The critical values of Z are 1.81 and 2.99. Scores of less than 1.81 are indicative of severe financial stress, and thus identify a potential bankruptcy candidate. The lower the score, the greater is the chance of failure. Scores above 2.99 indicate a stronger financial condition and little likelihood of failure in the near future. Scores between those barriers are less predictive. Altman referred to them as being in the "zone of ignorance," where classification was more difficult. If a single cut-off point was desired, Altman recommend the use of an index of 2.67 as the dividing line. The model, built on a sample of 33 bankrupt and 33 solvent firms, was 76% accurate in forecasting bankruptcy in advance of the event. It was been widely accepted in the financial community as a valid model, and the 76% accuracy rate has become a standard for comparative purposes. In addition to forecasting receivership, the model has been shown to be useful in assessing relative financial performance, (i.e. for ranking firms in a sample or population), and for managing turnaround situations (Altman, 1983).

3. Application of the Altman *Z* score model to major US airlines

The basic Altman *Z* score was applied to the US airlines, and successfully predicted the demise of both Continental and Braniff in 1982 and 1983, respectively, thus demonstrating is applicability to the airline industry (Gritta, 1982). It was later applied to the major carriers over the time horizon from 1966 to 1996 (Gritta et al., 2000) and then updated to the year 2000 to capture the financial status of the carriers just prior to September 11. Table 1 summarizes the results of those combined studies. Data for the year 2001 are also presented. Included are those carriers that have failed or were merged. The criterion for inclusion in the table was that the carrier be classed as a "trunkline" under the old definition, or as a "major" under the new classification (those airlines with revenues of US $1.0 billion or more). Only passenger carriers are scored. The carriers are grouped into two categories: the non-failed and the failed/stressed. The table also notes the dates of bankruptcies/reorganization, mergers, and the entry of "new" major airlines, such as Alaska, America West, Southwest and US Airways.

An example will help explain the derived scores. In 1999, Southwest's financial ratios were as follows:

$X_1 = -5.88\%$,
$X_2 = 42.15\%$,
$X_3 = 13.81\%$,
$X_4 = 335.46\%$,
$X_5 = 0.84$.

The calculated *Z* score for Southwest is thus:

$$Z = 0.012(-5.88\%) + 0.014(42.15\%) + 0.033(13.81\%) + 0.006(335.46\%) + 0.999(0.84) = 3.82.$$

Table 1
Financial status of US airlines as determined by Altman *Z* scores

	Year																	
Airline	1966	1973	1978	1981	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Non-failed																		
Alaska (a)					2.00	2.15	1.36	1.23	0.63	0.91	1.11	1.26	1.49	1.76	2.36	1.92	1.40	1.04
American	2.01	1.43	2.17	2.00	1.65	1.76	1.06	1.02	0.96	1.13	1.21	1.25	1.46	1.75	1.96	1.60	1.37	0.59
Delta	5.86	4.19	4.17	3.54	2.41	2.59	1.60	1.43	1.11	1.15	1.12	1.51	1.55	1.97	1.98	1.49	1.33	0.64
National (a)	4.64	2.91																
Northwest (a)	6.50	2.31	3.14	2.85	2.03	2.20	1.40	1.26	1.10	1.38	1.53	1.56	1.75	1.70	0.89	1.20	1.26	0.70
Southwest					1.80	1.92	1.86	1.79	2.33	3.36	3.07	2.71	2.85	3.03	3.61	3.82	4.21	3.83
Western (a)	4.10	2.24	2.43	1.67														
Failed/stressed																		
America West (b)							0.81	0.29	–0.08	0.66	1.47	1.75	1.52	NA	NA	NA	NA	NA
Braniff (b)	2.67	2.23	2.35	0.90	1.81	NA	NA	–3.8										
Continental (b)	3.64	1.08	2.05	1.57	0.93	1.08	0.30	–0.04	–0.20	0.88	0.50	0.99	1.61	1.75	1.83	1.46	1.59	0.90
Eastern (b)	2.16	1.10	2.21	2.19	–0.14	–0.95	–1.2											
Pan Am (b)					0.45	–0.23	–1.9											
TWA (b)	2.30	1.54	2.19	1.86	1.24	0.96	0.67	–0.26	0.02	0.68	0.44	1.20	0.79	0.88	0.75	0.12	0.04	–0.35
United (b)	2.34	1.67	2.57	2.08	1.92	2.16	1.60	1.28	0.94	1.26	1.16	1.39	1.50	1.42	1.24	1.17	0.82	–0.13
US Airways (b)					1.76	0.85	0.67	0.67	0.31	0.42	0.01	0.46	0.71	1.43	1.68	1.08	0.82	–0.04

Source: All scores calculated from data contained in the *Handbook of Airline Financial Statistics and Air Carrier Financial Statistics Quarterly*.
Notes:
(a) Alaska became a major carrier in the late 1980s; National was acquired by Pan Am in 1979; Southwest became a major in 1989; and Western was acquired by Delta in 1979.
(b) American West filed for bankruptcy in 1991; Braniff for reorganization in 1983 and again in 1989, then disappears; Continental files in 1983 and again in 1990; Eastern files reorganization in 1989, then bankruptcy in 1991 and disappears; Pan Am files in 1991, then becomes a regional airline; TWA files in 1992 and again in 1995; US Airways becomes a major carrier in 1989, files in 2002, then exits under Chapter 11 in 2003; United files in 2002 and now operates under Chapter 11.

Note that the ratios X_1 through X_4 are input as percentages. Thus, if the ratio of working capital to total assets was (US \$10/\$100), it would be input as 10.0%. X_5, however, is input as an index. If operating revenues and total assets were US \$125 and US \$100, respectively, the ratio would be input as 1.25, and not as 125%.

Three facts are evident from Table 1. First, the model does a good job of separating the solvent from the bankrupt/stressed. The scores of those carriers remaining solvent over the time are significantly higher that those of the failed/stressed group, especially since 1988. They ranged between 1.00 and 3.00 for most of this period, while the typical scores of the failed/stressed group were consistently below 1.00 for more than 9 years, and often were negative. Second, a comparison of the most recent scores with those of 1966, a banner year for air transport, and 1973 is interesting. The sharp decline in scores since the 1960s is clear evidence of the second point. The contrast to 1966 is quite striking. The mid-1960s were good years for the carriers, and all ten original airlines cleared the 1.81 hurdle. Northwest (6.50), Delta (5.86), National (4.64), and Western (4.10) all had very high scores. Five of the ten had Z scores above 2.99, and Braniff was right at the dividing line of 2.67. Scores just before and after 1966 (not reported here) were also high. Generally, low and declining scores characterize the period after deregulation, however. Third, the low scores, relative to the Altman barrier of 1.81, are indicative of the severe financial problems of this industry over a long time horizon. With the exceptions of Southwest and Alaska, virtually all the major airlines have had chronically depressed scores, even during the period of record profits in the latter part of the 1990s. It is a symptom of just how fragile this industry was and still is. The low scores indicate that most of the larger carriers must adopt different operating and financial strategies if they wish to survive. By late 2003, many are moving to lower their costs structures in an effort to compete, forced by the competitive nature of this industry. Fleets are being streamlined, cost-efficient regional jets are being introduced, wages and salaries are being slashed, route structures are being carefully pruned, and capacity is being reduced.

4. Causes of the US airline industry's problems

One advantage of the basic Altman model is that the key variables in the predictions are clearly visible. In the case of air transport, the key lies in the variable X_4. It is a leverage measure, an inversion of the popular debt/equity ratio used by most bankers to gauge the extent of financial leverage.

Financial theory (Brigham et al., 1999) suggests that there are three different levels of risk facing all firms: business risk, financial risk, and total or combined risk. Business risk is defined as the instability of operating profits over time. It has many causes. Fixed costs create instability, since costs cannot be cut immediately in response to a decrease in operating revenues. Demand for air transport is highly cyclical, thus adding to this instability. Finally, competition leads to increased risk

Table 2
Debt ratios X_4 for the major US airlines in 1966

Carrier	X_4 (%)
American	104.3
Braniff	241.2
Continental	275.7
Delta	565.2
Eastern	154.3
National	444.4
Northwest	745.2
TWA	166.7
United	179.6
Western	341.4

because of the inability to dictate pricing. Financial risk is defined as the added instability introduced in earnings after taxes and earnings per share to stockholders. It has only one cause – interest on the long-term debt used to finance assets. Total or combined risk is the total variability in earnings as revenues change. It is caused by all the factors mentioned above.

The problem is that business and financial risk interact in a multiplicative way. That is, one risk magnifies what the other risk has already magnified. The airline industry is particularly high in business risk. It does have fairly significant fixed costs, demand for its service is very cyclical, and intense competition exists today and has since deregulation in 1978. Financial theory suggests that industries, such as the airlines, that are high in this risk should moderate their use of debt finance. The failure to do so leads to very high levels of total risk. After-tax profits become very unstable, and the probability of failure increases substantially. One study comparing the airlines with a large sample of industrial firms has demonstrated the very high degrees of total leverage in this industry relative to other industrial groups (Gritta et al., 2003).

The X_4 variable is thus very important in this industry. As a group, the airlines have long relied heavily on debt to finance the acquisition of aircraft. The results are obvious in the Z scores already noted. A quick look at debt ratios, X_4 shows the root of the problem. Table 2 shows that ratio for the major carriers for the year 1966: all of the carriers have relatively sound equity to debt ratios. Table 3 shows the same ratio for 2000, except where noted. The sharp decline in the ratios for all the carriers, except Southwest, is noteworthy, as is the significant difference between the two groups. The low ratios for United and US Airways, and possibly American, are also of interest in light of the 2002 filings of both the former carriers and the threat of the latter to file in early 2003. The penalty for excessive debt finance has clearly been shown. It should also be noted that these ratios do not

Table 3
Debt ratios X_4 for the major US airlines in 2000

Carrier	X_4 (%)	Last year of operation (a)
Non-failed		
Alaska	88.5	
American	64.9	
Delta	56.1	
Northwest	25.7	
Southwest	335.5	
Failed/stressed		
America West	–11.4	1991
Braniff	–52.9	1991
Continental	–35.5	1990
Eastern	–41.3	1991
Pan Am	–42.9	1990
TWA	–35.2	1992
United	6.8	2002
US Airways	26.7	2001

Note: (a) The figures are for the year prior to bankruptcy for those carriers that have disappeared/failed and are based on the book value of equity in the case of the latter carriers.

reflect the impact of operating leases as debt. There has been a tremendous increase in the use of operating leases by the air carriers over the past two decades. Thus, the debt burdens of the carriers are understated to some extent (Gritta et al., 1995). There is no doubt that the inclusion of this lease data can aid the predictive ability of the model (Elam, 1975).

While an in-depth treatment of the causes of the reliance by this industry on debt finance are beyond the scope of this chapter, the sharp decline in Z scores after the late 1960s can be traced to the huge dependence on long-term debt to finance wide-body aircraft in the 1970s. This reliance thus significantly impaired the equity/debt ratio X_4. Airlines also mis-forecasted traffic growth during this period, resulting in excess capacity. Research has documented the effects of this debt on earnings over time during this period (Gritta, 1979). The deregulation of the industry in the late 1970s compounded the carriers' problems, as business risk in the industry increased as competition grew.

5. Other models of forecasting

5.1. The ZETA credit score model

The Altman Z score model is frequently used to assess financial condition because it is so well known and because it is in the public domain. In addition, the input

Table 4
ZETA credit scores for the major US airlines just prior to and after September 11, 2001

Carrier	Date December 2000	August 2001	December 2001
Alaska	–0.19	–0.56	–0.73
American West	–1.14	–1.84	–2.79
American	2.54	1.90	–0.99
Continental	1.02	0.37	–0.45
Delta	1.75	0.94	–1.06
Northwest	–1.62	–1.96	–2.15
Southwest	6.87	7.65	6.03
United	–0.10	–1.16	–2.02
US Airways	–2.72	–2.42	–6.21

Source: Courtesy of R. Haldeman of ZETA Services.

data is readily available from company reports. There are other models available to the analyst, however. Altman later added several variables to his original model in an effort to increase its predictive powers. Called the ZETA credit score model (Altman et al., 1977), it takes the form

$$\text{ZETA} = a_1X_1 + a_2X_2 + \mathrm{a}_3X_3 + a_4X_4 + a_5X_5 + a_6X_6 + a_7X_7,$$

where:

- X_1 is the return on assets (the same ratio as X_3 in the Z score model);
- X_2 is the earnings stability (the deviation around a 10 year trend line of X_1);
- X_3 is the debt service;
- X_4 is the cumulative profitability (the same as X_2 in the Z score model);
- X_5 is the liquidity (measured by the current ratio);
- X_6 is the ratio of equity to debt (using market values and a 5 year trend);
- X_7 is the firm size (measured by the log of the firm's total assets).

As in the case of Z model, the approach used to derive the ZETA was MDA. The ratio is centered on zero. Scores less than 0 indicate a financially stressed profile. The more negative the score, the greater the risk of financial problems.

The ZETA credit score model was applied to the airline industry in 1984, and was found to be reliable (Altman and Gritta, 1984). The scores gave warning signs for several carriers that subsequently failed. These were US Airways (–0.06), TWA (–0.13), Eastern (–3.85), Pan Am (–4.17), Continental (–4.99), and Braniff (–15.42). Unfortunately, the model is proprietary, and the intercept terms in the equation are not published. ZETA credit scores are only available by subscription from ZETA Services of Mountainside, New Jersey. This limits its availability, at least to academic researchers.

Table 4 lists scores for the major airlines just prior to and after September 11, 2001. The scores are based on market values in the equation. As is the case with the Altman Z scores, the ZETA credit scores show the weak and deteriorating conditions facing most of the major airlines just prior to September 11. Five of the nine major airlines had negative ZETA scores at year-end 2000, and in August of 2001. Eight of the nine finished the year with negative indices. As noted above, two have filed (United and US Airways), and American threatened to file in early 2003.

5.2. *The AIRSCORE model*

It could be argued, however, that a model derived from a sample of the same industry would be even more accurate than a generalized model such as the Altman Z or ZETA credit score approaches. With that in mind, an industry-specific model, AIRSCORE, was specified, using a sample restricted to the US airline industry (Chow et al., 1991). It included a significant sample of the large and smaller carriers (the latter referred to as regional airlines). Using the MDA approached utilized by Altman, the model was derived as

$$\text{AIRSCORE} = -0.34140X_1 + 0.00003X_2 + 0.36134X_3.$$

The three ratios predictive of insolvency or stress are defined as follows:

- X_1 is the interest/ total liabilities (the imputed interest rate on debt);
- X_2 is the operating revenues per air mile;
- X_3 is the shareholders' equity/total liabilities.

Because the distribution of the scores made the application of a single cut-off point difficult and inappropriate, several "gray zones" were defined, and the model yielded results similar to the Altman Z and ZETA credit score models. It was able to achieve accuracy rates of between 76 and 83%, depending on the zone used. While the model is accurate, it does seem to be somewhat biased toward the larger carriers in the sample.

5.3. *Logistics regression analysis*

Logistics regression analysis has also been used to forecast financial stress, and has become widely accepted (Ohlson, 1980). Logit models estimate the probability of bankruptcy, and are also useful in ranking firms in terms of financial strength. A logit model has been used to specify a model for airline financial stress (Pilarski and Dinh, 1999). Called the P score, the model takes the form

$$W = -1.98X_1 - 4.95X_2 - 1.96X_3 - 0.14X_4 - 2.38X_5,$$

Table 5
P scores for the major US carriers

Carrier	Year 1999	2000	2001	2002
American	0.0000	0.0008	0.0356	0.2493
Alaska	0.0005	0.0087	0.0036	0.0063
Continental	0.0100	0.0095	0.0274	0.0881
Delta	0.0100	0.0000	0.0100	0.1300
Northwest	0.0135	0.0056	0.0078	0.0418
Southwest	0.0000	0.0000	0.0000	0.0000
TWA	0.0914	0.1554	0.3791	Merged
United	0.0107	0.0093	0.1903	0.4893
US Airways	0.0080	0.0568	0.2296	0.9046

Source: Courtesy of J. Goodfriend of the US Bureau of Transportation Statistics.

where

- X_1 is the operating revenues/total assets;
- X_2 is the retained earnings/total assets;
- X_3 is the equity/total debt obligations;
- X_4 is the liquid assets/current maturities of total debt obligations;
- X_5 is the earnings before interest and taxes/operating revenues.

P is determined by

$$P = 1/(1 + e^{-w}).$$

Note that several of the input ratios (X_1, X_2, and X_3) are ratios from the Altman *Z* score model. *P* is the probability of failure. The higher the *P* value, the greater is the financial stress and the more likely is the chance of failure.

Table 5 provides *P* scores for the major carriers. The high *P* scores for American, United and US Airways are interesting. Preliminary research (Goodfriend, 2002) has shown a strong correlation between *P* scores and Altman *Z* scores.

5.4. Neural networks

Other efforts have been directed toward a different approach to the prediction problem. The use of artificial intelligence has gained widespread support for a variety of uses in finance, forecasting solvency being one major area of application.

Neural networks derive from research on the neural architecture of the brain, and attempt to mirror this architecture (Caudill, 1989). Neural networks are composed of interconnected neurons linked together through a network of layers. Inputs of data are provided to the network along with desired outputs. The network is then trained to classify new information based on the abilities it derives to separate the groups input into the model. Neural networks have several advantages over the MDA technique (Udo, 1993): they can be run on smaller samples than MDA models, they can tolerate "noise" or missing data, they can self-organize and learn by changing the network connections; and they can find and establish complex relationships among input variables. The actual construction of neural networks is, as the case with the MDA technique, quite complex.

Two separate studies were conducted on US air carriers using the neural network approach. Using 21 pieces of financial information from carrier balance sheets and income statements, a neural network was specified for the major US airlines. In a separate study, a neural network was trained to identify bankrupt/ stressed smaller carriers, known as large and medium regional airlines. The first study (Davalos et al., 1999) successfully classified all the major carriers that filed for receivership. The second study (Gritta et al., 2002) achieved an overall success rate of 88% in predicting stress for the smaller airlines. The use of neural networks thus provides an interesting supplement to the analyst in appraising financial health.

6. Conclusion

The financial situation of the majority of the US airlines has become very precarious over the last several years. Bankruptcy, an unheard of event in the USA during most of the airline industry's history, has become commonplace. Since 1982, over 50% of the members of the major carriers have failed, some more than once, and familiar names such as Braniff and Eastern have disappeared forever. While the events of September 11, 2001 have certainly been "the straw that broke the camel's back," this industry was financially unstable long before that sad date, as the data in this chapter attest.

Industry analysts could use tools to assist in forecasting stress in advance and in gauging the general financial fitness or health of this vital industry, to which this chapter has provided a brief introduction. The development and use of these tools is an important topic to many in the industry; creditors, stockholders, the flying public, and government regulators.

Several methods of assessing strength were presented. All were based on financial data and ratios derived from carrier balance sheets and income statements. The basic Altman *Z* score model has stood the test of time. It is easy to use, and the data necessary is readily available. It can be used to track the future health of this industry so vital to the US economy. Also outlined were the ZETA credit score, the AIRSCORE model, and the *P* score. The first is proprietary, but can be purchased from ZETA Services. All three yield results similar to the basic Z score. Also discussed was the use of neural networks.

There is room for further research in this area. The inclusion of other variables such as operating statistics and management variables could improve the predictive ability of some of the models. Attempts have been made to input less quantifiable variables, such as management factors, into predictive models, and these have been proven to aid the analysis (Gudmundsson, 2002).

Finally, this chapter has briefly examined the key variable in the US airlines' plight. The use of significant amounts of long-term debt in this industry has been the direct cause of the high failure rate of the carriers. The industry is high in business risk caused by the cyclical nature of demand for air transportation, fixed costs, and intense competition. It was argued that such industries should refrain from using too much financial leverage, the cause of financial risk. The failure to do so can cause wide fluctuations in returns to stockholders and will increase the likelihood of failure. The results of the carriers' failure to abide by this "sound principle of finance" are obvious.

References

Altman, E. (1968) "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *Journal of Finance*, 23:589–609.

Altman, E. (1983) *Corporate financial distress: a complete guide to predicting, avoiding and dealing with bankruptcy*. New York: Wiley.

Altman, E. and R. Gritta (1984) "Airline bankruptcy propensities: a ZETA analysis," *Journal of the Transportation Research Forum*, 25:150–154.

Altman, E., R. Haldeman and P. Narayanan (1977) "ZETA analysis: a new model for bankruptcy classification," *Journal of Banking and Finance*, 1:29–35.

Brigham, E., L. Gapenski and P. Daves (1999) *Intermediate financial management*. Hinsdale: Dryden Press.

Caudill, M. (1989) "Neural network primer," *AI Expert*, 4:61–67.

Chow, G., R. Gritta and E. Leung (1991) "A multiple discriminant analysis approach to gauging air carrier bankruptcy propensities: the AIRSCORE model," *Journal of the Transportation Research Forum*, 31:371–377.

Davalos, S., R. Gritta and G. Chow (1999) "The application of artificial intelligence to predicting bankruptcy risks facing the major US air carriers, 1979–1996: a neural network approach," *Journal of Air Transport Management*, 5:81–86.

Elam, R. (1975) "The effect of lease data on the predictive ability of financial ratios," *Accounting Review*, 7:25–43.

Goodfriend, J. (2002) *Analyzing the behavior of a potential financial risk index for the airlines*, Working Paper. Washington, DC: Bureau of Transportation Statistics, US Department of Transportation.

Gritta, R. (1979) "The effects of financial leverage on air carrier earnings: a break-even analysis," *Financial Management*, 8:53–60.

Gritta, R. (1982) "Bankruptcy risks facing the major US airlines," *Journal of Air Law and Commerce*, 48:89–107.

Gritta, R., G. Chow and E. Lippman (1995) "The impact of the capitalization of leases on financial analysis: an issue revisited," *Logistics and Transportation Review*, 30:189–202.

Gritta, R., N. Freed and G. Chow (1998) "Measuring the degrees of operating, financial, and combined leverage facing the US air carriers: 1979–1995," *Transportation Law Journal*, 26:51–71.

Gritta, R., S. Davalos, G. Chow and M. Huang (2000) "Small US air carrier financial condition: a back propagation neural network approach to forecasting bankruptcy and financial stress," *Journal of the Transportation Research Forum*, 56:35–46.
Gritta, R., G. Chow and S. Davalos (2002) "Air carrier financial fitness: an assessment over the 1966–1996 period," *Journal of the Transportation Research Forum*, 39:171–181.
Gritta, R., G. Chow and E.J. Freed (2003) "Business, financial and total risk in air transport: a comparison to other industry groups prior to September 11, 2001," *Journal of the Transportation Research Forum*, 57:149–156.
Gudmundsson, S. (2002) "Airline distress prediction using non-financial indicators," *Journal of Air Transportation*, 7:3–23.
Ohlson, J. (1980) "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of Accounting Research*, 18:109–131.
Pilarski, A. and T. Dinh, (1999) "Numerical scoring approach to credit risk analysis," in: G. Butler and M.R. Keller, eds, *Handbook of airline finance*. Washington, DC: McGraw-Hill.
Udo, G. (1993) "Neural network performance on the bankruptcy classification problem," *Computer and Industrial Engineering*, 25:377–380.

Chapter 33

VOLUNTARY TRAVEL BEHAVIOR CHANGE

PETER R. STOPHER
University of Sydney

1. Introduction

There has been much interest of late in a new strategy that is termed generally voluntary travel behavior change (VTBC). It has been given a number of different names. In Perth, Western Australia, when first introduced there, it was called TravelSmart. This term has become widely applied as the name by which voluntary travel behavior change programs are known. Components of TravelSmart may include the procedures known as Indimark, Travel Blending, Living Neighborhoods, and Living Change, among others. All of these procedures have in common a fundamental approach to the problem of today's over-reliance on the car as a means of urban travel. However, in implementation, they are, in certain areas, significantly different from one another.

Fundamentally, each of these approaches to VTBC is based first on the notion of seeking to generate voluntary change in travel behavior. Some may think that any changes in travel behavior are voluntary. However, the specific notion of "voluntary" in this application is that nothing is changed in the transport system, no pricing or taxing incentives or disincentives are provided, and it is rather a matter of providing better information to people about their transport (and non-transport) options. The idea is that, given this improved level of information, people will decide to make changes in their travel behavior. There is evidence from Perth in Western Australia that few people are aware of the options offered by public transport, and not many more are aware of bicycling and walking options (Socialdata, 2000). Voluntary behavior change may be defined as "change that occurs when individuals make choices for personal reward without a top-down mechanism, regulation of any sort or a feeling of external compulsion ... an individual decides to make a change so that he or she will improve their personal life in some way" (Ampt, 2003).

IndiMark is a social marketing approach. It is based on informing people about the transport-related infrastructure and services that are available in their locality. In this approach, the emphasis is generally on improving the practical understanding

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

of the transport system and services that is held by individuals. Where this understanding is incomplete, this improved knowledge and human capacity leads to shifting travel from environmentally unfriendly modes (principally driving the car, and especially driving alone) to environmentally friendly modes (public transport, walking, bicycling, or riding in the car as a passenger), without significantly altering the number of out-of-home activities in which people engage, and also probably not changing the amount of time spent traveling. Indeed, in this approach, it is often seen as most successful when there is no change in the number of trips, but a reduction in the vehicle-kilometers by car drivers, in particular from before the application of the approach to after.

The Living Change/Living Neighborhoods approach focuses on the problems that individual households have in travel and activities in their community. It describes problems in terms of their personal effects on the individual and households, and provides opportunities for individuals and households to solve their own problems. In so doing, members of the community also assist in solving community or government problems. This approach is as likely to lead to an overall reduction in travel as a shift in mode. In Living Change or Living Neighborhoods, however, success would be considered to have been achieved if the overall person-kilometers of travel were reduced, and possibly the number of trips and out-of-home activities were reduced, without necessarily changing the travel mode at all. Rather, this approach seeks to offer possibilities to reduce travel overall, albeit with the intent that travel by car, especially drive alone, will be reduced the most.

The travel behavior modification sought is thus a reduction in the use of drive alone, especially, either through a shift of some car driver trips into more environmentally friendly modes without changing overall levels of travel, or through blending of activities, destinations, modes, or, over time, to reduce the overall need for travel. The outcome in either case would be a reduction in the level of car traffic, without the imposition of politically unpleasant strategies, such as pricing or tolling options, and without the need for significant investment in transport infrastructure. Indeed, much of the emphasis of these approaches is toward making better use of the currently available infrastructure, and particularly using unused capacity, whether by mode or by time of day.

Providing improved information is much cheaper than investing in new infrastructure. If such informational enhancement can lead to significant shifts in travel, whether reducing overall travel or shifting travel out of the car, it offers the potential to alleviate some of the congestion that seems to have become ubiquitous in early twenty-first century cities, without incurring major capital outlays. It also can contribute in a substantial and significant way to the reduction of greenhouse gases, through reductions in vehicle-kilometers of travel involving cars, which are the major contributors to the production of carbon dioxide in most developed countries of the world.

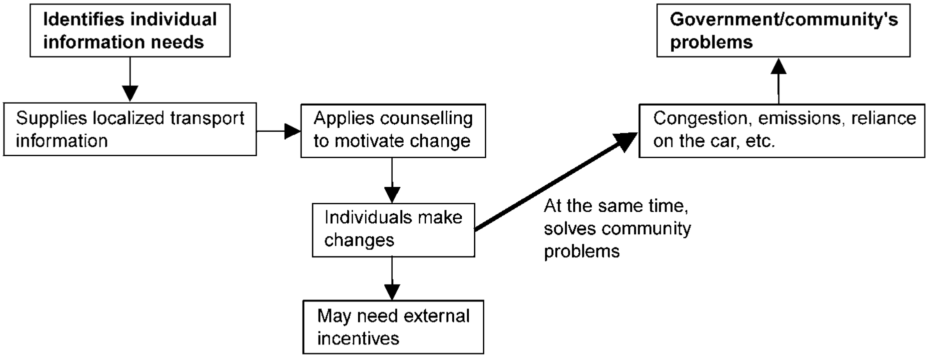


Figure 1. The social marketing approach.

In this chapter we will first explore the mechanisms employed for each of the two basic approaches to VTBM. We will then examine what is currently known about what the strategy can achieve. Third, we will consider under what circumstances these strategies may be most applicable, and finally we will address the issue of how to measure the effectiveness of the strategies. We will then draw conclusions as to the potential of this policy direction.

2. The social marketing approach (IndiMark)

Individualized marketing focuses on communicating to people the possibilities offered by the public transport system, walking, and bicycling as options to driving their cars. It focuses on providing information and opportunities for them to change their travel behavior so as to reduce reliance on environmentally unfriendly modes of travel. It focuses on changing the travel modes that people use without impacting their overall level of mobility, as measured by the number of out-of-home activities in which they engage, the number of trips undertaken, or the overall trip lengths. The approach can be broadly summarized as shown in Figure 1.

The strategy of IndiMark is first to segment the population into three groups (James, 1998):

- households that are interested in and willing to consider using other modes in place of the car (the I group);
- households that are already using alternative modes to car driver at least once a week (the R group);
- households that are not interested in or are not willing and able to change some of their car use to use of environmentally friendly modes of travel (the N group).

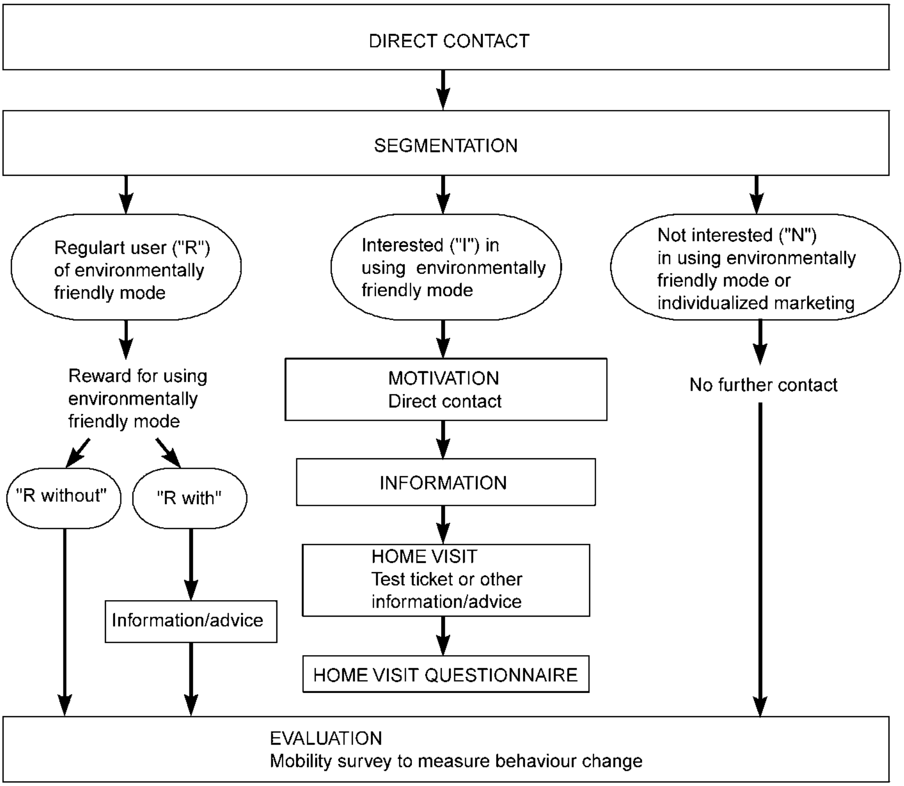


Figure 2. Elements of the individualized marketing approach. (Source: Socialdata, 2000.)

Figure 2 illustrates the basics of individualized marketing.

In the application of IndiMark, a customized treatment is determined for households in the I and the R groups, based on determining the particular needs and situations of those households. Usually, the R group divides into two subgroups, one of which desires no further information, while the other desires at least some further information. As its name implies, IndiMark or individualized marketing, involves tailoring to each household specific information that is relevant to that household. There is a range of levels and types of assistance that may be provided to households, depending on individual circumstances, including:

- no further contact (for the N group);
- providing generic brochures to those requesting them;
- providing location-specific information to those who request it;
- providing personalized documents (e.g. travel plans for public transport trips) or verbal advice;

- passing on comments about alternative modes of travel;
- providing a token reward;
- providing community/local service information (e.g. contacts for walking groups or discounts at local businesses, such as bike shops);
- for a small proportion, arranging personal home visits or test tickets for the public transport system and/or a personalized consultation on bicycle maintenance and riding.

Another important feature of IndiMark is that it is designed to be applied to an entire suburb. Ideally, there should exist (or be conducted as part of the implementation) a recent travel survey that would document existing travel patterns, and would show the potentials for travel behavior change. It has usually been the case that a geographic area is designated for the implementation, and attempts are made to contact all households in the designated area. As is seen later in this chapter, this is not always the approach taken in other travel behavior modification approaches.

2.1. Applications of the social marketing approach

The earliest application of this method in a multi-modal context was a pilot study in Perth, centered on the suburb of South Perth (James, 1998) which began in 1997. In the pilot, it was found that about 40% were either already using environmentally friendly modes to some extent or were interested in receiving information about alternatives to driving the car, while the remainder were not interested or did not respond to the initial survey. It was reported that this pilot achieved a 10% reduction in car use over the entire target population to which the pilot was applied.

Subsequently, IndiMark was subjected to a full-scale study again in South Perth, this time being applied to the entire suburb. This was undertaken in 2000. IndiMark has been implemented in the following locations, either as a pilot or as a large-scale implementation:

- South Perth, Australia (pilot - 1997, large-scale - 2000);
- Gothenburg, Sweden;
- Cambridge, Australia (large-scale);
- Viernheim, Germany (pilot and large-scale);
- Marangaroo, Australia (large-scale);
- Gloucester, UK (pilot);
- Portland, Oregon, USA (pilot);
- Subiaco, Australia (large-scale);
- Frome, UK (pilot);
- Marangaroo, Australia;

- Fremantle, Australia (large-scale);
- Armadale, Australia (large-scale);
- Vincent, Australia (large-scale);
- Melville, Australia (large-scale);
- Belmont, Australia (large-scale);
- Alamein, Australia (large-scale);
- Darebin, Australia (large-scale) (ongoing);
- Grange, Australia (pilot);
- Townsville Hermit Pk, Australia (pilot);
- Townsville Mundingburra, Australia (pilot);
- Adelaide Marion, Australia (pilot);
- Adelaide Onkaparinga, Australia (pilot).

The reported results of these various applications of IndiMark in terms of a reduction in car trips over the entire target population range from 4 to 14% (Ker, 2004). Target populations for these applications ranged from 200 households to more than 15 000. Reports also show that the non-contact rate in these applications was around 6% (ranging from about 2% to almost 13%), and non-participation was a further 6% (ranging from about 4% to about 10%). Broadly speaking, the take-up rate for these applications appears to be on the order of 40% of households, i.e. this is the proportion of households that were categorized as either interested, or already using environmentally friendly modes, but desiring further information.

3. The community development approach (Living Neighborhoods/Living Change)

The community development approach is focused on helping people to help themselves. It relies on the motivation coming from individuals and households

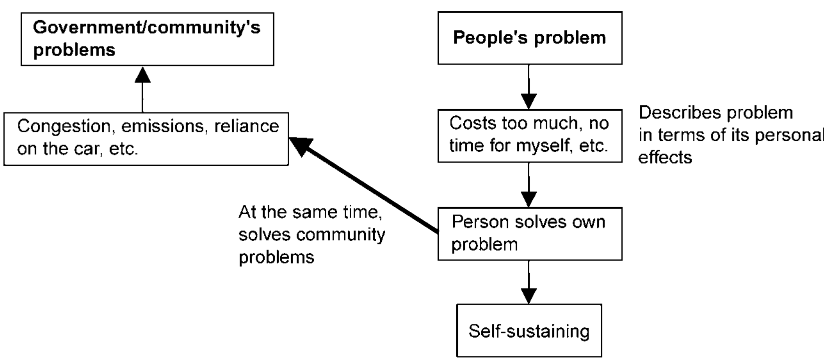


Figure 3. The community development approach.

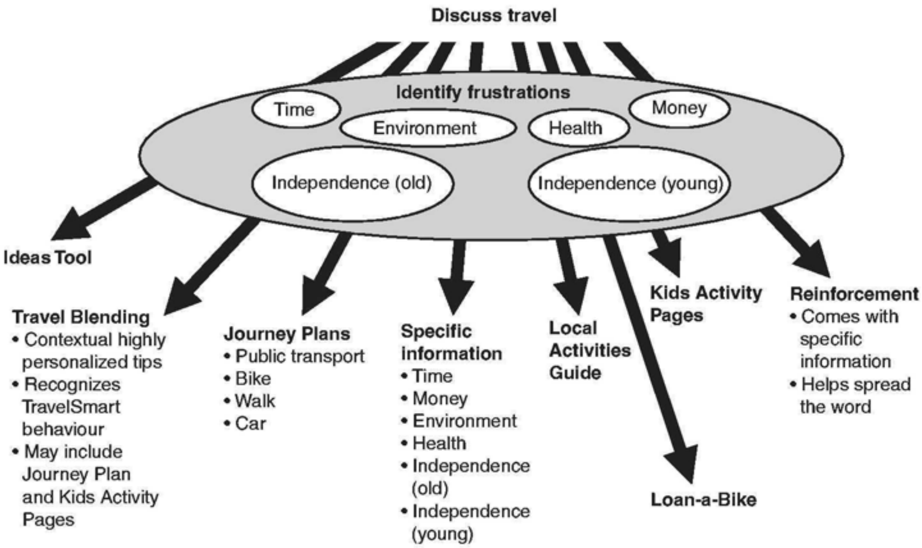


Figure 4. The Living Change process.

that perceive problems with travel and activities, and would like to solve them. It offers people assistance to get around their community and the larger region more easily, and possibly also opens up opportunities to participate in activities that were previously desired, but where knowledge of locations for the activities was lacking. Living Change may involve changing modes of travel for some trips, but may also involve doing things differently, or doing things at different times of the day. It is termed Living Change because it involves people in changing the way in which they live. Figure 3 illustrates the process of the community development approach.

The community development approach may involve the use of a number of different tools, any of which may be used alone or in combination to suit the particular household. As applied currently, this approach is initiated by a conversation about travel with a representative of the household. This conversation is aimed at identifying the frustrations that the household members may feel relating to their time, environment, health, or money. It also may have a great deal to do with the level of independence that individuals would like to enjoy. Various tools are then offered to the household, in an effort to provide information that would help household members to reduce their frustrations, and improve their living situation. Figure 4 illustrates how a full community development approach might be applied.

As shown in Figure 4, the Living Change approach includes the following tools:

- The Ideas Tool, to provide households with ideas on how to change current travel, activities, or timing of activities.
- Travel Blending, which involves members of households that select this option completing up to a 7 day diary, following which tips and suggestions are given on how to reduce overall travel, switch to alternative modes, etc. This is followed by a further diary and tips, and analysis that allow household members to practice the ideas suggested from the process.
- Journey Plans, which are designed to help people who have specific requests to try an alternative way of travel on foot, on public transport, or on a bicycle that they had not done previously, or routing information that may allow car journeys to be shortened.
- Specific information that consists of up to six brochures, dealing with how to save time or money in traveling, how to reduce the environmental impacts of travel, how to make travel less stressful, and how to become more independent (older and young people considered separately).
- Local Activities Guides and Kids' Activity Pages, to provide customized information on local activities, services and facilities that can permit people to find opportunities for things they need closer than they were aware.
- Reinforcement, which may include such items as a free ticket (only a day ticket is given, so that people do not feel obligated to give up the car, but rather just try public transport) for the public transport trip requested in a Journey Plan, information about what others have been able to do, and suggestions on how to get friends and colleagues to consider participating.
- Loan-a-Bike, which is an option that can be provided if a local bicycle shop is engaged in the project.

There are several other tools, not shown in Figure 4 that may also be included. These include information that might help people to arrange to work at home occasionally, tools such as pedometers, walking and bicycling information guides, and public transport timetables and maps.

A key element of Travel Blending is that it aims to provide people with "an achievable goal rather than a set of general possibilities" (Perkins, 2001). This emphasizes that some change is possible for everyone in the short term, that such changes can be incorporated into people's lifestyles without having a negative effect on lifestyle, that everyone can also identify possibilities for long-term change, and that changes can be sustained over the long term.

3.1. Applications of the community development approach

The earliest application of this method was in 1994 for the National Roads and Motorists Association (NRMA) in Sydney, Australia. Travel Blending was further developed and refined for application in Adelaide and the UK in 1996–1998. Living Neighborhoods, Travel Blending, and Living Change have now been implemented in the following locations:

- Adelaide, Australia;
- Melbourne, Australia;
- Sydney, Australia;
- Brisbane, Australia;
- Leeds, UK;
- London, UK;
- Nottingham, UK;
- Bishopbriggs, Inverurie, and Paisley, UK;
- New Jersey, USA;
- Santiago, Chile;
- Canberra, Australia.

Many of the applications of Travel Blending, Living Neighborhoods, and Living Change have been to target groups, such as employees of particular agencies or firms, schools, and community groups. As such, evaluation of the results have not been reported in terms of the likely impact on a metropolitan population or subgroup of the population. In addition, only three of the above applications have been subject to an external evaluation (Adelaide in 2002, Melbourne in 2003, and Canberra in 2004). Overall, take-up rates in the target populations are reported as having ranged from around 30% to 55%. This is a similar take-up rate to that observed for the social marketing approach. Evaluations, because of the nature of the targeting, have been reported only in terms of the participating households. The changes in travel behavior that have been reported from the limited evaluations indicate about a 10–15% reduction in car trips by participating households, a 10% reduction in vehicle-kilometers of travel, and increases of 10–20% in the use of public transport by participating households. As with the applications of the social marketing approach, the sample sizes for evaluation have been quite small, in the range of 100–200 households.

4. Evaluating voluntary travel behavior change

The most difficult aspect of voluntary travel behavior change is evaluating the extent to which it results in actual behavior changes, and assessing the sustainability of those changes. Assessing changes in travel behavior is not isolated to voluntary travel behavior change. However, it has taken on prominence because of the possibility that significant and substantial changes in car driver use and vehicle-kilometers of travel can be achieved with what is a comparatively very low-cost process. If it were indeed true that shifts of 10% or so of car driver trips to other modes of travel can be achieved by this mechanism, then the cost–benefit ratio of VTBM is extremely attractive and would warrant extensive adoption of this approach for such goals as reducing enhanced greenhouse gas emissions and delaying more capital-intensive investments in the transport system. For example, the National Travel Behavior Change Program (NTBCP) in Australia is looking for an outcome from voluntary travel behavior change that will have over 186 000 households participating in the Australian Capital Territory, Queensland, South Australia, and Victoria, who will, over a 5 year period, reduce vehicle-kilometers of travel by more than 3850 million km, and will thereby reduce carbon dioxide emissions by over 1.2 million tonnes for that 5 year period (Pramberg, 2004). If such reductions can be obtained through this mechanism, then it is indeed a very powerful one.

There are several difficulties involved in evaluating this policy. First, there is a need to be able to measure the travel behavior of participating households for a period before the intervention, sufficient to establish what are normal travel patterns, vehicle-kilometers of travel undertaken, and mode use for those households. Second, there is a need to be able to measure travel patterns at various times after the intervention, to establish whether or not, and how much, travel patterns have changed, and whether or not the changes are sustained over the long term. In particular, the NTBCP needs the monitoring to continue over a 5 year period from 2008 to 2013. Two difficulties arise immediately here, in that travel patterns are not static at the best of times. Therefore, there is a need to have control groups that will allow determination of what are the probable changes that would have taken place in household travel without the intervention, over the same period of time. There is also the further problem that households themselves are not static, and that changes in household structure may take place that would result in additional changes in travel behavior that will be confounded with the changes caused by the intervention itself and by the passage of time and the effects of external factors. Third, the amount of change that is expected is not large in statistical terms. To measure small changes accurately requires large sample sizes, which poses yet another difficulty. Finally, all of this measurement requires some quite burdensome activities by those who have agreed to participate in the VTBC program. Motivating the response to measurement and maintaining it over a period of time is a substantial challenge.

4.1. How to measure change

There are two primary options to consider for measuring a change in a population. One is to perform two or more surveys, using independent cross-sectional samples of the population for each survey. The other is to form a panel, and repeat measurements on the same sample on each occasion. There are pluses and minuses to each of these possibilities (Golob et al., 1997).

Repeated cross-sections

The sample size required to measure a certain change with a given level of precision will generally be substantially larger for the repeated cross-sectional samples than for a panel (Kish, 1967; Richardson et al., 2003). Therefore, each cross-sectional survey sample will need to be large. The determination of appropriate sample sizes is one of the issues that will be addressed in this project. Generally, the difference in sample size between a panel and repeated cross-sections for a phenomenon such as travel is likely to be on the order of a factor of two or more. This results because the variance of the difference between two occasions is reduced by the covariance between the two occasions. If the two samples are independent, the covariance is normally assumed to be zero. In a panel, however, the covariance term can be large, and may be close to the sum of the variances on the two occasions, thereby reducing the variance in the difference substantially. For example, under a particular set of assumptions, in which the population to which VTBC is applied is very large, the cross-sectional sample required for a level of accuracy of ±2% with 95% confidence, where the measured change is about 10%, could be as high as 8125 households. Using the same set of assumptions, the panel sample would need to be 4489.

On the other hand, cross-sectional samples are much easier to draw. Provided that the population from which the sample is to be drawn is large enough, the level of non-response is likely to be similar in each cross-sectional sample. Also, repeated cross-sectional samples will reflect changing demography in the population, changes in household composition, and other similar changes that may be taking place over the period of the surveys.

Panel surveys

Apart from substantially lower sample size requirements, the other principal value of the panel survey is that it measures the dynamics of change much more clearly than repeated cross-sectional samples. In a panel, comparisons can be made for the same household on each occasion that measurement is undertaken, and the magnitude of change is directly measurable. At the same time, it is also possible to track the magnitude of other changes that may have taken place in the characteristics of the household. In repeated cross-sectional surveys, it is only possible to make aggregate comparisons, and there is the underlying requirement to assume that the households in one cross-sectional survey are equally representative of the population as those in the next survey. Even if the samples are stratified into certain socio-demographic groups, comparisons are still at an aggregate level, and the errors are not significantly reduced. Uncovering underlying causes for dynamic changes is generally not possible in repeated cross-sectional surveys. For example, repeated cross-sectional surveys may show that bus ridership is remaining constant over a period of years. From this, it may be assumed that bus riders comprise a particular segment of the population, so that marketing efforts should be targeted to car users, to persuade them to shift from car to bus. However, a panel survey may reveal that, while the total number of bus riders remains unchanged, the individuals who are riding the bus are constantly changing. The panel survey might reveal that as young adults leave school and begin to work they become bus riders. However, as they begin to amass savings and to establish an earnings level that permits it, they purchase and begin to use a car, in place of riding the bus. This would suggest a completely different marketing approach, in which the effort would be to retain bus riders, as they become more wealthy, and to persuade them not to stop using the bus.

There are a number of problems that arise with panel surveys, however. First, panels are subject to attrition. This is the loss of panel members through several mechanisms, including households that move away from the study area, households that cease to exist due to some form of break up, and households that drop out of the panel. In the USA, for example, it has been estimated that panel attrition may run as high as 30% per year, for a panel conducted on an annual basis. More frequent measurement of the panel may reduce attrition due to moving and break-up, but will often increase attrition due to households becoming tired of participating. Another problem with panels is that of "conditioning." This is the process in which households involved in repetitive surveys either change their behavior because of what they learn through participation, or become able to provide false answers that they deem to be more politically "correct" or desirable in response to the survey questions. In this process, the households become less and less representative of the population, because their participation in the panel is changing what the panel measures. At the same time that this occurs, there is also panel fatigue, which results in panel members giving less attention to the survey task in repeated waves, and eventually dropping out of the panel.

Another issue is that the panel may become less representative of the population as time goes by, depending on how attrition is handled. There are three principal methods available to deal with attrition. The first is to oversample the panel initially, so that the panel declines in size over time, due to attrition, and each successive wave of the panel measures a decreasing subsample of the original panel. The second is to make up for the attrition at each wave of the panel, by selecting new members of the population to replace the lost households, where the replacement households are chosen to be as similar as possible to those who have been lost from the panel. The third is to make up for the attrition by selecting replacement households that reflect, as far as possible, the changed nature of the population. In this method, efforts would be made to change the constitution of the panel at each wave, so that it remains representative of the current population, as far as panel make-up will allow (Hensher, 1987).

To reduce attrition in panel surveys, it is also usually necessary to keep in contact with panel members between survey waves. This helps to keep the interest of the panel members, and may also help in finding out new addresses and other changes as they occur. However, this adds to the expense of the panel survey, offsetting some of the gains that are made by not having to draw fresh samples for each survey. Overall, however, panels appear to offer sufficient potential gains over repeated cross-sections that they should be considered seriously as the potentially best method to measure travel behavior change (Purvis and Ruiz, 2003).

4.2. Issues of variability

As mentioned earlier, there is a problem of variability in travel from day to day, from season to season, and in response to external stimuli. Variability is highest at a person level, and considering daily travel without controlling for the day of the week (Richardson et al., 2003). If a panel survey is conducted in which the same day of the week is used on each occasion, the variability decreases fairly significantly, dropping to about three-quarters of the uncontrolled day of week variability. If it were possible to measure travel for a whole week in each wave of the panel, the variability drops to a bit less than one-half of the uncontrolled daily variability. However, there has been very little success in the past with gaining cooperation of respondents to complete a travel survey for a whole week. There are also known problems with multi-day surveys of a drop in reporting completeness with increasing length of the survey (Madre, 2003). If a conventional travel survey approach were used, it would appear that the best option may be to use a 2 day diary, and to control this so that, in a panel, each household is asked to complete the diary on the same 2 days of the week as in previous waves. Two-day diaries, although still prone to reporting drop off on the second day, will provide some indication of day-to-day variability, without resulting in excessive respondent burden, whilst the use of the same days of the week on each wave of a panel would contribute to some substantial reduction in variability and concomitant improvement in accuracy for a given sample size.

4.3. Choosing a control group

This is another area of the evaluation that poses some considerable difficulties. The issue here is first and foremost how to select an appropriate control group.

The control group needs to be matched to the target population in terms of socio-demographics, location with respect to major employment centers, levels of public transport service and orientation of those services, car ownership and use rates, and trip-making rates and patterns. At the same time, the control group needs to be sufficiently physically removed from the target group that there will not be diffusion of the VTBC program to the control group population. Given that most applications of VTBC have been to specific suburbs within a large metropolitan area, the selection of an appropriate control group is by no means trivial. At the same time, failure to select a control group appropriately may make it impossible to determine if there are underlying travel behavior changes that need to be factored into the changes measured within the target population. There is no simple solution to this problem. Rather, there is a need to use all the means at one's disposal to select matching populations, considering such things as recent census information, journey-to-work travel information, geographic positioning, public transport service levels, and any other available information that is descriptive of the travel and locational situation, and to choose locations that are far enough away from the target areas to be immune from diffusion effects.

A related issue for the control group is motivation to participate in the repeated surveys. With the target group, there is at least the notion that there is some benefit being received by households that participate in the VTBC program. For the control group, there is no such benefit. As a result, response rates are likely to be lower from the control group, particularly from repeated measurements, resulting in the potential for more serious biases to exist in the control group. It is probably necessary to consider providing incentives to such respondents to participate in the surveys and to continue to do so.

4.4. Measurement error

A further problem that pervades all of the discussion on evaluating VTBC is that it is highly dependent on self-reporting. No matter whether households are interviewed through face-to-face surveys or over the telephone, or are provided with paper diaries to complete and return, or an Internet site to undertake the survey, all of these methods of survey involve self-reporting, usually relying on some degree of memory of recent travel and activities undertaken. It has been known since the first household travel surveys were conducted that people under-report their travel. Regular and repeated trips are usually reported well (e.g. trips to work or to school). Infrequent trips, especially short trips, are often forgotten, or are completely misreported. Recent work in the USA has shown that, using Global Positioning System (GPS) devices as a means to check on the completeness of diary reporting, the level of under-reporting may be as high as 20% of daily trip making, although the trips that are omitted are generally short trips (Wolf et al., 2003). Unfortunately, it may well be the short trips that are not reported that may be most susceptible to change under VTBC programs. At the same time, any omission of trips, which may be different among different individuals, jeopardizes the measurement of the extent of the changes that take place.

Possibly the least problematic survey method, and one that may lend itself to much of the evaluation of the long term effects of the NTBCP is the use of odometer surveys, in which responding households are asked to report the odometer readings of each of the household's vehicles at the beginning and end of a survey period. This will provide a means to determine the overall VKT driven by responding households, and would generally provide information about the reductions in VKT achieved. However, in order to determine whether there are biases in the sample, and to have a mechanism for expanding the sample, various household and person characteristics will also need to be collected each time that odometer readings are obtained. There is also no knowledge as to how accurately people report odometer readings in a telephone, written, or Internet survey. We do not know if they go and look at each car's odometer, or give a guess, based on the last value they recall seeing. A face-to-face interview, with an interviewer visit to the home for both the start and end of the survey period, would be the only method in which there would be reasonable certainty about the veracity of the reported odometer readings. This method will not permit assessment of shifts to public transport, walking, and bicycling, nor will it permit determination of whether travel has been reduced overall. It will also not provide indications of changes in vehicle occupancy, or of trip chaining. However, none of these effects appears to be of prime concern to the NTBCP, so this may remain as one of the most viable methods for the long-term monitoring.

4.5. External evidence

One of the other measurement devices that should be considered is that of external evidence of travel behavior change. By this, we mean evidence that is not gathered directly from households or persons who may have modified their behavior, but rather from external evidence, such as bus ridership, public transport ticket sales, and measurements of road volumes. Changes in bus ridership and ticket sales will again require lengthy periods of before data to establish trends, and also to cover areas that are not expected to be affected by the VTBC initiatives. This would allow more reliable estimation of whether there have been positive changes in the areas of application of VTBC in public transport use that may be attributable to the travel behavior changes. In the case of spot counts of road volumes, it is likely to be a little harder to establish, because the relative change in car driver use is much smaller than that for public transport use. For example, if public transport is carrying 4% of travel in a region, and car accounts for 64%, a shift of 5% of car trips into environmentally friendly modes is rather small. However, if two-thirds of those trips find their way onto public transport, the local share of public transport may increase by as much as 15 or 20%, which is much more readily evident and measurable. Corroborative evidence from public transport use and from road counts would be an important addition to the reliability of any measurement undertaken on the persons and households themselves.

4.6. Diffusion effects

Almost all of the evaluations to date have concentrated on applications of VTBC programs to households and communities. Measuring the diffusion effects of these applications has not been attempted. In addition, evaluating the diffusion effects of program applications in schools, workplaces, and other special locations has not been attempted either. The major difficulty here is in identifying the population to be measured. For workplaces, this may not be too difficult, since workers at many employers will not change substantially over a period such as the 5 years or so of this effort. Populations of schools and universities are, however, much more transient, and there are likely to be much more significant issues about how to measure these populations to determine if there is ongoing travel behavior change.

Further, it is the hope that such applications as these will be further diffused into the community through word of mouth, and reinforcement from "good stories" about changes that people have made. Given this, the measurement of the diffusion effects becomes challenging indeed. First, it would appear to be necessary to measure travel behavior in those communities from which workers, schoolchildren, or university students are drawn. Second, to determine if the VTBC programs at those workplaces, schools, and universities have had an effect in the community, it becomes necessary to ask if respondents have heard about TravelSmart, or any other brand name used for the VTBC programs. This almost immediately mandates that a panel cannot be used in this application. Asking this question of members of a panel will sensitize them to the existence of the program, and is more likely than not to cause subsequent behavior changes. Therefore, a program of non-overlapping cross-sectional surveys appears almost mandatory for this assessment. As noted before, however, this will have serious impacts on the sample sizes required and therefore on the expense of the survey.

One of the strategies that is likely to be necessary to measure the diffusion effects is to find suitable control groups. However, in this case, because it is unlikely that control groups can be found in reasonable proximity to the target populations, it will be necessary to begin measurement of travel behavior in both the potential target area and the control area at least 2 years before the VTBC program is to be initiated. What this will achieve is the establishment of trends in both the target and control populations, prior to any intervention. Then, as trends in the control group are observed after intervention with the target group, it becomes possible to estimate the probable changes in travel behavior that would have occurred in the target group, without the intervention. Hence, the actual diffusion effects can be estimated.

5. Evaluation of recent VTBC interventions

In those applications of VTBC that have been subject to an evaluation survey, there has been inconsistency as to whether or not a control group has been measured, and sample sizes for evaluation have generally ranged between about 100 and 700 households, although with most being around 200–400 households. In all cases, some form of travel diary has been the principal mechanism used to collect the travel data, and most of these have used self-completed diaries. There has generally not been control of the day of week between before and after surveys, and the after surveys have predominantly occurred only a few weeks or months after the intervention. Probably one of the longest time lapses has been in the case of the South Perth full-scale experiment, where an after survey was conducted 6 months and again more than 1 year after the original intervention (James et al., 1999), and has been repeated more than 2 years after, and is scheduled for a repeat more than 4 years after (C. Ashton-Graham, personal communication).

While government agencies interested in VTBC have begun to recognize the need for evaluation procedures, and, even more importantly, to recognize that the evaluation should be done by an agency that is different from the one undertaking the intervention, there still seems to be a considerable reluctance to budget sufficient time or money to undertake the appropriate level and reliability of evaluation, and there has been little attention paid to establishing the longer-term sustainability of the changes.

The evaluations that have been conducted to date are indicative of the effects of VTBC, but cannot be considered to provide reliable information on the overall levels of travel behavior change that can be achieved. In addition, it is not known at this point in time how sustainable these changes are. It is appropriate to note that human behavior, especially such habitual behavior as is involved in much travel behavior, is notoriously difficult to change. Furthermore, even where change can be effected in the short run, through mechanisms that may appeal to people as something worth trying, humans are very prone to lapse back into old habits after some period of time, either because they decide that they liked the old habits better after all, or because they realize that the gains they had hoped to achieve from the new habit have not actually eventuated. Therefore, the sustainability of VTBC must be considered to be somewhat speculative at this time.

6. Conclusions

VTBC appears to be a very cost-effective measure to change car-reliant travel behavior, when applied under the right conditions and in the right location. It tends to rely on the fact that people are not well informed about some of their alternatives, so will work much less well in situations where people have good information about travel and activity alternatives, and have made decisions on travel behavior that are based on this information. Second, it should work well where there is spare capacity in the public transport system, and where the environment is conducive to walking and bicycling as alternative modes. However, where there is no spare capacity in the public transport system, shifting travel from car to public transport will not be sustainable. In fact, it may be counter-productive in that those who attempt to use public transport that is already overcrowded are then likely to report back to others how bad it is, further cementing decisions to use cars. Similarly, in urban areas where there are poor pedestrian facilities, lack of bicycle paths and facilities, or many hills, shifts to walking and bicycling are also unlikely to be sustainable. On the other hand, there are probably few conditions under which people could not make many of the changes involved in Travel Blending.

VTBC is probably most effective when implemented as part of a package of transport initiatives (Ker, 2004). Implemented on its own, VTBC is likely to have only moderate impacts on the transport system, especially because many of the travel behavior changes are likely to affect non-peak travel. Even if the changes in travel behavior are only on the order of 3–6% of car driver trips shifting out of drive alone, the policy appears to warrant considerable further study and implementation.

Finally, there is a clear need for the development of much better short- and long-term evaluation procedures for VTBC. It is important that evaluation should be performed by a disinterested third party, and that the evaluation should involve both target and control groups. It is also necessary to investigate more carefully the sample sizes needed and the survey mechanisms that are most suited to evaluating these travel behavior changes. The potential of GPS as a mechanism for investigating and evaluating the extent of sustained travel behavior change appears to be well worth pursuing.

References

Ampt, E.S. (2003) "Voluntary household travel behavior change - theory and practice," *Proceedings of the 10th International Conference on Travel Behavior Research*. Lucerne (http://www.ivt.baum.ethz.ch/allgemein/pdf/ampt.pdf).

Golob, T.F., R. Kitamura, and L. Long, eds (1997) *Panels for transportation planning: methods and applications*: Boston: Kluwer.

Hensher, D.A. (1987) "Issues in the pre-analysis of panel surveys," *Transportation Research A*, 21:247–248, 265–284.

James, B. (1998) "Changing travel behavior through individualised marketing: application and lessons from South Perth," *Papers of the Australasian Transport Research Forum*, 22:635–647.

James, B., W. Brög, E. Erl and S. Funke (1999) "Behavior change sustainability from individualised marketing," *Papers of the Australasian Transport Research Forum*, 23:549–562.

Ker, I. (2004) "Household-based voluntary travel behavior change: aspirations, achievements, and assessment," *Transport Engineering in Australia*, 9:119–138.

Kish, L. (1967) *Survey sampling*. New York: Wiley.

Madre, J.-L. (2003) "Multi-day and multi-period data," in: P. Stopher and P. Jones, eds, *Transport survey quality and innovation*. Oxford: Pergamon Press.

Perkins, A. (2001) *The greenhouse abatement potential of travel behavior change initiatives. National Taskforce on National Greenhouse Strategy Measure 5.3: Promoting Best Practice in Transport and Land Use Planning.* Brisbane: Queensland Transport.

Pramberg, P. (2004) *National Travel Behavior Change Program (NTBCP): consultancy brief and offer specifications: long term monitoring*, Version 4. Brisbane: Queensland Transport.

Purvis, C.L. and T. Ruiz (2003) "Standards and practice for multi-day and multi-period surveys," in: P. Stopher and P. Jones, eds, *Transport survey quality and innovation*. Oxford: Pergamon Press.

Richardson, A.J., R.K. Seethaler and P.L. Harbutt (2003)"Design issues for before and after surveys of travel behavior change," in: *Proceedings of the 26th Australasian Transport Research Forum*. Wellington.

Socialdata (2000) *Potential analysis "Perth*." Perth: Department of Planning, Western Australia (http://www.dpi.wa.gov.au/travelsmart/pdfs/Report.PDF).

Wolf, J., M. Loechl, M. Thompson and C. Arce (2003) "Trip rate analysis in GPS-enhanced personal travel surveys," in: P. Stopher and P. Jones, eds, *Transport survey quality and innovation*. Oxford: Pergamon Press.

This Page Intentionally Left Blank

Chapter 34

NON-MOTORIZED TRANSPORTATION POLICY

MATTHEW PAGE
University of Leeds

1. Introduction

This chapter is concerned with transport policy toward cycling and walking, which are taken to be the main non-motorized modes currently in use in developed economies. It is possible that other non-motorized modes may emerge in the future, but it seems likely that these will share many of the main features of walking and cycling, which are limited by the power and endurance of the human body.

The chapter provides an overview of policy toward non-motorized modes. It includes some examples of the policy approaches adopted in various parts of the world after outlining the potential importance of these modes. It also discusses some important issues particular to these modes. This chapter is not meant to be a practical guide to actually providing for walking and cycling, which is a detailed subject in its own right.

Walking and cycling are often considered together because they share a number of characteristics, both being short-distance modes traveling at slow speeds. They also have low impacts on their surroundings, and are vulnerable when in conflict with motorized modes. Both could also be considered as more sociable and human scale activities than the use of motorized modes. When walking, for instance, one is much more part of the environment than just traveling through it; this is so much so that when talking about walking people often conflate the quality of the local environment with conditions for walking. As a walker one can enjoy public life in public spaces, without compromising and perhaps even enhancing the ability of others to do the same.

In the past, cycling and walking have been seen to some extent as road safety "problems," and therefore where policy has addressed non-motorized modes it has concentrated on specialized infrastructure to facilitate "safe" use. This has meant that the importance of wider transport policies for these modes has been overlooked as well as the impacts of policies in non-transport areas, which might have a significant impact on these modes. As the quality of the urban environment, equity of access and the regeneration of urban areas become more important,

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

walking and cycling could be recognized as having a very important role in policies that address these issues. It is such wider policies that might provide the all important setting and supporting context for walking and cycling and might have a greater impact than their traditional treatment as a transport mode.

In essence, the issues of most importance for non-motorized modes have a lot to do with the transport system as a whole, but also go much wider, concerning the nature and design of environmentally attractive, socially inclusive, sustainable urban areas.

2. Why walking and cycling are important

Walking and cycling have often been seen as the "Cinderellas" of transport policy, their needs often addressed only after the "main" modes of private and public motorized transport have been considered. However, even in developed countries, and despite declines in recent decades, they are still important modes of transport in terms of the number of trips made. In the developing world, both walking and cycling are extremely important; it is difficult to get reliable statistics, but many cities in the developing world are heavily dependent on non-motorized forms of transport (Whitelegg and Haq, 2003). In the UK, which has more walking and cycling than the USA, but less than much of Europe, walking accounts for 25% of all journeys and 80% of journeys under 1.5 km, with an average of 263 trips being made per person per year predominantly on foot. Despite significant falls in cycle usage over recent decades, more trips are made each year using cycle as the main mode (an average of 16 per person) than by surface rail (13) in the UK. However, the average distance traveled per person per year on foot has declined by 20% over the last decade, and the distance cycled by 6% (UK Department for Transport, 2003).

The importance of walking in particular goes beyond what the official statistics suggest, particularly as virtually every trip includes a walking element. Even so, walking and, to some extent, cycling tend to be undervalued as modes for a variety of different reasons.

Walking tends to suffer in terms of its visibility as a mode simply as a result of its very ubiquity. Almost everyone does it at some time or other, and most people think little of it. It is not always regarded, even by those who do a lot of walking, as a mode of transport in the same way as catching a bus or driving a car. Walking also seems to happen no matter what, in that some people will always walk, though it is easy to dissuade people from walking if they have a choice. Both walking and cycling are also actually quite difficult to measure, so they suffer from a lack of visibility in official statistics as well.

Both walking and cycling are low-cost forms of transport; this is both in the sense of being relatively cheap to provide for as well as being low cost to the user.

This means that the kind of well-funded lobby groups that exist for motorized transport do not exist (or not in the same form) to argue for walking and cycling. Facilities for walking and cycling do not require significant outlay or engineering challenge, and therefore tend not to attract attention, making it difficult to attract political or professional interest. Walking and cycling also have a relatively low impact on other road users and the environment, so they create few obvious problems of pollution or congestion that need to be addressed.

However, there are a number of compelling reasons why walking and cycling should be given greater prominence in transport planning and policy, and some of these are discussed below.

2.1. Walking and cycling are more efficient

Walking and, especially, cycling are efficient in terms of energy used to travel a certain distance – they do not directly rely on fossil fuels as a source of energy or directly produce any of the emissions responsible for climate change. They are also more efficient in terms of their requirements for road space and parking space at the destination. If drivers could be persuaded to leave their cars at home for some of their shorter journeys and walk or cycle instead then there could be benefits in terms of reductions in congestion. This argument underpins the justification for a greater role for walking and cycling in many policy documents, for example the UK's National Cycling Strategy (UK Department of Transport, 1996).

2.2. Walking and cycling are more sustainable

Sustainable development is commonly defined as encompassing equitable social progress, protection of the environment, careful use of natural resources and stable economic growth. In this sense, walking and cycling are more sustainable forms of transport in a number of different ways:

- they are less damaging to the environment, emitting negligible amounts of noise and air pollution;
- they require only a fraction of the expenditure of non-renewable resources required for the construction and operation of vehicles and infrastructure for motorized modes of transport;
- they represent more equitable forms of transport, both in the sense that they are available to a much wider proportion of society than other modes and that they impose little impact on others.

In the case of walking it could also be argued that a higher level of walking would enhance the local environment by improving feelings of local ownership and belonging and lead to a healthier community. Walking and cycling are also more sociable forms of transport – they are more exposed to what is happening around them and affected by it rather than being cut off from it.

2.3. *Walking and cycling impose less danger on others*

Walking and cycling are often regarded as more dangerous than motorized modes because they have higher levels of accidents per distance traveled, though the situation is slightly different if trips rather than simple distance are considered. However, it could be argued that these modes impose less danger on other road users than motorized modes due to their lower speeds and the lower levels of energy involved in their use. Common sense would also suggest that the behavior of walkers and cyclists is self-moderating to some extent because in any conflict, the walker or cyclist is at risk of injury and therefore has an immediate physical interest in avoiding a collision.

2.4. *Walking and cycling are healthy forms of transport*

In many developed countries there is increasing concern over the effects on public health of high levels of physical inactivity and sedentary behavior. Physical inactivity is associated with increased risk of cardiovascular disease, the development of diabetes, strokes, high blood pressure, various forms of cancer, and obesity. Walking is the most natural form of physical activity; it is also aerobic, weight bearing, and can be carried out at a moderate intensity. It also requires no specialized equipment, is possible virtually anywhere, and can easily be incorporated into daily activity. Cycling too can form a valuable part of a physically active lifestyle.

Cardiovascular disease is the main cause of death in Europe, and the cause of the greatest number of years of life lost due to an early death (Cavill, 2003). Over half of the adults in the EU do not meet the 30 minutes of daily moderate intensity physical activity recommended by World Health Organisation (WHO) as providing important health benefits, and lack of exercise is one reason for the high levels of obesity in the USA (Pucher and Dijkstra, 2003).

Concern about the public health effects of sedentary lifestyles is becoming an important issue, and looks likely to have a major influence on transport policy in the future. In the USA, the promotion of walking and cycling is emerging as a crucial policy to tackle the worsening epidemic of obesity (Pucher and Dijkstra, 2003). Using individual fitness as a reason for cycling and walking also appeals directly to the self-interest of the individual rather than to altruism or sensitivity toward societal benefits. This means that it could become a vital "lever" in encouraging even the most reluctant to increase their levels of walking and cycling.

There is a concern that increasing levels of walking and, especially, cycling might lead to greater numbers of road accidents. However, a basic analysis of the UK situation indicates that the public health benefits of higher levels of cycling far outweigh the loss of life through cycling accidents (British Medical Association, 1992). Also, in Denmark, an analysis of cycling levels and accidents in 47 towns suggests that where the level of cycling is high, cyclists have a lower accident involvement rate (Jensen et al., 2000). The message seems to be that cycle safety is best in towns where people cycle most.

3. Policies toward non-motorized modes

Policies toward non-motorized modes vary widely across the world, and have, in many places, changed significantly over the years.

In the developing world, levels of walking and cycling vary enormously. Statistics from individual cities suggest that in some places (especially in South-East Asia) walking and cycling together account for the majority of trips (Whitelegg and Haq, 2003). In some African cities, nearly half of all trips are made entirely on foot, and many of the remaining trips are by public transport, which will require a walked element (Interface for Cycling Expertise, 2000). However, many poorer countries are experiencing high rates of growth in car ownership and use, which often goes hand in hand with economic development. This directly puts pressure on non-motorized modes and tends to increase the demand for infrastructure for motorized modes, further marginalizing walking and cycling. One of the problems is that many developing countries themselves see non-motorized modes as "backward," and increasing motorization (especially higher levels of private car ownership and use) as the preferred development path. As a World Bank report observed (Interface for Cycling Expertise, 2000):

> Non-motorized transport plays a dominant role as the cheap, but sometimes unpleasant and dangerous, main mode of transport in some of the very poorest countries, with the consequence that it is frequently associated with poverty as something that countries aim to develop out of.

In the USA there has been a significant increase in interest in cycling and walking since the early 1990s, which resulted from the Intermodal Surface Transportation Efficiency Act, and its successor, the Transportation Equity Act for the 21st Century, and a change in the emphasis within the US Department of Transportation. The amount of money spent on walking and cycling projects has increased massively, and there has been a matching increase in research, planning studies, and the numbers of staff employed on cycling and walking projects.

This money has paid for a range of attractive facilities for walkers and cyclists, including high-quality segregated routes, cycle parking, bike-to-work campaigns, and education programs (Clarke, 2003).

However, while the funds devoted to walking and cycling have increased significantly, they still represent less than 1% of US federal transportation funding, a very large proportion of which still goes toward expansion of the road system for motorized modes. Much of this activity either directly undermines the popularity of non-motorized modes – by, for example, making it more difficult to cross roads on foot, or speeding up traffic – or worsens the relative position of non-motorized modes by making the car more attractive and encouraging car based development. In general, funding for walking and cycling facilities in the USA has gone on measures which do not conflict or prioritize these modes over motorized modes by, for instance, reducing capacity for cars. This is in contrast to the situation in some parts of Europe, where a policy of restraining car use has been implemented alongside providing facilities for walking and cycling. This can be doubly beneficial, since it can make the facilities easier to physically construct and increase their attractiveness relative to the car. In addition, there is a concern that those designing and constructing facilities for non motorized modes are not fully aware of the requirements of walkers and cyclists, as they are often more used to providing for motorized modes (Clarke, 2003).

In general in the USA, there is a feeling among advocates of cycling and walking that as much as possible has now been done without directly challenging the priority accorded to private motorized transport, which is felt to be politically difficult in the current climate (Komanoff and Pucher, 2003). Despite the recent interest in cycling and walking, they are still marginal modes in all but a very few areas. Overall they account for only 7% of urban travel in the USA (Pucher and Dijkstra, 2003).

In Canada the picture is a little different in that 8% of workers walk or cycle to work, but only 7% in the top five metropolitan areas (Transport Canada, 2002). There are areas where cycling is higher, such as Toronto, which is an older, denser North American city. There is also evidence that activity by advocacy groups and city officials in Toronto has been successful in encouraging more people to cycle and that a large proportion of cycling is for utilitarian purposes, rather than recreational (Pucher et al., 1999).

Walking and cycling levels vary significantly across Europe, but are generally higher than in the USA. Walking and cycling are highest in northern Europe, together accounting for 46% of urban travel in the Netherlands, 41% in Denmark, and 34% in Germany (Pucher and Dijkstra, 2000), although the levels for other European countries are lower. Most of this variation is accounted for by differences in the levels of cycling. Danes travel, on average, about 900 km per year by cycle, whereas citizens of the UK only manage 77 km, with the per capita figures for France, Portugal, and Spain (among others) being even lower (Commission for

Integrated Transport, 2001). The figures also vary significantly between different urban areas even within the same country – some compact urban areas of the Netherlands achieve about 50% modal share for cycling. It is important to note that for many European countries the majority of cycling trips are utilitarian trips to work, shopping, or education, in sharp contrast to the USA, where recreational trips dominate.

The EU is keen to promote cycling, especially in urban areas, through the encouragement of action at local level (Dekoster and Schollaert, 1999). In a recent White Paper the European Commission (2001) did recognize that current trends in transport cannot go on as they are, and current policy is to try and break the link between traffic growth and economic growth. Cycling and walking are mentioned in the White Paper, but mainly in terms of their ability to feed public transport rather than as a mode in their own right. The White Paper does not contain any targets for levels of walking or cycling across the EU.

The Netherlands is often thought of as a country where cycling has certain natural advantages due to the flat topography, relatively high population density, and compact pattern of development, and also because of a historical affinity for the bicycle. However, it is also an example of a country where sustained policy intervention has managed to reverse a historic decline in cycle use. During the 1970s and 1980s the Dutch government changed from a pro-car policy to recognizing the role that cycling could have in reducing the congestion and environmental problems caused by car use. This culminated in the Dutch Bicycle Master Plan (1990), but this was seen very much as part of a wider transport policy that aimed to restrain (car-based) mobility while improving and encouraging alternatives to the car. This has involved measures to reduce the attractiveness of driving (e.g. by increasing car-parking fees), land use planning to encourage compact development that is accessible without using a car, and high-quality provision for the cycle. In the Netherlands there is a dense network of cycle routes and facilities that are designed to be coherent, direct, attractive, safe, and comfortable for cyclists. The aim is to produce facilities that give the cyclist an advantage over car users for short-distance journeys and are safe and therefore attractive. Many of these facilities are segregated from the rest of the road network and have some priority over it at intersections. The sheer number of cyclists and the prevailing cycle-friendly road user culture mean that cyclists are accorded a certain respect and their rights are observed.

In Denmark, over 30% of trips are walked or cycled, with much higher levels of walking and cycling in many medium-sized urban areas. Danish guidance (published by the Danish Roads Directorate) includes a vast array of different measures to help make cycling more attractive, but the same document also emphasizes that "sticks" have to be employed to discourage car use at the same time (Jensen et al., 2000). These complementary measures include removal of car-parking spaces, road closures, and reductions in local speed limits. Measures to directly encourage cycling include improvements in the road layout for cyclists and the creation of logical, direct routes, but also promotional activities, financial incentives, and training for new cyclists. Danish guidance therefore suggests that the optimal strategy to encourage people to shift from car use to cycling is to combine promotional activities with physical improvements for cyclists as well as restrictions on car use.

The importance of urban planning to ensure journey distances are short is also emphasized in Danish guidance on cycling, because of a clear relationship with levels of cycling and walking. Danish urban planning aims to concentrate employment in town centers and to encourage smaller, more local shops. In Denmark it is very difficult to get planning permission for food outlets of more than 3000 m^2 and non-food outlets of more than 1000 m^2 (Jensen et al., 2000).

In recent years there have been a number of interesting developments in non-motorized transport policy in the UK. These are covered in more detail in the following sections because they illustrate policy development in a situation where a conscious effort has been made to encourage non-motorized modes in order to meet wider transport policy objectives. A number of interesting issues are raised about walking and cycling provision and the priority these modes should be given. This is in a country which is not as car dependent as the USA but does not have the level of walking and cycling in near-neighbor northern European countries despite sharing many other similarities.

3.1. Development of walking policy in the UK

Until the 1990s there was little explicit UK central government recognition that walking required consideration beyond simply providing facilities. The dominant consideration was safety (by which was meant the reduction or elimination of accidents), which led to a segregationist design philosophy. As the pedestrian was usually seen as the less important road user, this often meant that pedestrian convenience was sacrificed in order to remove the vulnerable pedestrian from the danger. Segregation tended to lead to the pedestrian having to forego convenience in favor of motorized modes. The most extreme form of this approach was pedestrian subways and footbridges, but this philosophy also underlies the use of guardrails and pelican and other light-controlled crossings.

At the beginning of the 1990s the impossibility of catering for ever increasing amounts of motorized traffic forced a reappraisal of transport policy, and it was suggested that a mix of new approaches was necessary. This was endorsed by the publication of the UK government's White Paper on transport (UK Department of the Environment, Transport and the Regions, 1998), which included an emphasis on the importance of walking and sought to make it a more attractive mode with the aim of encouraging people to walk instead of drive for short journeys. Late in 2000, a scrutiny committee of the UK government decided to examine the government's record on walking. Its concerns about walking and the importance it has been given in UK transport policy are of wider significance.

The committee's report was published in 2001 (Environment, Transport and Rural Affairs Committee, 2001). Overall, the report was critical of the government's policy on walking, and made a number of detailed recommendations, including a greater level of spending. However, the committee's strongest recommendations concerned changing the underlying philosophy of pedestrian provision, giving priority to walking and the establishment of a national walking strategy with targets. The committee explicitly challenged the "prevailing orthodoxy of accident reduction" that underlay the design of much current infrastructure for pedestrians; instead it suggested a policy of danger reduction because "It can be more effective in reducing pedestrian casualties, lead to better urban design and is more convenient for pedestrians." The committee explicitly challenged the philosophy underpinning the design of pedestrian facilities, and suggested separating out what had formerly been thought of as coincident concerns. These were the emphasis on accident reduction and design that was attractive to pedestrians. What the committee was implying was that the emphasis on providing for pedestrians should be shifted from an understandable desire to protect pedestrians for their own sake, with the unfortunate side-effect of making walking less convenient, toward reducing the danger from motorized vehicles at source. This might mean measures to reduce traffic volumes and speeds and redesigning the street environment to make it clear that the pedestrian is the most important road user. Underlying this approach is the idea of a hierarchy of road users that puts pedestrians at the top, and many UK local authorities have adopted this as a policy. The crucial aspect of any such hierarchy, however, is the way it is used and the weight it carries within the local authority applying it. For those planning the urban streetscape this requires a real commitment to the pedestrian and a willingness to inconvenience the motorist if required.

The committee was also concerned about the wider urban environment, the aesthetics of the street, and urban renaissance. In many ways these are concerns that people are more able to understand and respond to than "walking as transport," and they are therefore more politically important.

In its response to the committee's report the government was in agreement with many of the points made, and agreed to publish a national strategy for walking but not to set national targets for walking. The government was, however, not willing to countenance the redesign of the street environment suggested by the committee's more radical recommendations.

Since then, there has been continuing interest in the development of pro-walking policies and the promotion of walking. However, there has not been the support for walking at the national level that was expected after the publication of the White Paper, partly perhaps as a result of the fear of adverse publicity that an explicitly pro-walking policy initiative might incur (Goodman and Tolley, 2003). On the other hand, there is a clear recognition of the importance of a high-quality public environment.

3.2. Development of cycling policy in the UK

Cycling in the UK declined by about 80% from the early 1950s to the 1970s, and cycling is now a relatively minor mode in the UK, with only about 1.5% of trips and less than 1% of total travel (UK Department for Transport, 2003). This is in marked contrast to a number of other northern European countries that are otherwise similar to the UK, and suggests that if the conditions were right more journeys could be made by cycle. In the early 1990s the official attitude to cycling changed from one of seeing cycling as a road safety "problem" to a realization that cycling could play a part in the new approach to transport policy (McClintock, 2002).

This change in attitude culminated in the publication of the National Cycling Strategy (NCS) (UK Department of Transport 1996), which was developed by the UK Department of Transport in close consultation with a wide variety of representatives from a range of different backgrounds. The strategy adopted a headline target of doubling the number of cycle trips from 1996 to 2002 and doubling the number again by 2012. The strategy made it clear that the main aim was to replace short car trips with cycle trips.

Despite an increased interest in cycling at both the local and national level, the publication of useful guidance on designing facilities for cyclists, and efforts to change attitudes to cycling, progress so far in terms of actually increasing the amount of cycling has been patchy, and the number of cycle trips has not increased significantly over the lifetime of the NCS. Local authorities have also had significant problems in actually monitoring the levels of cycling in order to assess progress against the targets. Since the publication of the NCS and the realization that the first target would not be met, the longer-term target to increasing the number of cycle trips has been revised to a trebling of cycle trips between 2000 and 2010.

The quality of facilities for cyclists also varies greatly from place to place. Road space in the UK is often constrained, particularly in urban areas, and making space for cycling often involves taking it from motorized users. Some local authorities are more ready to do this than others, and many facilities seem to be rather token in their approach, with narrow cycle lanes, sharp turns, disconnected routes, lack of maintenance, and lack of attention to detail. Overall there is a lot of work going on at local level aimed at promoting cycling and providing facilities, but an unwillingness to inconvenience motorists means that the actual facilities provided for cyclists are often a compromise and sometimes reinforce the drivers' perception of cyclists as second-class road users. Cycling still has a somewhat poor image in the UK, and tends to be regarded in many areas as unusual, eccentric, and potentially dangerous, which also tends to undermine its status as a legitimate transport mode (Davies et al., 1997).

4. Discussion

Regardless of where policy-making is carried out, a number of important issues often emerge when considering walking and cycling. Many of the issues are obvious, but they are often overlooked as the emphasis of transport planning is often on motorized modes, and non-motorized modes have very different characteristics.

4.1. Differences from motorized traffic

Walkers and cyclists vary widely in their abilities in terms of speed, acceleration, and so on, and in terms of their sensitivities to different aspects of the environment, the most obvious example being feelings of safety (Davies et al., 1997). This is much more so than for motorized transport, which is more homogeneous in terms of abilities and motivations. Most motor vehicles are easily capable of the urban speed limits, and are simply trying to get from A to B as efficiently as possible. In contrast, pedestrians usually have the space and the motivation to display a wide range of different behaviors, including abruptly stopping and changing direction, or simply standing around. Cyclists and walkers are very sensitive to apparently minor diversions and detail in the streetscape, and they often take advantage of gaps in the traffic, especially if in a hurry. Pedestrians often use detailed body language to facilitate their efficient movement in proximity to other pedestrians.

This range of different pedestrian and cyclist behavior is related to the variety of journey purposes (the rushing commuter versus the window shopper), but pedestrians and cyclists also vary widely in their physical abilities. In a sense, while the infrastructure required is expensive and technically complex, motor vehicles are much easier to cater for because they tend to behave in much more predictable ways (at least in the urban context). A corollary to this is that many of the measures commonly used to facilitate the movement of motorized traffic, such as the use of lanes, traffic signals, large roundabouts, and so on, has undermined the position of pedestrians and cyclists because they tend to impose a more ordered way of using the street environment. Walkers and cyclists proceed at a more leisurely pace, exist at a smaller scale, and tend to use the road in a more informal, "chaotic" manner, which inevitably brings conflict with less maneuverable motorized users. The traditional way to solve this has been to constrain non-motorized behavior by, for instance, the use of guardrails to guide pedestrians to common crossing points. This, however, often works against the way that walkers would otherwise choose to use the road (Sauter, 2003).

As a result of their sensitivity to distance, walking and cycling thrive where trips are shorter, which implies higher-density, less-dispersed development. Motorized transport (public and private) implies different patterns of land use and a different scale of development, because higher speeds allows things to be more spread out, but also because more space is needed for the necessary access roads, parking, etc. This difference in scale makes it very difficult for walking and cycling to be viable alternatives in car-based developments (Jensen et al., 2000).

Walking and, to some extent, cycling are often seen as very much part of the purpose of the trip and not just a way of getting there. Walking in particular is often seen more as a way of interacting with the local environment, a way of seeing and being seen, rather than simply a way of getting from origin to destination. This is perhaps a reason for the apparently greater political concern for high-quality public spaces than for the plight of the pedestrian as transport.

For all modes of transport, people's choices are not only affected by physical factors but also by socio-cultural influences and social norms. This might be more so for walking and cycling, with their perceived low status leading to them not even being considered as options. Goodman and Tolley (2003) suggest that positive social and individual perceptions of the car and negative perceptions of walking may be as important as physical barriers, and the same might be even more true of cycling. For many people in the UK, cycling lies outside the mainstream, requiring special clothes and equipment, and it is associated with low social status and perceived as being used by an unusual minority (Davies et al., 1997). Car use, in contrast, is often associated with wealth, status, and power, and is seen more widely as "normal" behavior (Goodman and Tolley, 2003). This implies that encouraging people to walk and cycle more will require more than the physical barriers to be addressed, though the amount of effort required to change attitudes should not be underestimated. In some European countries, simple observation suggests that cycling has a more respectable "image." This must be partly as a result of it being an everyday form of transport for more people and a wider cross-section of society. This in turn is associated with a more utilitarian attitude to bicycle design and use. In contrast, a survey of UK cyclists and non-cyclists suggested that images of cycling were often negative in terms of encouraging people to cycle, and included historical, sporting, and eccentric images. In contrast, the same people had more positive "normal" images of cyclists in the Netherlands, Denmark, and Germany (Davies et al., 1997).

One of the main concerns of potential walkers and, especially, cyclists is safety, and it is important to realize how important safety appears to be as a deterrent to non-motorized modes. Sensitivities to this issue, however, vary considerably between different groups, as does willingness to accept inconvenience in return for a safer journey. Apprehension about safety dominates the concerns of would-be cyclists, but it is possible that to some extent this masks other reasons for not cycling, such as the physical effort involved and the low status of cycling as a mode of transport.

Evaluating the success, or otherwise, of a policy to increase the amount of walking or cycling requires the use of effective monitoring. Cycling and walking can be difficult to monitor because they often take place away from the traditional sites used for traffic monitoring, namely major roads. They also have different trip patterns and trip purposes and are difficult to monitor using automated methods. Users also often behave in somewhat unconventional ways compared with the orderly progress of motorized vehicles. Walkers can easily take unusual short cuts, and cyclists often behave in ways that more conventional vehicles cannot and which are difficult to categorize. Partly, of course, this is because we have become accustomed to considering the behavior of motorized vehicles as "normal," and because the street environment has often been adapted to facilitate their efficient movement.

4.2. The importance of sensitive design

Walkers and cyclists are much more sensitive to the details of design. They notice things more and are more affected by small features, because of their slow speed and the very personal means of locomotion. Some of the factors that make a facility successful might be specific to a location. What works in one place might not work in another. Walkers and cyclists are also more sensitive to anything requiring the expenditure of extra energy, such as diversions and hills. This sensitivity is recognized, for instance, in the design guides for European cycle facilities (Jensen et al., 2000). Cyclists in particular are very sensitive to loss of momentum, which is an important design consideration when deciding how off-road facilities should interact with the road network. A facility next to a main road that requires the cyclist to give way at intersecting side roads or to pedestrians is significantly less attractive in this respect than the main road (Fajans and Curry, 2001). Insensitively designed facilities often reinforce the perceived second-class status of walking and cycling by giving way to motorized modes or simply giving up when needed most. The design challenges are significant, but a well designed facility can act as a promotional tool for cycling and walking, whereas a badly designed one can reinforce prejudices and even be positively dangerous. Good design for walking and cycling may require that other modes are inconvenienced in order to provide a coherent, connected route. There is significant current debate over whether cycle facilities should be segregated from the road or whether on road facilities are a better solution. While high-quality segregated facilities may apparently offer the most desirable solution, it is often impossible in a constrained road system to offer the width, directness, and priority desired. Substandard segregated facilities can be less convenient than using the normal roadway for cyclists, and they can also undermine the cyclists right to use the road in the eyes of motorists. It has become fairly clear in the UK that encouraging cyclists to share space with pedestrians can cause problems due to the potential danger to pedestrians and the substandard nature of such facilities for cyclists. Cyclists tend to be different from pedestrians in that they travel at higher speeds and therefore are more concerned about sight lines, obstructions and surface. On the other hand, hard and fast rules are difficult to come up with, due to locational differences and constraints and the wide variety of different types of users and their concerns. For example, segregation from motorized traffic may be far more important for a facility that will be used predominantly by school children than for a fast commuter route.

The design challenges are significant and require a different approach from that taken for motorized modes. This implies specific training to address the needs of walkers and cyclists properly. Assessments about user willingness, for instance, to trade convenience for segregation from motorized vehicles need to be made carefully, and an assumption that walkers and cyclists should automatically defer to motorized modes should be avoided, even when this is for their own safety.

5. Conclusions

It is clear that there are sound reasons for wanting to encourage more people to walk and cycle. It is also clear that consideration of walking and cycling cannot be separated from wider transport policies and from other policies that affect transport and land use.

Facilities are important, but they have to be well designed in terms of coherence, directness, attractiveness, and safety. Detailed design can be crucial, and requires different skills from those used in constructing infrastructure for motorized vehicles. Creating attractive facilities for walkers or cyclists may involve taking space and priority from motorized modes. As well as directly benefiting walkers and cyclists, this can encourage the transfer of trips from car by improving the relative attractiveness of walking and cycling. However, challenging car use in this way might pose political problems. There is evidence from Europe that cycling can increase even in a situation of high levels of car ownership, and the same may be true of walking. However, this requires the robust application of pro-walking and -cycling measures along with complementary policies in a coherent integrated package.

Successful promotion of walking and cycling will also involve working with other agencies at a national and local level to ensure that walking and cycling are taken seriously. Agencies such as the police, education authorities, and tourist, leisure, and health agencies can also help in the promotion of cycling and walking. They can help to ensure that walking and cycling are not regarded as second-class forms of transport. Concerns over levels of fitness and obesity may become very important in promoting pro-cycling and -walking policies. Appealing to individuals' self-interest may be very important in persuading people to walk and cycle more.

Measures to restrain car use are important complementary measures because they can improve the relative position of walking and cycling, which are to some extent "in competition" with the car for short journeys. Enforcement of speed limits and speed reduction measures can also help to both make cycling and walking more pleasant directly and make them relatively more attractive. Land use policies are also crucial because they define the scale at which urban development takes place. To encourage cycling and walking these need to facilitate compact, high-density development in locations that are easy for people to cycle or walk to. The quality of public spaces is also important, especially to encourage walking.

Higher levels of walking and cycling have great potential in helping to make urban transport more sustainable, but walking, in particular, is more than just a way of getting there. As well as being the purpose of a trip, walking can play a significant role in the renaissance of urban areas by helping to create a people-friendly city. A cycle-friendly city can also facilitate slightly longer distance access in a way that does not compromise the environment of other people. The creation of safe and attractive conditions for walking and cycling and the design of urban areas to make walking and cycling more natural ways of getting about are not just an add-on, they are an essential part of creating sustainable urban areas for the future.

References

British Medical Association (1992) *Cycling towards health and safety*. Oxford: Oxford University Press.

Cavill, N. (2003) "The potential of non-motorised transport for promoting health," in: R. Tolley, ed., *Sustainable transport – planning for walking and cycling in urban environments*. Cambridge: Woodhead.

Commission for Integrated Transport (2001) *European best practice in delivering integrated transport – key findings*. London: CfIT.

Clarke, A. (2003) "Green modes and US transport policy: TEA-21," in: R. Tolley, ed., *Sustainable transport – planning for walking and cycling in urban environments*. Cambridge: Woodhead.

Davies, D.G., M.E. Halliday, M. Mayes and R.L. Pocock (1997) *Attitudes to cycling: a qualitative study and a conceptual framework. Transport Research Laboratory report 266*. Crowthorne: Transport Research Laboratory.

Dekoster, J. and U. Schollaert (1999) *Cycling: the way ahead for towns and cities*. Brussels: European Commission.

Environment, Transport and Rural Affairs Committee (2001) *Walking in towns and cities, 11th report, Volume 1. Report and proceedings of the Committee*. London: The Stationery Office.

European Commission (2001) *White paper – European transport policy for 2010: time to decide*. Luxembourg: European Commission.

Fajans, J. and M. Curry (2001) *Why bicyclists hate stop signs. Access magazine 18*. Berkeley: University of California.

Goodman, R. and R. Tolley (2003) "The decline of everyday walking in the UK: explanations and policy implications," in: R. Tolley, ed., *Sustainable transport – planning for walking and cycling in urban environments*. Cambridge: Woodhead.

Interface for Cycling Expertise (2000) *The significance of non-motorised transport for developing countries, strategies for policy development. Final report*. Washington, DC: World Bank.

Jensen, S.U., T. Andersen, W. Hansen, E. Kjaergaard, T. Krag, J.E. Larsen, B.C. Lund and P. Thost (2000) *Collection of cycle concepts*. Copenhagen: Road Directorate, Danish Ministry of Transport.

Komanoff, C. and J. Pucher (2003) "Bicycle transport in the US: recent trends and policies," in: R. Tolley, ed., *Sustainable transport – planning for walking and cycling in urban environments*. Cambridge: Woodhead.

McClintock, H. (2002) "The development of UK cycling policy," in: R. Tolley, ed., *Sustainable transport – planning for walking and cycling in urban environments*. Cambridge: Woodhead.

Pucher, J. and L. Dijkstra (2000) "Making walking and cycling safer: lessons from Europe," *Transportation Quarterly*, 54:25–50.

Pucher, J. and L. Dijkstra (2003) "Promoting safe walking and cycling to improve public health: lessons from the Netherlands and Germany," *American Journal of Public Health*, 93:1509–1516.

Pucher, J., C. Komanoff and P. Schimek (1999) "Bicycling renaissance in North America? Recent trends and alternative policies to promote bicycling," *Transportation Research A*, 33:625–654.

Sauter, D. (2003) "Perceptions of walking – ideologies of perception," in: R. Tolley, ed., *Sustainable transport – planning for walking and cycling in urban environments*. Cambridge: Woodhead.

Transport Canada (2002) *Transportation in Canada 2002. Annual report*. Ottawa: Transport Canada.

UK Department of the Environment, Transport and the Regions (1998) *A new deal for transport: better for everyone*, Cm3950. London: The Stationery Office.

UK Department for Transport (2003) *Transport statistics Great Britain 2003*. London: The Stationery Office.

UK Department of Transport (1996) *The national cycling strategy*. London: Department of Transport.

Whitelegg, J. and G. Haq (2003) "The global transport problem, same issues but a different place," in: J. Whitelegg and G. Haq, eds, *The Earthscan reader on world transport policy and practice*. London: Earthscan.

Chapter 35

FLEXIBLE TRANSPORT SYSTEMS

JONATHAN L. GIFFORD
George Mason University, Fairfax, VA

1. Introduction

In a dynamic economy the demands on a transport network will never be entirely known in advance. Breakthroughs in production, supply chain, and distribution, and the innovation of new products and services will invariably bring to bear demands on a transportation facility or system that its designers did not expect. The capacity of a transportation system to respond to such demands is a product of its flexibility. Generally speaking, flexibility is a favorable attribute. One of the strengths of the automobile highway system is its capacity to accommodate a highly diverse set of vehicles – some general purpose such as private automobiles and light trucks; some highly specialized for particular purposes such as hazardous materials transport – with a wide range of origins and destination. As long as a product can be fitted on a truck trailer, even if it requires special oversize or overweight permits, it can be delivered to its destination without the requirement of a specialized transportation facility for that purpose.

Flexibility is not unambiguously favorable, however. In some cases, a transportation facility serves as a complimentary asset. For example, consider a gas station located adjacent to a highway interchange. The interchange confers access to the gas station for highway users. If the interchange were removed from service, the value of the gas station and the parcel it occupies lose value. The flexibility to close the interchange is not viewed as favorable by the owner of the parcel or gas station. Similarly, a real estate developer may value proximity to a bus stop, which is fairly easy to relocate, differently from proximity to a light rail or heavy rail station, which is much more difficult to relocate.

This chapter begins with a narrative discussion of the issues surrounding flexibility. It then provides a set of numerical examples showing how conventional net present value analysis would be modified by using the theory of option value to formally represent the opportunity cost of waiting to invest, patterned after examples from Dixit and Pindyck (1994).

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

2. Predict and provide

Traditional approaches to transportation planning and decision-making are founded on the concept that the decision to locate and size transportation facilities should be based on estimates of future demand for the facility. The rationale is that long-lived facilities should be designed to accommodate future traffic. The size design criterion in a highway case, for example, is primarily the number of lanes to be provided. This problem can be broken down as follows:

(1) In a sparse network where alternative paths are not readily available:
 (a) determining the size of a replacement for an existing facility (bridge, highway link, etc.);
 (b) determining the location and size of a new facility.
(2) In a dense network with competing routes:
 (a) determining the size of a replacement for an existing facility (bridge, highway link, etc.);
 (b) determining the location and size of a new facility;
(3) Determining the location and size of a new network of facilities.

Situation 1(a) is conceptually the easiest. Service exists, and a facility is to be upgraded or replaced. Historical data, perhaps in combination with forecasts of population, employment, and other determinants, provide a basis for extrapolating a future design level of demand, which then provides the basis for the size determination. A refinement of this approach could include some form of staged construction, whereby provision is taken to build to a lower design traffic volume (fewer lanes), with an option to expand in the future. One complication is that if service quality (speed, reliability, comfort, etc.) improves (or declines) significantly, demand may increase (or decrease) accordingly. Correctly forecasting such a change would require a more complex forecast that takes into account how demand varies with service quality.

Situation 1(b) is somewhat more difficult. Because no service currently exists, no historical traffic data provide a basis for a forecast. Thus, the forecast must be based on a model of demand that represents the relationship of demand with other indicators of economic activity. A refined model could reflect particular economic sectors. Again, staged construction could be used to hedge uncertainty about future demand.

In a dense network – situations 2(a) and 2(b) – the presence of multiple possible routes introduces a further complication. Demand on the facility of interest is subject to service quality and price on competing facilities, as well as factors affecting secular demand. For an existing facility (situation 2(a)), historical traffic levels provide guidance, whereas for a new facility, traffic estimates must rely on more comprehensive models.

The design of a wholly new network of facilities poses the greatest uncertainty. Because of the fundamental change presumably offered by the new network, prediction of traffic flows is highly uncertain. To the extent that the new network offers services comparable to those available through other networks, demand for the new services may be easier to predict.

The difficulty of predicting future demand for future facilities is the underlying rationale for flexibility. The greater the uncertainty, the greater the value of the option to defer expansion until future demand materializes, utilizing strategies such as phased construction.

A shortcoming of the predict and provide approach is that it fails to recognize that supply may beget demand. Demand and supply are simultaneous processes, whereas predict and provide treats demand as essentially exogenous. Critics of predict and provide argue that the proper policy decision should be the level of travel demand that is desirable on a broad social basis, rather than a narrow engineering focus on what demand will materialize if a particular facility is built. On the other hand, most market economies do have difficulty directly managing transportation demand, since it is a product of social and economic processes that are deeply embedded in the private realm.

3. Control

As suggested by the mention of demand management above, one way to reduce uncertainty is to impose control on a situation. For transportation systems, options for control generally fall under the rubric of demand management. In US air traffic control, for example, controllers hold traffic on the ground when congestion due to weather is expected. Pricing is another often mentioned form of demand management, and also a variety of auctioning arrangements have been considered to optimize the allocation of a fixed supply (in the short term) of facility capacity. In a situation where facility capacity is not fixed – such as a planning problem where facility capacity is to be increased – control can be viewed as an alternative to increased capacity.

The extent to which control is available as an option for reducing uncertainty hinges on the inherent predictability of the factors giving rise to demand and capacity, and on societal and cultural norms and traditions about the acceptability of imposing controls. For example, using currently available techniques, weather is subject to some inherent uncertainty. For another example, the transportation demand requirements of a military conflict is subject to inherent uncertainty, since an enemy's actions and the performance of one's own forces is not entirely predictable.

Cultural limitations on the acceptability of demand management are exemplified by the practice in some Scandinavian countries of pegging highway speeding penalties to the driver's income, which would likely not be acceptable in the USA,

Table 1
Concepts of flexibility

Term	Defintion
Plain language	1a. Capable of being bent or flexed; pliable. 1b. Capable of being bent repeatedly without injury or damage. 2. Susceptible to influence or persuasion; tractable. 3. Responsive to change; adaptable: a flexible schedule (a)
Context-sensitive design	Flexible application of highway design standards to incorporate sensitivity to the surrounding community or natural context (b)
Flexible machines	Robots and computerized numerically controlled (CNC) machine tools that can produce new products or old products produced in new ways. Reduces the need to scrap fixed capital in order to produce new products (c)
Flexible manufacturing systems (FMSs)	An integrated combination of flexible machines to create small-scale "manufacturing cells" or computer-integrated manufacturing (CIM). FMSs may integrate phases of the production process, including design (using computer aided design, CAD), manufacture or distribution. Examples of distribution in FMSs include "just in time" production systems (c)
Flexible specialization and integration	A product strategy of small batch production of specialized products, which may be economical where there is a large range of specialized products over which to amortize the cost of FMS equipment, which is usually more expensive than conventional equipment (c)
Flexible accumulation	1. Sociologists and political scientists use the term to describe a national economic strategy using new technologies, products and services and "skill-flexible" labor
	2. Geographers use the term to refer to a "sea change" in the way capitalism works. "It rests on a startling flexibility [in] labour processes, labour markets, products, and patterns of consumption, … characterized by the emergence of … new sectors of production, new ways of providing financial and business services, new markets, and … greatly intensified rates of commercial, technology and organizational innovation" (D. Harvey, cited in (c))
Flexible regions	Regions with diversified industrial bases can adapt to changing economic conditions, whereas regions dependent on a single industrial sector can suffer when that industry is in a period of decline (d)
Military	One tries to maintain tactical and strategic flexibility on the battlefield by having logistics operations that can support a broad range of tactical and strategic options. There is a logistic element, and intelligence element. "Strategy must have at its disposal a variety of weapons and forces so that the particular combination most suitable to the situation, as it actually arises, may be quickly formed and swiftly and decisively employed in an appropriate manner" (e)

Table 1
Contd

Term	Defintion
Anthropology	Behavioral adaptation in response to changing environmental circumstances. Flexibility is the counterpart to adaptation. Without it, no adaptation; without adaptation, death or decline
Biology	Genetic and developmental adaptations to changing environmental circumstances
Management	Flexibility is a response to uncertainty about future conditions and the expectation that better information will be available in the future. Organizations may adopt incremental approaches to problems when there is uncertainty or conflict about goals or objectives. Organizations may also strive to be flexible in their response to problems and to be "learning organizations" (f)
Land use and urban planning	Future land uses are difficult to predict and control, so flexibility is important for the economic vitality of a region. Planned communities can be sterile and lack the vitality of vernacular, unplanned communities (c)
Austrian economics	Only decentralized decision making (through markets) can make best use of the information available in the economy. Capital goods are fossils of plans based on partially or wholly unrealized expectations. Prediction is inherently imperfect. What is important is that capital goods be redeployed as quickly as new information becomes available
Transaction cost economics	The decision to make or buy a good or service depends on the "specificity" of the good or service. If the good or service can be readily described and captured in a contractual document (using industry standards, specifications or plain language), it may be efficiently produced and traded in markets. If it is unique or has multiple intangible elements, it may be more efficiently produced internally

Notes: (a) *The American Heritage Dictionary*, 3rd edn; (b) US Federal Highway Administration (1997); (c) Gertler (1988); (d) Malecki (1991); (e) Eccles (1965); (f) Quinn (1978).

or the tradition in the USA of taxing motor vehicle fuels at levels far below most European countries.

4. Flexibility

The alternative to addressing uncertainty with control is flexibility. The concept of flexibility is one that is considered in a broad domain of inquiry, from geography to urban planning (Table 1). In transportation policy, two sorts of flexibility are particularly important. First is flexibility in the design and operation of a facility. Staged construction, mentioned above, is an example. The other sort of flexibility is in the decision process itself. Transportation planning processes, especially those lodged in the public domain, often involve funding and permitting decisions that are by tradition inflexible. Each of these is discussed in turn below.

Design and operational flexibility are closely tied to the basic technology of a transportation system. The nature of concrete being what it is, it is difficult (or expensive) to move a lane of highway once it has been put down. In that sense, the infrastructure itself is highly inflexible, and is often treated in economic terms as a sunk cost. The salvage value of the broken concrete is negligible, although the right of way often has considerable value. It may be parceled off, as in the case of Stapleton International Airport in Denver, Colorado, which was converted into a mixed residential and retail development after the city's new airport opened in 1995. Or the right of way may be redeployed for use by another transportation mode, as in the case of abandoned rail rights of way being converted into bicycle trails.

Design flexibility can anticipate future expansion or modification of a facility or system. For example, staged construction (as mentioned earlier) of a highway may entail constructing bridges and overpasses to allow the future addition of one or more lanes. A rail line may be designed so as to easily accommodate a future addition of an interlocking (switch) without rebuilding the existing line. Or a busway may be designed so as to easily accommodate conversion to light rail in the future.

Similarly, operational flexibility might include changes in the operation of a facility to accommodate future needs. Initial design can anticipate such future operational changes and build in the capacity to more easily accommodate them.

Clearly the particular technological features of a transportation facility or system would limit its flexibility. Buses don't fly. But considerable design and operational flexibility may exist within technological constraints.

Design standards and norms can operate to constrain design flexibility. In the area of highway design, for example, Interstate design standards in both urban and rural areas sometimes limit the ability of designers to adjust designs to mitigate impacts on surrounding areas. Such standards govern lane widths, shoulder widths, gradients, curvatures, and so forth. While such standards often are grounded in solid research and experience, the mere existence of a standard, even if it allows deviation in special circumstances, can act as a strong deterrent to design flexibility. Agency concerns about tort liability attributed in litigation to a deviation from a design standard is a particularly conservative force.

The US Federal Highway Administration, with strong support from interest groups concerned about the impact of excessive adherence to design standards, issued a report in 1997 on flexibility in highway design (Table 1). The matter remains controversial, however, as one man's flexibility may be another's dereliction of public safety. A similar phenomenon obtains in the area of operational practice. "Traffic calming" initiatives intended to reorder the balance between autos, pedestrians, and cyclists are viewed by some as placing pedestrians and cyclists in harm's way.

Flexibility in the decision process would seemingly be much easier to accomplish. After all, decision processes do not face the problems of the fundamental physical limitations of sound design and operation. However, the organizational and institutional constraints on decision making often impose severe limitations on decision flexibility. This is because many planning and decision-making processes are explicitly designed to reduce flexibility and lock in commitments as they proceed.

As a project moves through a planning process, it gains (or fails to gain) commitments of resources and approvals from various permitting bodies. The nature of such decisions is such that as a project moves forward and clears the series of hurdles and approvals it requires, new information that might condition or modify those approvals is essentially locked out of the process. Furthermore, the advocates and beneficiaries of a project can become increasingly filled with a sense of entitlement to the project.

Part of the reason for this dynamic is that communities and advocates often coalesce around particular projects. Planning processes, while they seek to apply rational, objective criteria to project selection and programming, often involve horse trading and back scratching. So once a deal is made to trade support of project A for support of project B, it can be very difficult to abandon project A or B if new information arrives indicating that one or the other is a turkey.

In a purely rational project selection and programming process, firm decisions about project commitments would be deferred until the last possible moment, in order to make a final decision on the basis of the best possible information. If projects could be broken into phases or stages, commitments to sequential stages could be similarly deferred until the last possible moment. Such "just in time" infrastructure decisions would build out the network incrementally following demand.

Such an approach flies in the face, however, of the notion that transportation facility investments should be used to guide development, rather than being guided by it. There is a substantial constituency that advocates the use of infrastructure investment to foster development of a particular form or functionality. Such a master planning approach views the infrastructure plan as one of a suite of planning tools that includes zoning and development codes and public facility investment (schools, parks, public buildings, etc.). The use of these tools to develop a community plan then provides guidance to private sector developers on what kind of development a community desires. Developers apply market tests on the viability of such development and opt to proceed or not under the terms offered, or in some cases to request revisions of codes and conditions.

Under such master planning approaches, the government commitment to invest public resources may play an important role in private sector decisions about where and how to invest. Public commitments to locate highways, interchanges, transit stations, and other transportation facilities, as well as expectations about how such facilities will be operated, may be critical. Investors often look closely at such commitments when evaluating prospective development projects. Hence the inflexibility of government decisions becomes a favorable attribute of a private investment.

5. Flexibility and the theory of options

The significance of flexibility in decision-making and design is therefore quite important. Transportation facility planning is, like most investment decisions, subject to some uncertainty, largely irreversible, and hence the timing of the investment may be critical. Yet traditional approaches to transportation planning and decision making have resembled a master planning approach that only crudely incorporates flexibility. The theory of option value used widely in financial markets accounts for irreversibility, uncertainty, and timing by recognizing that the option of waiting to invest or decide is an opportunity cost that may be significant. Option theory values this opportunity cost explicitly and incorporates it into the decision to invest now or to wait for better information. This section works through some simple examples to illustrate how option value can depart from the orthodox theory, adapting examples from Dixit and Pindyck (1994). The section is adapted from Gifford (2003).

Consider an agency that is trying to decide whether to build a highway segment. The investment is completely irreversible – the highway can only be used to carry traffic, and if no traffic materializes, the agency cannot recover its investment. For simplicity, assume that the highway can be built instantly for cost I, and that trips using the highway are worth P = \$1 each. Currently the expected traffic is 200 trips, but next year the traffic will change. With probability q, it will rise to 300, and with probability $(1 - q)$, it will fall to 100. The traffic will then remain at this level forever. While these are strong assumptions, the conclusions they imply are fairly robust and are developed more fully in Dixit and Pindyck, from which the example is adapted.

The general formula for the conventional net present value (NPV) is

$$\mathrm{NPV} = -I + \sum_{t=0}^{\infty} \frac{D_0 P}{(1+i)^t}.$$

Assume a risk-free interest rate, i, of 10%, with I = \$1600 and q = 0.5. Assume also that demand, D, is unrelated to what happens to the overall economy.

Given these values, should the agency build the highway today, or should it wait a year and see whether the traffic goes up or down? Investing now, noting that the expected value of traffic is 200, yields

$$\mathrm{NPV} = -1600 + \sum_{t=0}^{\infty} \frac{200}{1.1^t} = -1600 + 2200 = \$600.$$

It appears that the NPV of this project is positive. The current value of the highway segment, V_0, is equal to \$2200, which exceeds the \$1600 cost of the highway. Hence, it would seem desirable to go ahead with the highway under orthodox theory.

This conclusion is incorrect, however, because the calculations ignore the opportunity cost of investing today rather than waiting and retaining the option of not investing should traffic fall. To illustrate this, calculate the NPV of waiting one year and then investing only if traffic increases. In this case,

$$\mathrm{NPV} = 0.5 \times \left(\frac{-1600}{1.1} + \sum_{t=1}^{\infty} \frac{300}{1.1^t} \right) = \frac{850}{1.1} = \$773.$$

(Note that there is no expenditure and no revenue in year 0. In year 1, the \$1600 is spent only if traffic rises to 300, which will happen with probability 0.5.) Waiting a year before deciding whether to invest in the highway yields a project NPV of \$773, whereas it is only \$600 for investing in the highway now. Clearly it is better to wait than to invest right away.

Note that if the choice were only between investing today or never, it would be better to invest today. If there were no option to wait a year, there would be no opportunity cost of killing that option. Likewise, if it were possible to liquidate the \$1600 investment and fully recover the funds, it would make sense to invest today. Three things are necessary to introduce an opportunity cost into the calculation: irreversibility, uncertainty about future demand, and the option to wait.

The value of the option is simply the difference in the two NPVs: \$773 – \$600 = \$173. That is, one would be willing to pay up to \$173 more for an alternative that includes the option.

5.1. Varying construction cost

An alternative way to value this flexibility is to ask how much more one would be willing to pay in construction cost I to have an investment that is flexible rather than inflexible. To answer this question, find the value of I, call it $\bar{I}$, that makes the NPV of the project when waiting equal to the NPV when I = \$1600 and investing now. Substituting $\bar{I}$ for the \$1600 and substituting \$600 for the \$773 in the previous equation:

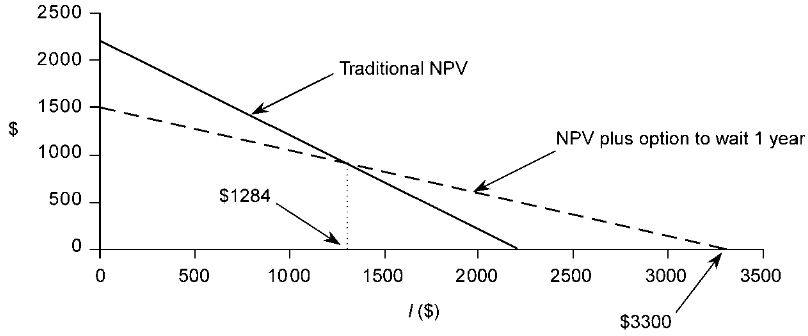


Figure 1. Traditional NPV and option value.

$$\mathrm{NPV} = 0.5 \times \left(\frac{-\bar{I}}{1.1} + \sum_{t=1}^{\infty} \frac{300}{1.1^t} \right) = \$600,$$

$$\bar{I} = \$1980.$$

In other words, the opportunity to build a highway now and only now at a cost of \$1600 has the same value as an opportunity to build the highway now or next year at a cost of \$1980.

Two additional values of I are noteworthy. First, there is a value of I, I', at which it makes sense to invest today, that is, the option of waiting for a year sinks to zero. To find that value, equate the NPV of starting now with the NPV of waiting for one year and solve for I':

$$0.5 \times \left(\frac{-I'}{1.1} + \sum_{t=1}^{\infty} \frac{300}{1.1^t} \right) = -I' + \sum_{t=0}^{\infty} \frac{200}{1.1^t},$$

$$I' = \$1284.$$

Second, there is a value of I, I'', at which it makes no sense to invest even if demand rises to 300. To find that value, equate the NPV of waiting for one year to zero, and solve for I'':

$$0.5 \times \left(\frac{-I'}{1.1} + \sum_{t=1}^{\infty} \frac{300}{1.1^t} \right) = 0,$$

$$I'' = \$3300.$$

Hence, depending on the value of I, we should either undertake the project immediately if $I < 1284$, wait for a year to see if demand rises to 300 if $\$1284 < I < \$3300$, or abandon the project altogether if $I > \$3300$, as illustrated in Figure 1.

5.2. Varying demand

One can also examine how the value of the project and the option to wait a year vary with the level of demand. In the original example, the expected value of demand, D_0, was 200, since there was a 50:50 chance that demand would rise to 300 or fall to 100 in year 2. At what value of D_0, call it D_0', would one be willing to invest now even if demand were to drop by 50% in year 2? Intuitively, this means that the benefit to be earned in year 0 by investing now is sufficient to offset the value of the option to wait and see what happens to demand in year 2. Therefore, equate the NPV of waiting with the NPV of investing now and solve for D_0':

$$0.5 \times \left(\frac{-1600}{1.1} + \sum_{t=1}^{\infty} \frac{1.5 D_0'}{1.1^t} \right) = -1600 + \sum_{t=0}^{\infty} \frac{D_0'}{1.1^t},$$
$$D_0' = 249.$$

Similarly, to determine what value of D_0, say D_0'', would justify bypassing the project completely, set the NPV of waiting for one year to zero, and solve for D_0'':

$$0.5 \times \left(\frac{-1600}{1.1} + \sum_{t=1}^{\infty} \frac{1.5 D_0''}{1.1^t} \right) = 0,$$
$$D_0'' = 97.$$

Hence, if $D_0 > 249$ invest immediately; if $97 < D_0 < 249$, wait for a year and invest if the demand rises; and if $D_0 < 97$, abandon the project.

5.3. Increasing uncertainty over demand

One may also vary the extent to which demand might rise or fall. The original example assumed demand would rise or fall by 50%. How does the value of the option vary if that variance rises or falls? Assume the probability of a rise in demand, q, is 0.5, but that the demand will either rise or fall by 75% in period 1. Then

$$\text{NPV} = 0.5 \times \left(\frac{-1600}{1.1} + \sum_{t=0}^{\infty} \frac{1.75 \times 200}{1.1^t} \right)$$
$$= \$1023.$$

The value of the option is \$1023 – \$600 = \$423, compared with the \$173 in the original example. Why does the value of the option value increase with the variance of future demand? Because it increases the upside potential payoff from the option, leaving the downside risk unchanged, since we will not exercise the option if demand falls.

5.4. The "bad news principle"

An important implication for infrastructure investment of option value is the "bad news principle." By modifying the example by allowing both the probability of an upward move in demand, q, as well as the magnitude of the upward and downward moves to vary, it is possible to determine how "good news" (an upward move) and "bad news" (a downward move) separately affect the critical demand level, D_0', that warrants immediate investment. (In the foregoing examples, the magnitude and probability of upward and downward moves increased and decreased together.) It turns out that the critical demand level, D_0', depends only on the size of the downward move, not on the size of the upward move. The reason is that the ability to avoid the consequences of "bad news" makes waiting desirable.

Suppose that the initial demand level is D_0, but in period 1 the demand level becomes

$$D_1 = \begin{cases} (1+u)D_0 & \text{with probability } q, \\ (1-d)D_0 & \text{with probability } (1-q). \end{cases}$$

Keeping the cost of the investment at I, the NPV of investing now is

$$\begin{aligned} \text{NPV} &= -I + D_0 + q\sum_{t=1}^{\infty}\frac{(1+u)D_0}{1.1^t} + (1-q)\sum_{t=0}^{\infty}\frac{(1-d)D_0}{1.1^t} \\ &= -I + 10[1.1 + q(u+d) - d]D_0. \end{aligned}$$

On the other hand, the NPV of waiting is

$$\text{NPV} = \frac{1}{1.1}\{q\ \max[0, -I + 11(1+u)D_0] + (1-q)\ \max[0, -I + 11(1-d)D_0\}.$$

Clearly, the point of indifference between investing now and waiting is in the range of D_0 where investment in period 1 is warranted if the price goes up but not if it goes down. In this case, the NPV of waiting simplifies to

$$\text{NPV} = \frac{q}{1.1}[-I + 11(1+u)D_0].$$

Equating the NPV of investing now and the NPV of waiting and solving for D_0 gives

$$D_0' = I\left(\frac{1}{1.1}\right)\left(\frac{0.1+(1-q)}{0.1+(1-q)(1-d)}\right).$$

This equation has one important detail: D_0' does not depend in any way on u, the size of an upward move. It only depends on the size of a downward move, d, and the probability $(1-q)$ of a downward move. Also, the larger is d, the larger is the critical demand level, D_0'; it is the magnitude of the possible "bad news" that drives the incentive to wait.

Moreover, if the current net benefit can be negative and a society is contemplating a costly disinvestment or abandonment of a project, the bad news principle turns into a good news principle: the size and probability of an upturn are the driving forces behind the incentive to wait to disinvest.

5.5. *Scale versus flexibility*

Economies of scale can be an important source of cost savings. By building a large facility instead of two or three smaller ones, society may be able to reduce its average costs and increase social well-being. This suggests that society should respond to growth in demand for services by bunching its investments, that is, investing in new capacity only infrequently, but adding large and efficient facilities each time.

What should society do when there is uncertainty about demand (as there usually is)? If society irreversibly invests in a large addition of capacity, and demand grows only slowly or even shrinks, it will find itself holding capital assets it does not need. Hence, when growth of demand is uncertain, there is a tradeoff between scale economies and the flexibility that is gained by investing more frequently in small increments to capacity as they are needed.

This problem is important in road building, and transportation generally. It is generally much cheaper per unit of capacity to build a four-lane facility than it is to build a two-lane facility. At the same time, road building authorities face considerable uncertainty about how fast the demand for a road will grow. Hence, it is important to be able to value this flexibility.

To adapt the earlier example, assume that the highway segment in question can be built either in a four-lane configuration at a cost of I_0 = \$1600 (option A), or in a two-lane configuration at a cost of I_0 = \$1000 (option B). If built in a two-lane configuration, it could be expanded at a cost of \$1000 to four lanes in year 2. Demand is uncertain. It is presently at 200, but with a 50:50 chance it may grow to 300 in year 1 or remain at 200.

The NPV of the four-lane option is

$$\mathrm{NPV_A} = -1600 + 200 + \sum_{t=1}^{\infty} \frac{250}{1.1^t} = \$1100.$$

The value of the two-lane option is

$$\begin{aligned}\mathrm{NPV_B} &= -1000 + 200 + 0.5 \times \sum_{t=1}^{\infty} \frac{200}{1.1^t} + 0.5 \times \left(\frac{-1000}{1.1} + \sum_{t=1}^{\infty} \frac{300}{1.1^t} \right) \\ &= -800 + 1000 - \frac{500}{1.1} + 1500 \\ &= \$1245.\end{aligned}$$

Hence, uncertainty about demand growth makes the option to build incrementally preferable, even at a loss of economies of scale. The total construction cost of option B is

$$1000 + \frac{1000}{1.1} = \$1909,$$

compared with the \$1600 of option A, a difference of \$309, or 16%, and the option value present in option B is \$145 (\$1245 – \$1100).

Extensions to this logic include varying the probability of demand increasing or decreasing, extending the uncertainty beyond a single period, and incremental projects, where the first few steps of a project reveal information necessary for assessing its total value, as in the case of preliminary engineering studies. These are not explored in detail here but are available in Dixit and Pindyck (1994).

5.6. Investing under uncertainty: summary

The examples used here to illustrate the importance of option value are subject to many simplifying assumptions. Dixit and Pindyck (1994) relax these assumptions, and assert the following general insights from the theory of investing under uncertainty:

- The value of the option to wait can be significant and may justify waiting rather than proceeding immediately.
- The option value increases with the value of the sunk cost, *I*.
- The option value increases with the degree of uncertainty over future demand (i.e. the variance of future demand), the downside risk being the most important aspect.
- Conditions that merit killing an option to wait and proceeding immediately with a project often imply conventional NPVs considerably higher than unity.
- "When a project consists of several steps [and] the uncertainty about it will be revealed only as the first few steps of the project are undertaken, ... then these first few steps have information value over and above their contribution to the conventionally calculated NPV. Thus it may be desirable to start the project even if orthodox NPV is somewhat negative."
- "Investment on a smaller scale, by increasing future flexibility, may have a value that offsets to some degree the advantage that a larger investment may enjoy due to economies of scale."

These general insights from the option theory have profound implications for infrastructure investment.

6. Public policy implications

This chapter has examined the issues of flexibility in the design and operation of transportation facilities, first with a narrative discussion of the issues and then with a set of numerical examples showing how conventional net present value analysis would be modified by using the theory of option value to formally represent the opportunity cost of waiting to invest, patterned after examples from Dixit and Pindyck.

This discussion raises a number of important policy implications. Traditional master planning approaches to infrastructure development have generally followed a "predict and provide" approach. Increasingly, observers (and many critics) have come to recognize that those predictions are somewhat self-fulfilling, because development and traffic often follow the infrastructure.

Two possible responses to this condition suggest themselves. Given that demand is endogenous, one response is to subject forecasts to a public process and not treat them as purely technical activities. A second response is to move towards a more "just in time" infrastructure provision process, waiting to deploy infrastructure until investment criteria that recognize the option value of waiting indicate that it is favorable.

The first approach is more consistent with a view of public facility planning as a tool for guiding private investment. The latter is more consistent with a view of it as a market driven process. In practical terms, a pure version of either approach is not feasible. Master planning requires private sector participation and if the forecasts are incorrect, no matter how devoutly wished by the participants, private capital will not materialize. Pure "just in time" infrastructure provision is also not feasible. While possibly appropriate for determining the scale of facilities, it provides little guidance on whether to deploy wholly new systems. The decision to deploy an interstate highway system, for example, does not afford a lot of room for flexibility. However, the pace and scale of that deployment might well be informed by the tools of flexibility discussed here.

References

Dixit, A.K. and R.S. Pindyck (1994) *Investment under uncertainty*. Princeton: Princeton University Press.

Eccles, H.E. (1965) *Military concepts and philosophy*. New Brunswick: Rutgers University Press.

Gertler, M.S. (1988) "The limits to flexibility: comments on the post-Fordist vision of production and its geography," *Transactions of the Institute of British Geographers*, 13:419–432.

Gifford, J.L. (2003) *Flexible transportation systems*. Oxford: Pergamon.

Malecki, E.J. (1991) *Technology and economic development: the dynamics of local, regional and national change*. New York: Wiley.

Quinn, J.B. (1978) "Strategic change: "logical incrementalism," *Sloan Management Review*, fall:7–21.

US Federal Highway Administration (1997) *Flexibility in highway design*. Washington, DC: US Department of Transportation, Federal Highway Administration.

This Page Intentionally Left Blank

Chapter 36

STIMULATING MODAL SHIFT

PETER BONSALL
University of Leeds

1. Background

1.1. Introduction

This chapter is written against a background in which the private car is most travelers' mode of preference because it meets their perceived need for convenience, freedom, security, privacy, and status. The use of cars is increasing in most countries, while walking, cycling, and the use of public transport is generally declining. In the UK, for example, the 1990s saw an 11% increase in the distance traveled by car, an 11% reduction in the distance traveled by local bus, and a 20% reduction in the distance walked (UK Department for Transport, 2003).

These trends coincide with reductions in the real (wage rate adjusted) cost of car ownership and use and increases in the real cost of public transport use. These underlying cost trends are likely to continue because technological change and globalization bring reductions in the costs of car ownership and use while the price of public transport is more closely tied to the rising costs of labor. Recognizing this, the UK government's 10 year plan for transport (UK Department of Environment, Transport and the Regions, 2000) assumed that, by 2010, the real cost of motoring would fall by 20% while the cost of bus and rail journeys would remain fairly constant. Nonetheless, the plan sought to achieve a reduction in congestion, a 10% growth in bus passenger journeys, a 50% growth in rail passenger-kilometers, and a tripling in the number of cycle trips. It is against this background of government aspirations to overcome adverse trends that we discuss methods to achieve changes in the modal split.

1.2. Why should we want to alter the modal split?

The objectives of transport policy vary from country to country, and may change over time, but they generally seek to maximize positive aspects such as delivered

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

accessibility while minimizing negative aspects such as congestion, accidents, personal insecurity, pollution, and consumption of energy and other resources. More latterly, the list of desirables has been expanded to include the health-giving aspects of travel that involves physical exercise, while the list of undesirables has been extended to include the loss of activity space and visual amenity in a car-dominated landscape or townscape.

For a given individual, travel by private car tends to score well in terms of convenience, flexibility, time efficiency, comfort, personal security, privacy, and cost. Although the rest of society may benefit from the car user's ability to participate in economic and social activities, they will suffer from the noise, pollution, visual intrusion, accident hazard, and congestion caused by that car user. Public transport can achieve greater efficiencies in terms of passengers per unit of road space, lower emissions, and fewer casualties per passenger-kilometer, but its efficiency advantages will be illusory if it cannot achieve high load factors, and its environmental credentials presuppose an efficient and well-maintained vehicle stock. The individual user can benefit from an attractive public transport service, but personal security may be a problem, particularly when waiting for services or walking to or from the access points. Also, there will always be people for whom, due to infirmity, public transport is not an attractive prospect, and there will always be certain journeys for which it is simply uneconomic to provide a public transport service. Walking and cycling can bring health benefits to the individual and have few negative impacts on the environment or the rest of society, but they are not feasible for long journeys, may have an unacceptable accident risk (in a mixed traffic environment), and may not be available to the infirm. Similar summaries for other modes would highlight the speed of air travel, set against its serious environmental consequences; the efficient use of road space by motorcycles, offset by their poor accident record and noise nuisance; and the convenience of taxis, set against their inefficient use of road space.

It would clearly be an oversimplification to suggest that any one mode is wholly good or bad – although the environmental and road space efficiency arguments have led many transport authorities to adopt targets for reducing car use while increasing use of public transport, walking, cycling, and even motorcycling. Other authorities, recognizing a link between economic activity and travel, take levels of traffic, taxi usage, and air arrivals as indicators of economic success, and would be loathe to see them decline.

Although it would generally be agreed that an increase in trips by public transport, walking or cycling associated with a reduction in trips by the private car is likely to be a good thing, the means of achieving this result may not be acceptable. For example, if an increase in public transport use is achieved only by providing a heavily subsidized service that achieves very low usage levels, the overall efficiency of the service will be compromised and the opportunity cost of the subsidy may be too high. Similarly, if the reduction in the use of private cars is achieved only by restricting people's free choice, it may cause serious inconvenience, perhaps amounting to an injustice, for certain classes of traveler, and, if such people are important to the local economy, it may compromise the vitality of that economy. And, of course, there are potential conflicts between public transport, walking, and cycling and between the different public transport modes – an increase in bus use might be welcomed by bus operators but it would not please transport planners if it came about only by encouraging people to make more journeys or to use the bus instead of walking, cycling, or going by train.

The remainder of this chapter recognizes that attempts to influence the modal shift are likely to be controversial; not only is there room to dispute the desirability of any particular modal mix but, even when a particular goal can be agreed, there will be those who hold that the achievement of a particular end cannot justify the means of achieving it.

The chapter considers methods of influencing modal split under three broad categories: those that discourage or restrict "undesirable" modes, those that improve the attractiveness of "desirable" modes, and those involving publicity or propaganda. Conventional wisdom is that the "sticks" are more influential than the "carrots," and so it is with the former that we will start. The chapter is written primarily in the context of mature and affluent cities, but several of the issues raised will be of relevance to rural areas and to cities in less developed economies.

2. Reducing the use of "undesirable" modes

There are two main methods by which to reduce the use of particular modes: increased taxes or charges; and regulations or physical restrictions. We will deal with each of these in turn.

2.1. Taxes and charges

Governments are likely to favor the use of increased taxes or charges on the purchase, ownership, or use of vehicles: not only are increased taxes and charges relatively simple to introduce and generally very effective, but they can be said to leave the driver with a degree of free choice and can produce a very attractive revenue stream for the authorities. Economists point out that, only by setting charges to reflect the true costs of use, can we expect to maximize overall welfare and achieve maximum system efficiency. The existing levels of charges and taxes vary from country to country. Some barely cover even the direct costs of providing and maintaining the road network, some make a contribution toward the costs of police, hospital, and other services provided for road users, and others again seek to cover the less tangible costs of congestion and environmental impacts that may be attributed to the use of private vehicles. Few governments would admit that the total tax burden on motorists is set so as to discourage usage, even though motorists' organizations invariably claim that their tax contribution far exceeds the costs that they impose on society.

Most countries impose taxes on car purchase. Although this tax is usually in line with that for other goods and services, there are some interesting exceptions. In Singapore (Tanaboriboon, 1993), anyone wishing to purchase a new car is faced with a significant import tax, an "additional registration fee" (currently 160% of the purchase price) and a "quota premium" (currently in excess of S. $26 000 – see Section 2.2). The Netherlands has introduced a €45 bond payable on new cars to cover the eventual cost of recycling the component materials. Several countries, including Hong Kong and Brazil, impose a graduated tax rate to discourage the purchase of large petrol-engined vehicles. Some countries impose special taxes on imported vehicles. Taxes on car purchase can certainly influence the purchase of new cars, but the effect may not always be what was intended. In Australia, the import duty on four-wheel-drive vehicles is lower than that on ordinary cars because they were deemed to be primarily for farm use, but this differential has contributed to four-wheel-drive vehicles becoming the fastest-growing component of the private car fleet. More generally, increased taxes on new cars may delay the scrappage of older, more polluting vehicles.

Taxes on car ownership, usually expressed as an annual vehicle registration fee, vary from country to country, but a typical level example is the £150 annual fee payable in UK. This fee is roughly equivalent to 0.6% of the national average wage and, allowing for differences in purchasing power, it is about a sixth of the fee charged in Hong Kong, about half the fee charged in the Netherlands but about twice that charged in Luxembourg. The use of differential charges for different types of vehicle is increasing – the UK now has four different levels designed to reflect the environmental damage caused by different sizes and types of engine. A substantial proportion of motorists in some countries drive vehicles owned by their employers, and thus escape the taxes on car ownership. This loophole can be closed by treating access to a company-owned vehicle as a taxable component of the individual's salary. One problem with taxes on car ownership is that, while they might discourage purchase or continued ownership, once paid, they represent a sunk cost to motorists and provide no incentive to reduce their use of the vehicle – indeed, having paid the charges, motorists may be intent on getting their money's worth by making full use of the vehicle.

Taxes and charges on car use are more likely to have a direct influence on car use. They may be in the form of fuel taxes, parking charges, tolls, or congestion charges. Fuel tax varies from country to country: in the USA it is about 25% of the pump price, in Greece it is about 50%, and in the UK it recently reached almost 75%. Fuel taxes often discriminate between different fuels – usually with the specific aim of encouraging drivers to switch to fuels that are deemed to be less environmentally harmful or more beneficial to the national economy. Compared with the tax on unleaded petrol, several countries impose higher taxes on leaded petrol and lower taxes on diesel fuel, and several impose very low or zero taxes on biomass fuels and liquid petroleum gas. Increases in fuel tax are administratively simple, cheap to impose, and provide a fairly direct means of discouraging car use. Fuel taxes are a useful instrument because they tend to fall most heavily on the cars that are contributing most to congestion and environmental problems (because larger cars tend to consume more fuel than smaller ones, because cars tend to consume more fuel in congested conditions, and because emissions are broadly proportional to fuel consumed). However, since fuel taxes cannot be varied according to the time or place in which the vehicle is used, they cannot be used to target particular journeys, and increased charges may create serious problems for drivers who have no realistic alternative means of going about their business. Fuel tax increases sufficiently large to discourage drivers in congested urban conditions would cause serious problems for motorists in uncongested, sparsely populated, rural areas where public transport services can never be viable. Thus, fuel taxes would need to be accompanied by other measures to further dissuade car use in congested areas or to compensate motorists in rural areas. Fuel taxes are generally seen as politically unattractive because they have tended to generate considerable public opposition. Increased petrol taxes led to riots in Harare in 1998 and in Lagos in 2003; and fuel tax protests in the UK in 2000 led the government to abandon its previous strategy of progressive increases in fuel tax levels (Harman, 2002).

Parking charges can be an effective means of dissuading car owners from using their cars to travel to congested urban areas or, if imposed in residential areas, may even discourage car ownership in those areas. For example, Cullinane (2003) noted that 8% of young Hong Kong residents said that the main reason for delaying car purchase was the cost and shortage of parking facilities. Government levies on parking spaces can simultaneously reduce demand for parking and raise revenue; for example, in Perth, Australia, a levy on parking in the central business district is used to part-fund free public transport services within the city. The effectiveness of parking charges as a policy instrument is, however, limited if the public authority does not have jurisdiction over all the available spaces and cannot impose a levy on privately owned spaces. Recognition of this problem led the UK government to give local authorities the power to impose levies on employee parking spaces, but it is not yet clear what effect this will have (Bonsall and Milne, 2003). Another, more fundamental, problem with the use of local parking charges is that, rather than switch to other modes of transport, drivers may opt to park somewhere that offers cheaper parking, and, although this may reduce a local congestion problem, it may have a devastating effect on the local economy and may simply transfer the congestion problem to another location. Parking charges have a limited role in demand management unless they can be introduced in a coordinated fashion over a large area.

Tolls have long been imposed on users of particular roads, tunnels, or bridges in order to recoup the costs of providing that facility and as a means of raising revenue. Although the technology used to collect the payments has advanced, the principle behind the charges imposed electronically on users of certain autoroutes in France, freeways in Canada and the USA, and motorways in Japan are no different from those collected in cash from users of roads and bridges in mediaeval Europe. The idea that tolls or charges should deliberately be used to manage the demand also has a long pedigree, but has come very much to the fore in the context of Singapore's Area Licensing Scheme (Holland and Watson, 1978) and London's congestion charging scheme (Greater London Authority, 2001). Singapore's scheme was first introduced in the late 1970s in an attempt to reduce peak hour congestion, it required drivers to purchase a S. $3 per day supplementary license to enter the city center during the morning peak hours. The scheme resulted in a very significant (73%) reduction in car use, and the charge period was subsequently extended to cover the whole working day, with premium charges during the peak periods. Since 1998, the supplementary license has been replaced by a stored-value card that is debited electronically. The London scheme, which was introduced in February 2003, requires drivers to pre-pay UK £5 to use roads within enter an 8 km^2 area of the city center during the working day. Initial indications (Transport for London, 2003) are that it has reduced city center traffic by as much as 38% and has resulted in significant increases in bus use.

The theory behind congestion charging is that motorists should be charged an amount that reflects the full costs that, by using the road, they impose on others. Prominent among these costs are motorists' marginal contribution to congestion and environmental degradation. Provided that they are charged these costs, motorists can make a rational decision on whether to make a journey, and society will benefit whether or not they decide to travel – in one case they would provide revenues that can be used for the public good and in the other case they will no longer be imposing the costs. In practice, of course, it is not possible precisely to match the charges to the costs, but the net effect should still benefit society. The usual objections to congestion charges are that they might cause diversion of trips to other roads that might be ill-equipped to cope with the extra demand and that they might reduce the competitiveness of the area to which they are applied. It is also suggested that the charges would represent a serious problem for those individuals who have no alternative but to drive in the charge area. It has been suggested that the only way to avoid unwanted traffic diversion and loss of competitiveness would be to introduce a universal system, perhaps based on satellite tracking, whereby all journeys could be charged at a rate that results in an optimum distribution of traffic across the network (Commission for Integrated Transport, 2002). Skeptics doubt that, given rivalries between competing towns and cities, there could ever be agreement on what this optimum distribution might be, and they question whether the cost of such a technologically advanced mechanism is warranted.

A pragmatic approach to the problems identified with each of the fiscal instruments identified above might be to reduce the level of fixed costs and charges, including perhaps the motorist's annual third-party insurance costs, to raise fuel taxes by a somewhat greater amount, and then to use the surplus revenue to subsidize the provision of local services in areas not well served by public transport and to provide rebates for drivers who are unable to use public transport by virtue of some physical disability. Interestingly, a recent survey of public opinion in the UK (Royal Automobile Club, 2004) suggested that motorists would prefer to pay for their road use through increased petrol charges than through annual taxes, tolls, or congestion charges.

2.2. *Regulations and physical restrictions*

One of the objections to fiscal measures of restraint is that, although preserving the semblance of free choice, they may be socially divisive and discriminate against the poorer members of society. A more equitable, if less economically efficient, solution may be to apply restrictions to everyone irrespective of their ability or willingness to pay. Regulations are commonly used to forbid the use of unsafe, polluting, or excessively noisy vehicles (by means of construction-and-use regulations, licensing rules, or periodic tests). Regulations and physical restrictions are also commonly used to restrict the use of certain classes of vehicle in particular areas; for example, to exclude general traffic from bus lanes or pedestrianized areas. Many cities have a deliberate policy of reducing the amount of parking or road space capacity available to general traffic. Restrictions in the number of parking spaces can have a direct impact on travel choices but can lead to increases in congestion – due to vehicles searching for spaces or queuing to enter parking lots – and illegal parking. City managers are generally wary of reducing their parking stock for fear of reducing their attractiveness with respect to competing cities or out-of-town locations. When road space is removed from use by general traffic it may be re-designated as environmental amenity space or, more commonly as we shall see later, for use by pedestrians, cyclists, buses, or high-occupancy vehicles. Although many authorities had feared that reductions in road space would lead to increased congestion or diversion of traffic to other routes, recent research (Cairns et al., 1998) has suggested that this is rarely the case. Physical restrictions, typically in the form of speed humps, road narrowing, or street closures, are increasingly used to restrict traffic, or certain types of traffic, in sensitive areas.

Other, less common, examples of the use of regulations to reduce traffic levels include the exclusion of most vehicles from areas designated as low-emission zones (National Society for Clean Air and Environmental Protection, 1999) and the selective exclusion of vehicles from city centers on the basis of their license plate number – the so called "odd-or-even" system whereby odd-numbered vehicles are allowed in only on odd-numbered dates and even-numbered vehicles only on even-numbered days. Restrictions of this kind may be permanent or, as in the case of Athens, may be imposed as a temporary measure during periods of severe air pollution – in which case any impact on mode split may be short lived.

Some countries have regulations to restrict the number of vehicles owned in a particular area. In Singapore, prospective car owners must bid for one of a limited number of "certificates of entitlement" for their new vehicle. Successful bidders in the monthly auctions must pay the "quota premium," which is equal to the lowest accepted bid for each type of vehicle (Tanaboriboon, 1993), and must renew the certificate after 10 years. In Tokyo, people wishing to purchase a vehicle must first obtain a permit, and they can do this only if they can demonstrate that they have somewhere to park the vehicle.

3. Positive encouragement of the use of "desirable" modes

3.1. Provision of facilities and services

The most obvious way to encourage people to use a particular mode is to ensure that it is available and offers a good level of service. In the case of public transport this means ensuring that the services are reliable, swift, comfortable, safe, and clean, that they run when and where required, and provide adequate capacity. In the case of walking and cycling, it means ensuring that the individual can get from origin to destination with maximum safety, security, and comfort and minimum diversion and delay, and, in the case of cyclists, with appropriate facilities for secure storage of cycles and for showering/changing after the journey.

Unfortunately, it is simply not possible to provide this level of service for all potential journeys by all potential travelers. To attempt to do so would imply an investment of resources out of scale with the likely number of users. The conclusion must be that, even though use of public transport, walking, and cycling may be desirable, they are only justified if they can be achieved at reasonable cost. Investment must be targeted where it is likely meet a real need. A too-generous public check book may simply lead to waste and inefficiency. A knowledge of the costs of provision, together with an understanding of travel patterns and their relationship to personal characteristics, activities, and land use patterns, will provide a basis for investment decisions. Data analysis and modeling will help determine what level of usage might be achieved at what cost, and so indicate what facilities and services might best be provided.

Rules of thumb have been devised to indicate what type of public transport system might be most appropriate in different sized cities with different forecasts of potential ridership (Armstrong-Wright, 1993; Chartered Institute of Transport, 1996). The traditional hierarchy puts heavy rail at the top, being the only realistic option where the ridership exceeds 20 000 per hour, and follows through metro, light rapid transit, trams, trolley buses, conventional buses, demand-activated services, and taxis. The advent of various forms of guided bus, busways, and bus priority measures have extended the range of riderships for which the bus is likely to be the preferred solution up into territory that was previously the preserve of rail-based modes (Smith and Hensher, 1998; Mackett and Babalik Sutcliffe, 2003). The inherent flexibility of bus services is an important advantage, and, given a degree of segregation from general traffic, bus services can offer journey times and visibility that approach those of much more expensive solutions.

Whatever version of public transport is under consideration, analysis of costs and potential revenues will usually suggest that the facilities and services be concentrated along major corridors of movement and that the services should be most frequent during peak hours. The revenue per unit cost is certainly greatest for these services, and a purely commercial operator may be well advised to look no further. However, in the absence of feeder services and off-peak services, many potential users will be unable to rely on public transport for all their journeys ,and so may "need" to buy a car, and having done so may decide to use it for all their journeys. Maximum use of public transport is not achieved by relying on commercially justified levels of service. Ideally, the socially necessary but loss-making feeder services would be planned as part of an integrated timetable of services and would be cross-subsidized by the more profitable services. The alternative approach, characteristic of the UK outside London, is for such services to be added in only after the commercial network has been defined, but although this avoids the worst excesses of uncontrolled cross-subsidy, it cannot produce the most effective or efficient pattern of services.

The performance of individual services will, of course, be dependent on the conditions encountered *en route*. Except where they involve street-running, rail services generally have an exclusive right of way, and journey times will depend on the quality of rolling stock, track, and signaling. For street-running rail services and buses, the journey times will depend crucially on local traffic conditions and on the amount of priority over other traffic. The provision of segregated track (for light rapid transit or guided bus), exclusive busways or bus lanes (for conventional buses) can protect public transport from general congestion – although bus lanes may not be effective unless adequately enforced. Dedicated road space and priority at signals not only gives public transport vehicles a very real advantage over general traffic but, because their effect is very visible, the perceived advantage will be even greater, and so the consequences for mode choice are likely to be enhanced. The same general argument applies when priority is extended to all multi-occupancy vehicles: the solo driver becomes acutely aware of experiencing more delay than the higher-occupancy vehicles.

Priority for buses and multi-occupancy cars does not need to extend throughout the route – some advantage can be gained even if the priority is concentrated at bottlenecks. Different arguments apply in the context of facilities for cyclists and pedestrians. A would-be pedestrian may welcome the introduction of stretches of well-made and well-lit pathway and priority crossings but, unless the path is continuous from the origin to the destination, the individual may decide that the journey would still be too dangerous or circuitous on foot. The provision of cycle lanes and priority measures might similarly be welcomed by potential cyclists but might fail to convince them to travel by cycle if they still have to negotiate a dangerous intersection or if there is no provision for cycle storage at the destination. This is not to say that an incomplete provision is of no value, or that a comprehensive network is always justified, but it should be recognized that the usage of such facilities as do exist will, in part, be dependent on the existence and quality of complementary facilities.

This argument can sometimes lead to unexpected conclusions. For example, since non-car owners are more likely to walk, cycle, or use public transport, and since people may be able to forgo purchase of a car if there are sufficient taxis or car club vehicles, the use of public transport, walking, and cycling might actually be promoted by providing taxis or encouraging car clubs. Although it may seem perverse to promote the use of car-borne modes as part of a strategy designed to reduce car use, such provision can be an important part of a low-car-use lifestyle. This argument is certainly not lost on some public transport operators, such as Swiss National Railways, who actively promote the membership of the Swiss car club (Bonsall et al., 2002). There will, of course, be people who decide to use the taxi or car club vehicle in preference to public transport, and this may result in loss of revenue for the operator. If the public transport service thereby becomes non-viable, this could generate problems for the operator, the body responsible for financing the service, or, if the service is withdrawn, the other users. If everyone's needs can be met through taxis and so forth, then this may represent a better use of resources, but there are likely to be some who would have preferred the continuation of the public transport service. There is clearly a delicate balance to be struck when encouraging the use of these "complementary" modes.

3.2. Financial inducements

Many people are clearly dissuaded from traveling by the costs incurred. Price always features prominently among quoted reasons for not using public transport more frequently. Even though there is evidence to suggest that some people quote price as an easy excuse rather than a true reason, and that sensitivity to price is not as high as naive public opinion surveys might suggest, there is ample evidence to show that reductions in price generally lead to increases in use. Evidence suggests that the long-term elasticity of bus use with respect to price is about –0.6 for commuting journeys and about –0.9 for leisure journeys (Balcombe et al., 2004), and thus suggests that a 10% reduction in price will bring about a 6% increase in commuter usage and a 9% increase in leisure usage.

In the context of public transport, price reductions may be in the form of across the board reductions, discounts for use of particular services, e.g. off-peak services, or for particular classes of user, e.g. pensioners or leisure travelers. Although an across the board reduction in price could in certain circumstances produce sufficient new passengers to more than offset the increased operating cost and loss in revenue per head, such price reductions usually cost money and require external subsidy or cost savings elsewhere. If these are not available, then general price reductions cannot normally be contemplated. However, if there is spare capacity in the system, a price reduction might not imply additional costs, and a price reduction targeted at use of the spare capacity need not cost anything. The justification for discounts for particular classes of user may be social; for example, pensioners, job seekers, and children are often identified as deserving of subsidy, and elaborate mechanisms may exist to allow operators to offer reduced ticket prices to these groups and claim the lost revenue back from the government. The offer of discounts to particular groups, however, may reflect a commercial judgement that the recipients of the discount are more price-sensitive and so a price reduction is likely to generate additional passenger numbers that, given spare capacity, might bring a net increase in profit. In practice, of course, the argument is often put the other way round – that business travelers and commuters are less price sensitive and so a price increase aimed at them will not lead to significant loss of patronage.

The offer of discounts for off-vehicle sales, including reduced price or free transfer tickets, and for off-peak journeys can be seen as means of encouraging patterns of use that are less costly to provide. Discounted season tickets and multi-journey tickets are often seen as a means of rewarding loyal customers or encouraging repeat journeys, but, apart from the reduced administrative costs and operational delays that can be associated with off-vehicle sales, there may be little commercial logic in subsidizing customers who use the services at peak times. The logic may, however, become apparent when considering the transport system as a whole; although provision of capacity for peak period users may be costly, it is less costly to provide this capacity on public transport than as extra road space for private vehicles.

Although the most obvious way of reducing the cost of journeys by "desirable" modes may be for public transport operators to reduce their ticket prices, there are other ways. For example; employers might offer their employees loans at preferential rates to enable them to purchase public transport season tickets or cycles, tax authorities might treat these loans as a business expense, and employers and tax authorities might allow pedestrians and cyclists to claim an amount per kilometer traveled on business. All these examples are derived from allowances that, in many countries, have long been available to private car users and that, unless extended to cover all modes, exist as a hidden subsidy to car users.

People who use public transport, walk, or cycle can of course be exempt from tolls and charges, but it is also possible to offer exemptions or discounts to those classes of vehicle or driver whom one wishes to encourage. For example; the Singapore area licensing scheme initially allowed exemptions for vehicles carrying two or more people (Holland and Watson, 1978), tolls on certain highways in the USA apply only to sole-occupancy vehicles (US Federal Highway Administration, 2000), and the London congestion charge allows exemptions for taxis and motorcycles (Greater London Authority, 2001).

4. Marketing

4.1. Basic information and advice

If people are unaware of the existence of modes and services, or if they perceive them not to offer an attractive option, they will be unlikely to use them. If usage is to be encouraged, provision of basic information about public transport routes and services, cycle routes, footpaths, taxi ranks, high-occupancy vehicle lanes, etc., should be seen as a necessary accompaniment to the introduction of such facilities. Where such basic information does not exist, its provision should be given high priority. Revisions to route maps, timetables, fares information, and bus stop signage should accompany any change in the provision of public transport services, just as revisions to maps and signage should accompany any change in the provision of cycle routes and other facilities.

There are, of course, many different ways of providing the information. The cheapest, but least effective, option is to provide the information only when it is asked for. A more proactive approach would involve distribution of posters, timetables, maps, and information leaflets, provision of a telephone inquiry service, information kiosks, and a website. Particular success has been reported with the use of information packs designed for visitors and new residents – people who are least likely to have the background information, who are likely to be actively considering their travel options, and whose travel patterns have not yet become a matter of routine.

A choice also has to be made about the content and sophistication of the information provided. The most basic information relates only to routes, average fares, and frequencies, while the most sophisticated would provide real-time information on service arrival and departure times. The cost of providing real-time information is considerable, although, owing to new tracking technology, this is much less than once it was. Conventional wisdom has been that, in the case of rail-based modes, the cost of providing real-time information is covered by the revenue from extra passengers attracted to the service, but that this was not true for bus services. The recent and expected reductions in the cost of providing this information are likely to make the provision of such information worthwhile even on a purely commercial basis.

Although telephone inquiry services have long been able to provide would-be passengers with individually tailored information or advice, the advent of the Internet has allowed a step-change in the quality and sophistication of such services. It now becomes possible to offer almost instantaneous information on which service to use, where and when to access it, how long the journey will take, and what it will cost. The most sophisticated systems allow inquirers to specify their preferred departure or arrival time, their preferences for journey duration, straightforwardness and price, and will then make a recommendation that might involve use of several different modes and services. The Internet is fast becoming the inquiry medium of choice for large sectors of the population

It should be recognized, however, that the availability of equivalent information for car journeys is resulting in raised aspirations on the part of the customer; unless public transport operators can offer information of the kind now readily available to motorists, the drift away from public transport is likely to accelerate. A number of Internet sites offer advice on how to make a journey by public transport as well as by car. If the inquirer had intended to make the journey by car, this additional information would make him or her aware of the public transport option and might persuade him or her to use it, but if the comparison is not flattering to public transport and if the initial intention had been to use public transport, the net result could be that some potential passengers are lost. Provision of comparative information increases choice, but the result may not always be what was hoped for.

4.2. Presentation and marketing

The importance of effective presentation and marketing of public transport services in an increasingly consumer-oriented society cannot be overemphasized. What constitutes good presentation will vary depending on the nature of the service, but will generally imply high standards of vehicle maintenance and cleanliness, polite and helpful staff, easily understood service and fare structures, effective provision of information, and customer-friendly methods of ticket purchase. Recognition of the importance of these factors has led many operating companies to invest in customer-relations training for their drivers and other front-line staff, and to simplify their service and fare structures.

The London Underground map is often quoted as an example of a successful presentation of a complex network structure, and several operators have sought to achieve something similar for their services – hence the promotion of bus services as an "overground" network and the branding of individual services with names or colors rather than simply with numbers. The importance of clear and distinctive network information has also been picked up by organizations responsible for promoting cycling, and this is reflected in the increased availability of maps showing cycle routes and facilities.

The widespread use of fixed or zonal fares, rather than fares proportional to distance, testifies to the fact that simple fare structures bring operational and administrative savings as well as simplifying life for the customer. However, the attractions of a simple, easily understood, fare structure can be lost if, in order to appeal to niche markets or exploit differing price sensitivities, the operator introduces an array of special ticket types and discounts. The problem is that although each new ticket type or discount may be attractive to its target market, it may have to be so hedged round with restrictions to prevent it being used by groups who are assumed to be prepared to pay the full price that the overall picture becomes confusing and intimidating to the potential traveler. Some operators, struck by the success of low-cost airlines in attracting additional passengers, are experimenting with fare structures featuring a headline-grabbing low price that might not be available for the specific service that the potential passenger actually requires. This is an interesting development although it obviously risks alienating disappointed customers.

Conventional marketing seeks to distribute its message via posters, leaflets, flyers, and advertising in printed publications, and on TV and radio. Although these media, particularly leaflets and flyers, can be targeted toward potential markets – e.g. existing users, business travelers or local residents – the approach is essentially to broadcast the message in the hope that recipients will respond to it. Recognizing the difficulty in getting people to change established patterns of behavior and of overcoming the negative image that many non-users have of public transport, many operators have sought to target their marketing toward people who are likely to be considering a change in their travel arrangements. As noted above, this may take the form of a "welcome pack" for new residents containing service information and free tickets, to encourage them to try out the services on offer. Some success is reported for an even more interventionist approach to get people to adopt a new mode: the provision of "buddies," whose role is to help people learn how to use the new mode. Two such schemes exist in Leeds: a "cycle buddy" scheme whereby an individual who wishes to try cycling for a regular journey is allocated an advisor who helps him or her prepare and then accompanies the individual on the route for a day or two while he or she builds confidence and familiarity with the route; and a "bus buddy" scheme that is designed to assist people with learning difficulties to make independent use of bus services.

Another development that has broader application and has been adopted in several cities involves personalized assessment of an individual's current travel patterns and the provision of practical suggestions of ways in which that individual might reduce their car journeys, perhaps by increased trip chaining, increased use of public transport, increased acceptance of lifts, walking, cycling, or making greater use of telecommunications as an alternative to travel. These suggestions are accompanied by appropriate information on the location of the relevant bus stops, services, and so forth. This type of targeted marketing or advice has been found to work well with that section of the community predisposed to wanting to reduce car use, do not have constraints that make alternatives difficult, and have alternatives available. It is based on the idea that voluntary travel behavior change is more sustainable than a change imposed from outside (Ampt, 2003). It is suggested that, if someone is encouraged to invest his or her own time and resources into considering the benefits of a change in travel arrangements, he or she will have a sense of ownership of the solution, and that a change of behavior prompted by a rational consideration of the impacts on one's own lifestyle, health, and finances is likely to persist longer than one that is made on a whim or out of a sense of obligation to society. Such schemes have reported reductions of 5–14% in the number of car trips, and of up to 14–17% in the distance traveled by car, across the target populations, not just among those actively participating (Brog, 2000; Rose and Ampt, 2001).

These changes have been achieved in a wide range of situations and without substantial improvements to public transport, cycling, or walking facilities. Personalized interventions of this kind are, of course, very resource-intensive, but do appear to be effective. Even with fairly conservative assumptions about durability, voluntary behavior change programs are calculated to yield significantly higher benefit–cost ratios than are typical of more conventional urban transport investments (Ker and James, 1999).

Attempts to change people's behavior by appealing to their sense of social responsibility can be successful (TAPESTRY, 2003), but, at least in the transport arena, they tend to have a limited and short-lived impact. During a national crisis – such as a fuel shortage or environmental emergency – people may be prepared to forego their own convenience and may feel embarrassed to be seen not to be helping, but if the crisis is extended or is perceived as having receded, socially motivated behavior tends to give way to more selfish concerns. It is not easy to find examples of reductions in car use over the medium or long term that have been brought about solely by appealing to drivers' sense of civic duty. Although there is a long tradition of public service advertising that seeks to appeal to people's sense of social responsibility, conventional marketing wisdom is that it can be counter-productive to alienate the audience by appearing to criticize their existing behavior. However, the repetition over the last several years of the message that cars cause environmental problems has perhaps contributed to a change in the climate of opinion that, in certain sections of society at least, has made it *de rigueur* to be seen to want to reduce one's use of the car. Whether this climate of opinion actually encourages a change in behavior is less certain.

Any attempt to market the benefits of public transport, cycling, or walking will face stiff competition from well-funded attempts to sell new cars. Car advertisements pervade the media, and not only extol the virtues of a particular brand or model but also evoke the supposed lifestyle benefits of car ownership and use (Wiltshire, 2003). Attempts to promote the use of alternative modes are unlikely to succeed if they seek to compete on the same ground. They need to identify advantages that are credible and relevant to the target audience – such as the freedom from the need to find a parking space, the ability to sleep, read, or conduct business while traveling, the ability to travel out and back from different locations or by different modes, the ability to drink before traveling, or the ability to travel without a large up-front investment – each of which might appeal to different groups within society (Guiver and Bonsall, 2003).

In practice, public transport marketing tends to be more successful in encouraging existing customers to make more journeys than in attracting new customers out of their cars. For operators, of course, one extra passenger is one extra passenger, no matter what his or her previous mode, and indeed, other things being equal, they will prefer extra passengers to be already familiar with the procedures of buying tickets and alighting the service and to want to travel outside the commuter peak. But, for planning authorities wishing to reduce car use during the peak periods, it seems that conventional marketing of public transport among existing passengers has little to offer. It has recently been suggested (e.g. by Wright and Egan, 2000) that, in order to achieve a significant change in mode use, marketing efforts should focus on discouraging car use rather than on encouraging the use of alternatives.

5. Who can make it happen?

We now turn to the question of who should be involved in efforts to influence mode use – noting that the objectives of different stakeholders may conflict.

Governments or courts generally set the ground rules that determine the powers of the various bodies whose actions might influence modal split. National or state governments will generally give some direction to transport policy and usually retain direct control over taxation while leaving detailed policy and implementation to lower tier authorities. In 2000, the UK Government went further than its predecessors and many of its contemporaries in making it clear that it wanted to influence modal split and in passing the legislation that it thought necessary to achieve this (Bonsall 2000).

In the context of mode split, the main policy levers directly controlled by central government are the taxes on fuel and annual car ownership, the tax status given to company cars, business travel expenses and employee car parking spaces, and the provision of subsidies or tax advantages to public transport operators. The sums of money involved can be considerable. For example, in the UK, the government is currently raising approximately UK £28 billion per year from taxes on motorists while providing an annual subsidy of around UK £1.8 billion to the rail industry and about UK £300 million of support to bus operators. Financial flows on this scale can have a significant impact on the economy and on government revenues, and so it should be no surprise that they are not determined simply with regard to their effect on modal shift!

Central government can influence the actions of local authorities in various ways. First, it can require local authorites to make specified provisions or take specified actions. For example, to provide subsidized public transport for school children or pensioners, to ensure that public transport is accessible to disabled people, to enforce traffic regulations, to monitor environmental conditions, or to set and monitor targets for modal split. Secondly it can encourage local authorities to take particular actions by providing funding or other assistance. For example, local authorities might be given grants to introduce public transport priority measures or facilities for pedestrians and cyclists, or they might receive assistance in the design of road charging schemes or training for travel plan coordinators (see later). Financial support has traditionally been provided in the form of counterpart funding or grants tied to government-approved projects or programs, but, for innovative schemes, it has become common to channel funds into demonstration projects chosen following competition between alternative schemes. The third way in which central government can influence local authorities is by providing powers that facilitate the achievement of particular objectives. For example, local authorities may be given the power to make contracts with public transport operators, to operate services themselves, to borrow money for approved investments, or to impose congestion charges.

Most local authorities have little influence over the costs of car ownership – unless they have the power to impose a local registration tax. Their influence over the costs of car use is often restricted to setting the tariff in publicly owned car parks, although some may be able to impose direct charges or tolls on the users of certain roads, and some may even be able influence fuel prices through their control of local taxes. UK local authorities now have the power to introduce congestion charges and to impose levies on employers who provide parking for their employees, but, in the absence of such powers, local authorities must accept that the costs of car use are fundamentally determined by world oil prices, advances in automotive technology, central government decisions on fuel tax, and the decisions of private sector suppliers of parking space.

Local authorities' main mechanism for influencing mode split is through their role in the provision of infrastructure for the various modes of transport. This role enables them, subject to legal and financial constraints, to influence the relative performance and attractiveness of the various modes, and this is of obvious importance if they wish to achieve a modal shift – for example, to persuade some peak period car drivers to switch to public transport. There is of course a key difference here between those local authorities who have the power to operate public transport services and those who can only seek to influence the actions of commercial operators. Where the local authority is the operator, it can, subject again to legal and financial constraints, offer a service that is designed to meet whatever objectives it sees as appropriate.

Left to their own devices, commercial operators of public transport services can be expected to wish to maximize their net revenue, but are unlikely to choose to maximize ridership, let alone ridership by former car drivers who would otherwise be contributing to peak period congestion. A local authority may have the power to influence a commercial operator's activities under the terms of an operating license, tender, or franchise. This power may, as in London, include the full specification of services, timetables, and fares or, as in the UK outside London, may be limited to inviting tenders to supply specified, socially necessary, services over and above those that the operators have chosen to supply on a commercial basis in a competitive market. It is interesting to note that, in the years since this geographical distinction was introduced in the UK, bus ridership has increased in London but has fallen elsewhere in the country – suggesting that the introduction of a competitive market may not be compatible with a desire to see increased use of public transport. Partly in recognition of this fact, more recent UK legislation (Her Majesty's Stationery Office, 2000) has provided for more cooperation between local authorities and public transport operators. "Quality partnerships,"whereby the local authority undertakes to provide facilities (e.g. bus stops, interchanges, bus lanes and other priority measures) while the operators undertake to provide a specified quality and frequency of service, have proved particularly successful, and are reported to have led to significant increases in ridership (e.g. Huntley, 2001).

Commercially motivated public transport operators will be interested in increasing the profitability of their operations by making appropriate use of new equipment, technology, and procedures. For example, new ticketing systems may reduce costs and improve management information; real-time monitoring of bus locations should improve fleet utilization and allow an improvement in the level of service; and enhanced staff training may improve efficiency and result in improved customer relations. These investments can work to the benefit of the operator and the customer, and may help to change the mode split.

Travelers' mode choice decisions are obviously influenced by the quality of public transport provided, but, if they have access to a car, they are perhaps more influenced by the availability and price of parking facilities. The decisions of employers, retailers, and other service providers to provide or subsidize parking for their employees, customers, and clients can have a profound effect on the relative attractiveness of different modes of transport. These decisions will reflect commercial judgements in the light of land values and market conditions, but can also be directly influenced by local planning authorities. Local planning regulations have traditionally required developers to provide parking spaces, but, recognizing that additional parking brings additional traffic (Shoup, 1999), several countries now allow local authorities to set limits on the maximum amount of parking provided in new developments.

Local authority control over new development can also be used to influence the design and location of new developments, and so influence the nature and mode of the associated traffic (e.g. Carson et al., 1999). The influence of development policy on traffic patterns may be very significant in new or fast-growing cities, but can have an incremental effect in more mature cities. Many countries have regulations that allow the planning authority to refuse permission for new developments that would generate more traffic than could be carried by the existing infrastructure. This concept was extended in the Netherlands, where the "ABC" system classified areas in terms of their transport links as suitable for specific types of development. For example, an area with poor public transport would not be considered suitable for developments likely to attract or generate significant numbers of person-trips (Bakker et al., 1997). The UK has planning guidance (UK Department of Environment, Transport and the Regions, 2001) that seeks to ensure that new developments that are likely to attract or generate significant numbers of person-trips are concentrated at locations with good access to public transport and that new housing is built at densities that are likely to make public transport viable. Another lever available to UK local authorities is the ability to put conditions on a planning consent such that the developer must take specified actions to restrict car traffic or to promote the use of public transport, walking, or cycling at the new site. This might typically involve a requirement on a retailer to subsidize a bus service to the site or on an employer to establish a travel plan among its employees.

Company travel plans provide an interesting example of local action to influence modal split. The concept has been applied in several countries, including the USA, the Netherlands, and the UK. It involves action by individual companies or organizations to discourage solo car use and promote use of public transport, walking, cycling, and car sharing among their employees. Typical company travel plans include the provision of a car-sharing matching service, information about public transport services, discounted tickets, preferential car parking for car sharers, cycle storage and changing facilities, and individual advice on ways of reducing car use. Employers may become involved voluntarily, as a service to their employees, or in the hope that they can reduce their expenditure on parking facilities, but it is clear (e.g. Steer Davies Gleave, 2001) that government "encouragement" can make all the difference. In the USA, the Federal Clean Air Amendment Act (1990) together with matching legislation at the state level, required large employers to have a commute trip reduction plan in place. In the UK, legislation in the late 1990s allowed local planning authorities to require developers to produce travel plans for their sites, and public sector employers have received strong advice that they should be setting an example by establishing travel plans among their own employees.

6. Conclusion

We have seen that many governments are seeking to reverse the trend toward increasing car use and declining use of bus, walking, and cycling, but that objectives differ. We have noted that while on environmental grounds one might wish to emphasize reductions in car use, on health grounds one might wish to emphasize increased walking and cycling, on social grounds one might wish to increase the use of walking, cycling, and public transport, and on commercial grounds one will generally be looking for increases in the use of public transport rather than decreases in car use.

We have identified various approaches by which to influence the modal split, and have noted that, since "carrots" are more popular than "sticks," the public will generally prefer the authorities to seek to influence behavior by enhancing the attractiveness of an existing option or providing a new one, than by taxing, restricting, or otherwise discouraging those options that they deem undesirable. However, we have noted that the sticks are generally more effective than the carrots. Taxation and regulation seem to have been particularly effective, but we have noted that it may not be politically possible to adopt a wholly negative approach. The traveling public is not slow to complain if it thinks that it is being exploited by government or monopoly operators.

If people are to be persuaded to use other modes it is important to provide an attractive, accessible, and visible service for public transport users, pedestrians, and cyclists. Although the problems of funding cannot be overlooked, it is clear that the efficiency of public transport services can be approved through investment in new technology and procedures. We have drawn attention to the need to offer transport services that appeal to identifiable markets, and to make sure that the target users are appropriately informed. We have discussed the role of marketing, and, while not underestimating the challenge posed by the ubiquity of pro-car images and opinions, have noted that it is possible to capitalize on the fact that walking, public transport, or cycling can offer real advantages to some people for some of their journeys. Encouraging such people to make voluntary changes in their travel arrangements can be cost-effective, and, because the response is voluntary, the policy is likely to be politically attractive.

Of course, no single policy is likely to be effective in isolation, and the best option is likely to include a package of measures.

References

Ampt, E.S. (2003) "Voluntary household travel behaviour change – theory and practice," in: *Proceedings of the 10th International Conference on Travel Behaviour Research*. Lucerne: IATBR (http://www.ivt.baum.ethz.ch/allgemein/pdf/ampt.pdf).

Armstrong-Wright, A. (1993) *Public transport in third world cities – state of the art review*, No 10. Crowthorne: Transport Research Laboratory.

Bakker, D., A. van der Hoorn and B. van Wee (1997) "Office suites suit the railways," in: *Proceedings of the 1997 European Transport Conference*. London: PTRC.

Balcombe, R., R. Mackett, N. Paulley, J. Preston, J. Shires, H. Titheridge, M. Wardman and P. White (2004) *The demand for public transport: a practical guide*. Crowthorne: Transport Research Laboratory.

Bonsall, P.W. (2000) "Legislating for modal shift," *Transport Policy*, 7:179–184.

Bonsall, P.W. and D.S. Milne (2003) "Urban road user charging and workplace parking levies," in: J. Hine and J. Preston, eds, *Integrated futures and transport choices*. Aldershot: Ashgate.

Bonsall, P.W., A. Jopson, A. Pridmore, A. Ryan and P. Firmin (2002) *Car sharing and car clubs: potential impacts*. Final report to DTLR and the Motorists' Forum. London: Commission for Integrated Transport (http://www.cfit.gov.uk/mf/reports/carclubs/report/).

Brog, W. (2000) "Switching to public transport," *Proceedings of Second UITP Asia Pacific Congress*. Melbourne: UITP.

Cairns, S., C. Hass-Klau and P. Goodwin (1998) *Traffic impact of highway capacity reductions: assessment of the evidence.* London: Landor.

Carson, G., M. Dix, J. Callaghan and R. Slevin (1999) "Encouraging sustainable development by linking public transport accessibility, new parking standards and developer contributions," *Traffic Engineering and Control*, 40:370–379.

Chartered Institute of Transport (1996) *Better public transport for cities*. London: CIT.

Commission for Integrated Transport (2002) *Paying for road use*. London: CfIT.

Cullinane, S.L. (2003) "Attitudes of Hong Kong residents to cars and public transport: some policy implications," *Transport Reviews*, 23:21–34.

Huntley, T.P. (2001) "Partnership for the future," in: A. Grayling, ed., *Any more fares? Delivering better bus services*. London: Institute for Public Policy Research.

Greater London Authority (2001) *The Mayor's transport strategy*. London: GLA.

Guiver, J. and P.W. Bonsall (2003) *Marketing public transport*, Working paper. Leeds: Institute for Transport Studies, University of Leeds.

Harman R. (2002) "Fuel tax protests and fuel rationing – a historical perspective," in: G. Lyons and K. Chatterjee, eds, *Transport lessons from the fuel tax protests of 2000*. Aldershot: Ashgate.

Her Majesty's Stationery Office (2000) *The Transport Act 2000*. London: HMSO.

Holland, E.P. and P.L. Watson (1978) "Traffic restraint in Singapore," *Traffic Engineering and Control*, 19:14–17.

Ker, I. and B. James (1999) "Evaluating behaviour change in transport: a case study of individualized marketing," in: *Proceedings of the Australian Research Forum*. Perth..

Mackett, R. and E. Babalik Sutcliffe (2003) "New urban rail systems: a policy-based technique to make them more successful," *Journal of Transport Geography*, 11:151–164.

National Society for Clean Air and Environmental Protection (1999) *Low emission zones*. Cedar Rapids: NSCA.

Rose, G. and E.S. Ampt (2001) "Travel blending: an Australian travel awareness initiative," *Transportation Research D*, 6:95–110.

Royal Automobile Club (2004) *Report on motoring 2004*. London: RAC.

Shoup, D. (1999) "The trouble with minimum parking requirements," *Transportation Research A*, 33:549–574.

Smith, N. and D. Hensher (1998) "The future of exclusive busways: the Brazilian experience," *Transport Reviews*, 18:131–152.

Steer Davies Gleave (2001) *The take-up and effectiveness of travel plans and travel awareness campaigns*. London: Department of Environment, Transport and the Regions.

Tanaboriboon, Y. (1993) "Demand management implementation in South East Asia," *ITE Journal*, 63:21–28.

TAPESTRY (2003) *Travel awareness, publicity and education supporting a sustainable transport strategy in Europe* (http://www.eu-tapestry.org).

Transport for London (2003) *Central London congestion charging scheme three months on*. London: TfL (http://www.tfl.gov.uk/pdfdocs/cc/cc-three-month-report).

UK Department for Transport (2003) *Personal travel factsheets*. London: DfT.

UK Department of Environment, Transport and the Regions (2000) *Transport 2010 – The Ten Year Plan*. London: DETR.

UK Department of Environment, Transport and the Regions (2001) *Transport: planning policy guidance note 13*. London: DETR.

US Federal Highway Administration (2000) *Report on the value pricing pilot program*. Washington, DC: FHWA.

Wiltshire, P. (2003) "How lifestyle marketing keeps cars the number one travel choice," *Local Transport Today*, 374:8–9.

Wright, C. and J. Egan (2000) "Demarketing the car," *Transport Policy*, 7:287–294.

Chapter 37

MARKETING PUBLIC TRANSPORT

GENEVIEVE GIULIANO and SARA HAYDEN
University of Southern California, Los Angeles, CA

1. Introduction

This chapter provides an overview of marketing strategies aimed at increasing public transport ridership. We start from the premise that marketing is a broad concept; it encompasses any policy or strategy intended to change individual behavior, in this case convincing more people to use public transport more frequently.

Increasing transit ridership has become ever more challenging, as rising real per capita income and decentralization of metropolitan areas make the private vehicle ever more attractive. These trends not only make public transport a less suitable substitute for the car, but, as ridership declines, subsidies must increase to maintain the level of service. Governments typically respond by reducing service and raising fares, setting off a cycle of decline. In the USA, these trends led to the near demise of the public transport industry in the 1960s. Since then, massive investment and subsidies resulted in a bottoming out in the 1970s, and modest increases in ridership since. More recently, 2002 was the first year in which transit ridership increased more rapidly than the population. In European metropolitan areas, public transport has retained a larger market share, and in cities where public transport is of particularly high quality (e.g. Zurich or Munich) significant increases in market share have occurred (Transportation Research Board, 2001).

Some key factors have come together to make transit more competitive. First, there is a growing public perception that solving congestion, air quality, and other environmental problems will require increased reliance on public transport. This has led to broad political support for investment in public transport in most developed countries. Second, advances in information technology provide new tools for improving service, integration across services, and flexibility in pricing and service structure. For the first time in decades, technology provides opportunities for transforming traditional, fixed-route service. Third, there is a growing realization among transport providers that attracting choice riders (those who have a car) requires a customer perspective. Marketing tools have therefore become more important now than in the past.

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

Where does marketing fit in? The decision to use public transport depends on two primary factors: service quality and price. Service is not simply route coverage or frequency; it also includes ease-of-use, information availability, safety, and many other factors. Likewise, price is not simply the transit fare but also includes time costs, the forgone comfort and convenience of driving one's car, as well as the value of reduced stress afforded through travel by transit. Thus, marketing may focus on informing the public of the availability and quality of transport services or on strategies that make the price more attractive. A third strategy is market segmentation, or providing specific services tailored to specific markets. The remaining sections of this chapter discuss these three categories of marketing.

2. Informing the public

Conventional transit marketing takes many forms, but has two primary purposes: informing the public about specific services offered by an agency and benefits derived from transit provision by the community, including both riders and non-riders. Providing the former information can induce ridership among people needing the specific services publicized. Providing the latter can improve the image and visibility of transit in a community, leading to greater numbers of people willing to consider transit as a transportation option, or at least to support it as a valuable community resource.

Educational programs can fulfill both informational purposes. An employer-based strategy of providing specific transit information to newly hired workers has proven to be very effective, decreasing car use by about 5%. Many transit agencies employ representatives who provide training and information at local companies.

Many agencies have programs for school-aged children. Such programs may include transit agency speakers who visit local schools, or the free use of regularly scheduled transit service for field trip transportation. One innovative program is offered by Kitsap Transit in Washington State in the USA, which developed a curriculum designed to introduce sixth-grade students to ecological issues connected with car use and the solutions offered by alternative transportation. After studying the curriculum in class, the students take a field trip to the agency for a tour. Participating teachers rate the program as excellent, and student travel pass sales have increased since the program began (Texas Transportation Institute et al., 1999). Educational programs can familiarize children with and give them a positive view of transit, increasing the likelihood that they will use transit as adults.

Programs aimed at new residents also perform an educational function, informing new residents about services that are available and accessible to them. Portland TriMet in the USA contracts with a local utility to send out a packet of information to all new residents. The packet includes three free day tickets as an inducement to ride. Moving is a particularly good time to attract riders, since old commuting patterns have been broken but new habits have not yet been formed. Taking a different tack, the Capital Area Transportation Authority in Lansing, Michigan, places employees strategically throughout the system to act as guides and answer questions. Having begun as a program to orient new university students to the system, the program has contributed to seven straight years of ridership increases (Starcic, 2003a).

Free passes are useful in inducing non-riders to try transit. Ben Franklin Transit in Richland, Washington, offers free rides system-wide on Wednesdays and Saturdays. Increased ridership on the free days has carried over to the rest of the week, leading to a 14.6% growth over the first year of the promotion. Pierce Transit, also of Washington State, did a 1 day promotion entitled "Step off the Gas, Step on the Bus" in which any passenger who presented a gas receipt at boarding was given a free ride, thus reinforcing the expense of driving and transit as an economical alternative (Texas Transportation Institute et al., 1999). Free passes go one step beyond the goal of informing potential passengers of specific transit services by giving them first-hand riding experience.

Advertising is an essential technique in transit marketing. Available media include print, broadcast, direct mail, the Internet, and various out-of-home media, such as billboards, bus advertising, or advertisements in movie theaters. Frequency is vital in getting a message out via advertising; transit agencies often implement campaigns that use a variety of media simultaneously. A 2003 awareness campaign, "Pulling Together," run by MARTA in Atlanta, employed television commercials, radio spots, print advertisements, and outdoor advertisements to communicate its message.

Positive messages that focus on benefits enjoyed by the transit customer or community have the most impact. For instance, the above-mentioned MARTA campaign carried the message that transit is good for the economy and improves traffic, air quality, and quality of life in the region. A campaign developed for nine counties in Washington State focused on the stress reduction provided by the use of alternative transportation; using radio spots, billboards, newspaper advertisements, and transit space. The campaign yielded 75% brand recognition/awareness upon completion (Robinson Research, 1999). Both of these campaigns were based on the premise that building awareness of a transit agency is a prerequisite to selling specific transit services. Advertisements conveying specific service-related information are also produced; however, there are fewer of these due to the complexity and amount of information that this generally involves.

A new marketing strategy that employs personal contact to provide detailed, customized service information is called individualized marketing (Brog, 2003). The process begins with a survey that is used to segment respondents into groups based on their willingness to use transit or alternative transportation. Those who are interested are contacted directly, either by telephone or through a visit, and provided only with the information and help they request. The dialog is tailored to individual interests and maintained only as long as the interest continues. This strategy has proven effective in European and Australian trials, where it reduced car usage by 10–14%. A pilot program in Portland in the USA has yielded an 8% reduction in car trips.

2.1. *Changing the image of public transport*

Public transport – especially bus transport – has a poor image in the USA. It is often seen as the mode of last resort, used only by the very poor, the disabled, or the frail. Hence, many bus transit providers seek to change the image of their service by using special vehicles or designing logos to identify special services. In Santa Monica, California, the "Big Blue Bus" has become the marketing image, identifying buses as always clean, on-time, and available at low fares. The subtle message is the contrast with the much larger regional operator, whose service does not have these positive attributes. Buses designed to look like early twentieth-century trolley cars are used for tourist routes along shopping or entertainment streets in many destination cities. In Stockholm, light-rail service on a major tourist route is provided with restored train cars. The new express bus service in Los Angeles is called "Metro Rapid," and has an entirely different logo and color theme from the regular route service; the bus stops have special designs – with the practical purpose of speeding up boarding time – that further differentiate the service. The intended message is that the new service is a new kind of transit, used by commuters and shoppers.

2.2. *Information availability*

The complexity of most urban transport systems can be a significant barrier. Prospective travelers must acquire a large amount of knowledge in order to know which train or bus to take, where to transfer, how much to pay (and in what form), etc. Providing more accessible information is a practice that has grown exponentially in usefulness due to the advent of Internet technologies. Transit agency Web sites are unparalleled as a means of information dissemination, and can be used both to provide detailed user information, such as routes/schedules and public relations information, as well as new service information, press releases, etc.

At this point in time, most, if not all, US transit agencies have a Web site available to inquirers. Fare information is the most frequently available feature, offered by 88% of transit Web sites. Other common features are route schedules (81%) and route maps (50%) (Multisystems, 2003). As experience with Web capabilities has grown, new functions have been added by some agencies that have greatly enhanced the Web site's marketing efficacy, such as online itinerary planning.

Advances in technology should lead to more ways to make information easily available. For example, cell phones incorporating Global Positioning System (GPS) technology should make it possible for travelers to obtain real-time route guidance. This would make using the system much easier for tourists, or for people making a new trip.

Itinerary planning

Online itinerary planners use Web-based interfaces that allow a transit user to input origin and destination points and receive back detailed travel directions/ instructions. Besides the origin and destination, programs ask users for their preferences regarding time (date and time of departure or arrival), itinerary characteristics (fastest, fewest transfers, lowest price, minimal walking) and maximum walking distance. Planner outputs are generally written directions, including itinerary fares and schedule; some transit agencies also offer detailed walking maps and localized transit system maps for points of destination and embarkation.

Online itinerary planners at many agencies have grown out of telephone operator-based trip planning call centers. Trip planning call centers have been tremendously useful in assisting riders, increasing the propensity of travelers to choose transit for new trips. Compared to call centers, Web-based planners have the advantage of being in 24 hour operation, allowing users to retrieve information at their convenience. Web-based planners also eliminate the potential for phone lines to reach capacity, leading to busy signals and users being unable to find information at times of heavy system use. Being fully automated, online planning programs can also reduce the need for hiring telephone operators to answer calls.

Web-based trip planners also provide a convenient way to integrate the networks of multiple transit agencies within a greater metropolitan region. The Metropolitan Transportation Commission (MTC) in California's San Francisco Bay Area operates an itinerary planner (Multisystems and Coogan, 2003) that includes 19 transit providers in nine counties. Integrating provider information simplifies the travel process and effectively expands the area over which transit users can travel, making transit a more useful mode for all users, both those who ride by choice and by necessity.

Online trip planners are very useful in encouraging transit use by tourists. User interfaces that allow the entry of landmarks as origin and destination points are particularly helpful in this context, when potential users are much less familiar with the actual layout of a city than they are with points of interest.

Households and businesses in many developing countries do not have Internet access; however, computer kiosks in transit stations, shopping malls, and other public locations can be installed to increase access to online services. In Europe, almost universal access to cell phones makes possible a similar service via automated telephone access, or the Internet, using the Wireless Application Protocol (WAP).

The benefits of online itinerary planners are tangible, though not easily quantifiable. Public response to the Bay Area MTC trip planner has been "tremendously positive." Although growing ridership cannot be directly attributed to the itinerary planner, more than 1 million itineraries were generated in the first year of operation, and currently 7000 itineraries are produced each day (*Urban Transportation Monitor*, 2003). The Washington Metropolitan Area Transit Authority's planner is currently completing 2 million itineraries each year (Multisystems and Coogan, 2003). Even agencies in smaller transit markets are reporting heavy usage of their itinerary planners. One such agency is Metro Transit in Minneapolis, which provides an average of 80 000 itineraries each month (*Urban Transportation Monitor*, 2002).

Real-time service information

Real-time service information is another form of information dissemination made possible by technological innovations, available to users over a wide variety of media. Real-time service information is now available on most urban rail systems. Displays at stations announce the expected arrival of the next train. Such information continues to be rare for bus transport, because providing it is a far more complex problem. Automatic vehicle location (AVL) equipment and transponders are installed on vehicles. Vehicles are polled periodically, and GPS is used to locate the vehicle along the route. Updated location information is then fed to users via a Web site or some form of automated messaging service.

Examples of real-time service Web sites are Seattle's MyBus.org and Busview.org (Maclean, 2002), both are joint projects of King County Metro and the University of Washington. Busview is a Java applet, available for download to any computer; it displays real-time bus location on a map of King County. Users can choose to display all system buses or only those buses from a particular route or routes. Vehicle progress is updated every few minutes, allowing users to track vehicle speed and approximate arrival time at their stop. MyBus takes an alternative approach to vehicle location; instead of following the bus, it allows the user to choose a bus stop for which the route number, destination, and expected arrival time for imminent buses is then displayed. The system is also accessible to WAP-enabled cell phones and PDAs. Between October 2000 and October 2001, the MyBus Web site received over 7.5 million requests for departure time information from standard computers and 18 000 requests from cell phones. Also of interest is the fact that a greater percentage of wireless users (compared with computer users) access data during the morning peak, suggesting that the time constraints of starting up a computer may restrict usage of Web sites providing real-time information, while the convenience of wireless access provides real benefits to users.

AVL systems also have other benefits that can be highly attractive to potential transit riders. They make possible email or cell phone notification of delays or other changes in specific routes when a user registers to receive updates on that route. Safety can also be enhanced through AVL implementation; many large transport agencies connect silent alarms to the AVL system, making it possible to precisely locate the bus. Casey (2003) reports that King County Metro's use of an AVL system has led to a shorter emergency response time.

Real-time vehicle arrival information can also be provided to travelers *en route* by installing dynamic messaging signs (DMSs) at stops or stations, which display the length of the waiting period until the next vehicle arrives. One of the earliest of such installations was the Countdown system in London in the UK, launched in 1994. A survey taken on the pilot route revealed that, with real-time arrival information displayed, 89% of passengers felt waiting for the bus to be more acceptable, 65% of passengers perceived a shorter waiting time, the bus service was felt to be more reliable, and waiting at night was felt to be safer. The same research found that the Countdown system generated at least 1.5% of new revenue (Multisystems, 2003). Results have been similar in other cities. In both Brussels and Liverpool, routes with real-time arrival information had increased ridership of between 5 and 6%. When Delaware's DART First State tested a real-time arrival system for four months in 2001, ridership on its Beach Bus service increased 13.5% (Starcic, 2003b). A Wisconsin study conducted by Peng et al. (2002) showed that knowing when the bus will actually arrive at the bus stop was ranked highest of ten factors affecting the decision to ride the bus in two separate study areas, proving the marketability and desirability of AVL systems. A study of public response to Portland TriMet's "Transit Tracker" real-time bus arrival notification system, showed that 60% of respondents rated the scheme a "5" on a value scale of 1 to 5, while the mean rating was 4.5 (TriMet Marketing Information Department, 2002).

3. Pricing strategies

Marketing by price is a simple concept: the lower the cost for public transit, the more willing people will be to use it.

3.1. Free fare strategies

Taken to the logical conclusion, doing away with fares entirely will bring the maximum possible ridership growth. Some transit agencies have, in fact, abolished all fares. For some small systems in the USA, the cost of collecting and processing fares represents a large portion of fare revenue. In other cases, local governments see external benefits that may justify free transit. For example, Chapel Hill Transit in North Carolina began operating on a fare-free basis in 2002 (Multisystems et al., 2003); ridership in April 2002 was 54% higher than the same month in the previous year.

Few transit agencies are willing to use such extreme measures to recruit riders. However, special “fare-free zones” or “fare-free times” may be an effective strategy to attract riders. Portland and Denver in the USA have established areas of downtown where transit is offered at no charge. The intent is to promote transit use within the most congested part of downtown and hence reduce car traffic and congestion, as well as to provide an opportunity for people to “test” the system. In Long Beach, California, the transit agency offers a special free shuttle service between major tourist and entertainment attractions (with special vintage-type small buses). People who would not ordinarily experience transit have the opportunity to do so as part of weekend recreational activities. Many transit providers in Europe offer free rides to family members of those holding a monthly pass on weekends or during off-peak periods (Transportation Research Board, 2001).

3.2. *Simplified fare structures*

As an alternative, fare simplification has been used with good outcomes. Fare simplification involves reducing the variability of fares. This is a two-edged sword, since the cost of providing transit varies greatly as a function of distance and time of day. Pricing transit fares based on costs would therefore lead to great variation in fares. Transit patrons favor flat fares, without time of day differentials or special charges for long trips. Thus, pricing strategies must find a way to differentiate fares but not overly complicate fare payments for customers. Various forms of day or multi-day passes provide a good alternative.

A US example is the Maryland Mass Transit Administration’s fare overhaul in 1996 (Taylor and Carter, 1998), which resulted in the reversal of a 2 year trend of falling ridership and revenue. As a result of market research, the agency changed the zonal structure from five zones to one while eliminating transfers and implementing a day pass. Under the previous system, fares ranged from US $1.25 to $2.25 depending on the zone. After simplification, all riders paid a flat fare of $1.35. The price for a day pass was $3.00, just $0.10 more than a round trip fare with two $0.10 transfers. Six months after implementation of the new fare policy, system ridership (including bus, light rail, and metro) had increased 6.3% with the metro ridership rising 16.1%. Moreover, growth in ridership has been sustained (Multisystems et al., 2003), increasing another 11% between 1997 and 2000. Both drivers and riders report higher satisfaction with the simplified fare structure.

The Orange County Transportation Authority (OCTA) is another US agency that has reaped the benefits of simplified fares. Due to the complexity of paid transfers, instances of transfer fraud, and vehicle operators' desire to avoid confrontations with passengers over transfer validity, OCTA eliminated transfers in July, 1999 and began selling day passes as a substitute (Multisystems et al., 2003). As a result, both revenue and ridership climbed consistently, despite a day pass price increase just 6 months after their introduction. Sales of 30 day passes have been particularly impacted by the change.

Connecticut Transit (CTTransit), again in the USA, embarked on fare simplification in 1998 (Lee, 1999). As in Maryland, zones were eliminated, and while free transfers were preserved, convenience fares including a day pass and a 7 day pass were added to the fare structure. Within 6 months of policy implementation, 520 000 day passes and 83 000 seven-day passes had been purchased. Day pass sales also resulted in a 7.5% decrease in transfers issued, amounting to 208 000 trips. Although reliable ridership counts are not available, indicative of success is the fact that during the first 6 months after simplification, passenger revenue was over budget by nearly $300,000, or 3%. Thus, in Connecticut, as in both Orange County and Maryland, fare simplification did not result in lost revenue or ridership despite a rise in prices for at least some customers, but instead resulted in gains for all three agencies.

European transit operators have long offered seasonal passes for regular customers. Agencies from the USA and Canada are beginning to follow their lead and offer more multi-month pass options, as well: the Champaign-Urbana Metropolitan Transit District offers one pass each for Fall, Spring, and Summer; King County Metro offers a 3 month version of the Puget Pass that can be used for any three consecutive months; and London (Ontario) Transit offers a 2 month summer-long pass to secondary students. Studies show that seasonal pass holders make more transit trips than other regular users of public transport (Axhausen et al., 2001).

3.3. Smart and accessible fare payment

Changes in fare payment systems, made possible by new technologies, can have ridership effects as dramatic as changes in fare structure itself. Smart cards, credit card-sized cards containing a microchip, have proven to be a revolutionary transit payment method. Smart cards make possible a wider variety of pricing, integration across services, shorter dwell times at stops, and more convenient access to purchase. Smart cards have more capabilities than the more common stored fare card (e.g. the magnetic strip card); value can be added as well as subtracted, and fares can be tailored to individual travelers. Smart cards also give the transport agency a rich source of data for analyzing patterns of service demand. Smart cards can be sold via the Web or telephone, as well as at commercial outlets; and they can be sold in any amount. Smart cards can be structured to be used as credit or debit cards, say for parking or taxis or shopping purchases.

Because information is exchanged with the card reader through a short-range wireless interface, smart cards are "contactless," and can even be read without being removed from a bag or wallet. Transactions are completed in about one-fifth of a second, making vehicle and/or station entry considerably more efficient.

Hong Kong has been using a smart card system known as Octopus since 1997. The Octopus fare card is valid on all public transportation in the territory, encompassing several different train, bus, and ferry networks. About 10 million Octopus cards are currently in circulation, and nearly 8 million transactions are conducted each day. In addition to transit fare payment, the Octopus card is now accepted by public telephones, taxicabs, convenience stores, and even some grocery stores and restaurants.

In New York City, the New York Metropolitan Transit Authority used magnetic card technology to enable sweeping changes in their fare structure. Prior to 1997, tokens were the only fare media used in the subway; no intermodal transfer mechanism was available and no discount for bulk purchase was offered. With the launch of the MetroCard in 1997, free bus transfers were introduced, and several unlimited use pass functions were gradually added. The effect of these changes on the bus system was marked: by 2000, bus ridership had increased by approximately 35%, rising from 1.6 million daily trips to a peak of 2.4 million, and 80% of all New York subway riders were using the MetroCard (Duffy, 2000a,b).

Washington Metro Area Transit Authority became the first US city to introduce a smart card as a fare medium in 1998. Despite the fact that it can only be used on a single-agency network, it quickly became widely used. By 2000, 1000 cards were being sold each day. In 2000, the Chicago Transit Authority instituted a smart card program, which is inter-agency as well as intermodal (Constantinou, 2000).

Some banks are beginning to issue their own stored-value smart cards, for use as payment at participating retailers. The Bank of Nova Scotia implemented such a system in Barrie, Ontario (Crawford, 2001), and the local transit agency, Barrie Transit, signed on as a participating vendor. Card readers were installed on the entire fleet of buses, allowing riders to pay for single rides with their stored value card. During the first phase of implementation in 1998, the system achieved a 15% reduction in cash handling, and the average transaction time for bus boarders dropped to 1 second.

Smart cards can achieve greater operational and informational efficiency for transit operators, and they attract customers because of their convenience and flexibility. As Giglio (2003) notes, they also have the added effect of divorcing the act of buying from the act of paying, which often results in higher sales to individual customers. This is the principle at work in credit card sales at traditional retailers: customers are not bound by the amount of cash in their purse, making impulse purchases possible. In the same way, smart card-using transit riders are not restricted in their transit use by the tokens or change in their wallet, and are therefore free to ride more often.

4. Market segmentation

A basic tenet of marketing is segmentation: providing a specific product at a specific price for a target group. Public transport providers have engaged in market segmentation quite extensively: senior riders or school children often pay a reduced fare; frequent travelers are offered discount passes; and on some systems peak period travelers are charged a premium. More recently, innovative transit providers are targeting new groups: tourists, university students, and groups of employers. As noted earlier in this chapter, transit providers can manipulate service or price. Some examples of both will be presented in this section.

A strategy that has proven highly successful in attracting riders while preserving an agency's fare revenue is that of class passes. A class pass is a program in which a whole group of people is provided with an unlimited-use transit pass at a highly discounted cost. This is viable for the transit agency because when an entire group of people is given a pass, only a portion of the group will use it. The cost, however, is spread between all members of the group, and thus not burdensome.

The most common application for class passes is at universities. Universities provide a very good market for transit: students represent a large, low-income market, and campus parking charges are typically high. A university assesses every student to determine a fee (the average in 2000 was US $30 per student per year), which provides the students with unlimited transit usage on participating providers (Levin, 2000). The University of California at Los Angeles (UCLA) partnered with the Santa Monica Big Blue Bus to begin a class pass program called BruinGO in 2000 (Brown et al., 2003). Under UCLA's agreement with Big Blue Bus, faculty and staff are entitled to unlimited transit access along with the students. Each time a BruinGO participant rides the bus, he or she swipes his or her university ID card through an electronic reader, and the university pays the transit system US $0.45. During the 8 month pilot program, UCLA's total fare payment to the Big Blue Bus was US $640 000. When averaged among UCLA's 36 900 students and 26 800 faculty and staff, the cost comes to US $1.27 per person per month.

At this low cost, both the university and the transit system reaped many benefits. Among students, the number using transit facilities increased 43%; and among employees the number commuting by bus rose 128%. Across the university as a whole, the number of those riding the bus increased 56% while the number of those driving alone to campus fell 20%. Demand for campus parking decreased by

Table 1
Comparison of class pass programs at various US universities

Transit agency	University served	Year begun	Pass price per term	Yearly ridership
Chicago Transit Authority (a)	29 institutions	1998	Varies by school	15.3 million to 1.7 million
CityB (b)	Purdue University	1999	Contract with school for US $1.1 million (no individual counts)	60% of system riders are students or faculty
Milwaukee County Transit System (c)	University of Wisconsin – Milwaukee	1994	US $31.00	35% increase
Regional Transportation District (b)	University of Colorado at Boulder	1991	US $27.85	1 427 476
Santa Monica Municipal Bus (d)	UCLA	2000	US $0.45/ride	56% increase

Notes: (a) Chicago Transit Authority (2002); (b) as of 2003; (c) Meyer and Beimborn (1996); (d) as of 2001.

1300 spaces. The Big Blue Bus received increased revenue from the university, and also found an added bonus in the electronic payment system, which saves 1 second per rider of boarding time, increasing network efficiency. Among 35 US universities offering class passes in 1998, the number of eligible students was 825 000; the increase in student ridership rose from 71 to 200%; parking demand fell from 1000 spaces to 400; and the reduction in the cost of college attendance for students was an average of US $2000 per year (Brown et al., 2001).

Many US cities and regions have now instituted university-based class passes or are in the process of doing so. Chicago's program is valid for 26 universities in the city. King County Metro's UPASS program at the University of Washington in Seattle, while not a strict class pass, has resulted in only 25% of all faculty, staff, and students driving alone to campus (Santschi, 2002). Table 1 provides a comparison of several university-based class passes and their ridership impacts in the USA.

The use of class passes is expanding beyond the university setting. Several US cities (Denver, Salt Lake City, Seattle) have instituted programs offering class passes to large employers, enabling companies to offer employees a valuable benefit at a low cost. Some cities have gone even further and are offering class passes to all employees in the central business district. White et al. (2002) describe such a program that Ann Arbor, Michigan, began in 1999, calling it "Go! Pass." A joint project of the Ann Arbor Transit Authority, the city, and the chamber of commerce, the Go! Pass provides unlimited use passes to all employees of participating employers at a cost of only US \$5.00 per employee. All employers are eligible, regardless of size. Surveys conducted in June 2000 and June 2001 show a 9.2% increase in bus trips into downtown and a corresponding decrease of 3.5% in the private vehicles driving into downtown.

The Santa Clara Valley Transportation Authority also offers a class pass called EcoPass to all employers in the service area. The program has been expanded to include a Residential EcoPass, which allows "any residential community with 25 units or more that is defined by a geographical boundary, such as an apartment building or condominium complex" to purchase unlimited use passes for all development residents.

4.1. Public–private partnerships

Class passes are one way that public transit agencies are partnering with the private sector to increase transit ridership and revenue. Another is to offer services to specific organizations that are tailored to customer needs. Knoxville Area Transit (KAT) in Knoxville, Tennessee, has been aggressively using this strategy (Schneider and Hairr, 2002). In 1999 KAT began a new route connecting the University of Tennessee to an apartment complex that catered to students but was 4 miles (6.4 km) away from the campus. Partially funded by the apartment complex, ridership on the route has grown 121% since its establishment. The success of this initial line has led to more public–private partnerships providing transit between the university, other housing developments, and the downtown business and entertainment areas.

The Lubbock, Texas, transit agency Citibus partners with United Supermarkets, a major local grocery chain, which underwrites routes providing bus service from retirement homes, community centers, and other institutions to the markets. The service runs twice each weekday, providing grocery shopping opportunities to seniors, disabled, and other individuals living along the routes, which are free to all riders and "guaranteed" by the market chain, although open to the public (Texas Transportation Institute et al., 1999).

CityBus in Lafayette, Indiana, is a recent transit success story, having doubled ridership in the 4 years between 1998 and 2002 (Duffy, 2003). This growth was due primarily to a class pass agreement with Purdue University; however, the agency has also used partnerships with private entities in creative ways. One approach is pre-paid contracts with apartment complexes for the establishment of shuttles between the developments and the Purdue University campus. A second strategy was to work with a private developer on a downtown building less than a block from the transfer center. The developer owns the top floor, which is rented out as student apartments, while CityBus rents the ground floor at a reduced rate to a childcare provider. Having childcare available at a bus transfer point reduces the need for extra trips, enhancing the usefulness of the bus for commuting parents. The childcare center has a waiting list to get in, but bus riders and welfare-to-work participants are given priority.

5. Conclusions

There is a wide array of marketing strategies for increasing the patronage of public transport. This chapter has taken a broad view of marketing, considering not only promotions and information dissemination, but service improvements, pricing strategies and market segmentation strategies to better serve various submarkets. Recent technology improvements have made possible significant improvements, from real-time information to flexible fare payment cards to more reliable travel times. There is considerable evidence that better service and more attention to serving the customer is working: public transport use is increasing in many metropolitan areas in the USA, Europe, and other parts of the developed world.

References

Axhausen, K., A. Simma and T. Golob (2001) "Pre-commitment and usage: cars, season tickets and travel," *European Research in Regional Science*, 11:101–110.

Brog, W. (2003) "Reducing car use!? Just do it!" *27th Nottingham Transport Conference*. Nottingham.

Brown, J., D.B. Hess and D. Shoup (2001) "Unlimited access," *Transportation*, 28:233–267.

Brown, J., D.B. Hess and D. Shoup (2003) "Fare-free public transit at universities: an evaluation," *Journal of Planning Education and Research*, 23:69–82.

Casey, C. (2003) "Real-time information: now arriving," *Metro*, 99:56–59.

Chicago Transit Authority (2002) *2001–02 U-Pass student ridership*. Chicago: Market Research Department, Chicago Transit Authority.

Constantinou, T. (2000) "Chicago Smart," *Mass Transit*, 26:52–53.

Crawford, D. (2001) "Barrie reduces the barriers," *ITS International*, March/Apr.:39–40.

Duffy, J. (2000a) "The Big Apple's new MTA NYCT," *Mass Transit*, 26:10–34.

Duffy, J. (2000b) "Double digit ridership increases on NYCT buses," *Mass Transit*, 26:38–48.

Duffy, J. (2003) "Lafayette Indiana: CityBus has something to teach about ridership," *Mass Transit*, 28:13–27.

Giglio, J.M. (2003) "Creating customers," *Public Works Management and Policy*, 7:231–234.

Lee, D.A. (1999) "Introducing fare simplification and new convenience fares at Connecticut Transit," *Transportation Research Record*, 1669:109–112.

Levin, J. (2000) "Distributive cost pricing: an effective strategy toward building transit ridership quickly among targeted markets," in: *American Public Transit Association Bus and Paratransit Conference Proceedings*. Houston.

Maclean, S.D. (2002) "Wireless internet access to real-time transit information," *Transportation Research Record*, 1791:92–98.

Meyer, J. and E. Beimborn (1996) *Evaluation of an innovative transit pass program: the UPASS*. Madison: Wisconsin Department of Transportation.

Multisystems (2003) *Strategies for improved traveler information. Transit Cooperative Research Program report 92*. Washington, DC: Transportation Research Board, National Academy Press.

Multisystems and M.A. Coogan (2003) *e-Transit: electronic business strategies for public transportation. Transit Cooperative Research Program report 84*, Vol. IV. Washington, DC: Transportation Research Board, National Academy Press.

Multisystems, Mundle and Simon and Simon Research (2003) *Fare policies, structures and technologies: update. Transit Cooperative Research Program report 94*. Washington, DC: Transportation Research Board, National Academy Press.

Peng, Z.R., D.L. Yu and E. Beimborn (2002) "Transit user perceptions of the benefits of automatic vehicle location," *Transportation Research Record*, 1791:127–133.

Robinson Research (1999) *Washington State Department of Transportation Advertising Awareness Study: executive summary report*. Olympia: Washington State Department of Transportation (http://www.wsdot.wa.gov/relax/research.pdf).

Santschi, M. (2002) "University of Washington: ten years of results," *Parking Today*, 7:36–38.

Schneider, R. and M. Hairr (2002) "Discrete market segments and public–private partnerships: a model of transportation development supporting urban revitalization," in: *American Public Transit Association Bus and Paratransit Conference Proceedings*. Minneapolis.

Starcic, J. (2003a) "How to fill those empty seats," *Metro Magazine*, 99:26–28, 30.

Starcic, J. (2003b) "Designing bus shelters for the savvy passenger," *Metro Magazine*, 99:50–53.

Taylor, S. and D. Carter (1998) "Maryland Mass Transit Administration fare simplification: effects on ridership and revenue," *Transportation Research Record*, 1618:125–130.

Texas Transportation Institute, Southwest Transit Association, and University of Wisconsin-Milwaukee (1999) *A handbook of proven marketing strategies for public transit. Transit Cooperative Research Program report 50*. Washington, DC: Transportation Research Board, National Academy Press.

Transportation Research Board (2001) *Special report 257 – making transit work: insight from western Europe, Canada, and the United States*. Washington, DC: Transportation Research Board, National Academy Press.

TriMet Marketing Information Department (2002) *Transit tracker evaluation June 2002: final report*. Portland: TriMet.

Urban Transportation Monitor (2002) "Web-based transit trip planners very successful," *Urban Transportation Monitor*, 16:1–2.

Urban Transportation Monitor (2003) "This week's survey results: Web-based transit trip planners," *Urban Transportation Monitor*, 17:7–11.

White, C., J. Levine and M. Zellner (2002) "Impacts of an employer-based transit pass: the Go! Pass in Ann Arbor, Michigan," in: *American Public Transit Association Bus and Paratransit Conference Proceedings*. Minneapolis.

This Page Intentionally Left Blank

Chapter 38

IMPLEMENTING INTELLIGENT TRANSPORTATION SYSTEMS

BRIEN BENSON
George Mason University

1. Introduction

Intelligent transportation systems, or ITS, are the application to surface transportation of information technologies – sensors, computers, and telecommunications (Taylor, 2001). Examples include new and sophisticated traffic signals, electronic toll collection, use of Global Positioning System technology to keep track of fleets of trucks and buses, and use of acoustic and infrared sensors and video cameras to monitor traffic conditions.

Information technologies are of course used extensively in aviation and the maritime industries, for example in air traffic control, passenger and baggage screening, scheduling and reservations, fare calculations, and tracking of freight movement, but in neither of these industries is there a national policy to promote the use of information technologies.

ITS technologies of one kind or another are used throughout most of the world, and in Japan and Europe governments have adopted formal ITS policies akin to those in the USA. Japan, with its strong central government, densely concentrated population requiring carefully planned transportation, and impressive consumer electronics industry, has paid particular attention to national ITS policy, and has well-developed ITS traffic control and vehicle navigation systems.

Europe, like Japan, has generally denser populations than the USA, and also like Japan, more of a tradition of strong central government than the USA. For these reasons some European nations have been particularly aggressive in promoting ITS – Germany, for example, in the development of government–industry partnerships to deploy traffic control systems; France in deploying the Minitel computer system permitting household access to transportation information; and the UK in the posting of variable traffic speeds on electronic variable message signs. Yet, as a whole, national differences have frustrated efforts at European-wide ITS planning, and, on balance, the European ITS program is at best on a par with that of the USA, on which this chapter will concentrate.

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

The US ITS program began in earnest with the 1991 Intelligent Vehicle-Highway Systems (IVHS) Act, which led to US $250 million being spent annually on ITS research, testing of pilot projects, and development of national standards. The program was extended by the 1998 Intelligent Transportation Systems Act, which called for the same spending levels but a reorientation from research to deployment of ITS. Today, government at all levels in the USA spends an estimated US $500 million annually on ITS, and private industry spends a similar amount. During the 13 years since a formal ITS program was established in the USA a number of important policy issues have arisen, to which we now turn.

2. Federal leadership

The overarching public issue confronting ITS in the USA is the appropriate role of the federal government; in particular, the degree to which the federal government should guide the evolution of the ITS industry. US national legislation provides that the federal government should direct preparation of a 10 year program plan, and should also develop a national systems architecture with supporting standards and protocols.

These requirements flow from a conviction on the part of many ITS leaders in the USA that without federal leadership, uncoordinated stovepipe applications of IT to transportation will proliferate, and that federal guidance of ITS will be successful, for two reasons. First is the extraordinary success of the US interstate highway system, which showed that a national vision could be carried out, that rigorous national standards of design and construction were feasible, and that federal, state, and local transportation agencies could work closely together to achieve a national goal.

The second source of confidence in a nationwide application of IT to transportation in the USA was the development of the systems integration discipline during the Apollo moon shot program, which demonstrated dramatically that huge, complex systems could be successfully integrated. Finally, the ability of the computer and consumer electronics industries to achieve national product standards clearly beneficial to the market, for example facilitating an abundance of plug-and-play audio and video systems, gave ITS visionaries confidence that national IT standards could function throughout the USA, not just where the federal government was the dominant purchaser but also where market forces prevailed.

2.1. National program plan

The federal 1991 IVHS Act, which was the foundation of the national ITS program in the USA, required preparation of a 20 year program, and the requirement for a program plan was renewed in the 1998 ITS Act, while being trimmed back to 10 years:

> The National ITS Program Plan shall-
>
> (A) specify the goals, objectives, and milestones for the research and deployment of intelligent transportation systems in the context of major metropolitan areas, smaller metropolitan and rural areas, and commercial vehicle operations;
>
> (B) specify how specific programs and projects will achieve the goals, objectives, and milestones referred to in subparagraph (A), including consideration of the 5- and 10-year timeframes for the goals and objectives.[a]

The idea behind this requirement is that a national vision for the USA should be developed by all interested parties, and that this vision should be more than just vague speculation, but should include specific milestones. There is ample precedent in US transportation law for long-term plans, a result of the fact that major road and rail construction projects frequently take years to complete and are likely to be used for many more years, and even decades.

However, in a sector dominated by the information technology industry, any effort to look 10 years into the future raises serious questions of credibility. To begin with, many of the technologies involved are rapidly evolving and will likely permit the development of many new products, particularly involving miniaturization, which there is very little ability of accurately foreseeing. The often-cited Moore's law states the computing speeds roughly double every 18 months, and while this rate may not necessarily go on for as long as 10 years, certainly rapid progress can be expected.[b]

Similarly, revolutionary developments in telecommunications could well lead to new products and services that a 10 year plan could hardly expect to foresee. Recent examples of such developments include the substitution of traditional transmission technology by fiber optics, the exponential growth in cell phone utilization, and the recent emergence of VOIC, permitting the Internet to replace traditional telephone networks.

Equally rapid evolution of electronic sensing devices and systems can be expected, with, for example, the capabilities of monitoring traffic conditions in radically new ways, the ability to revolutionize the tracking and monitoring of commercial vehicles, and potentially dramatic new applications of in-vehicle monitors, for example of driver drowsiness and emotional condition.

But rapidly evolving technologies are not the only challenge to producing an even remotely realistic 10 year US program plan. Equally challenging is

[a]Transportation Equity Act for the 21st Century. Title V: Transportation Research. Section 5205: National ITS Program Plan.

[b]For a full discussion of unpredictable factors facing emerging technologies, see Benson and Sage (1995).

foreseeing how various new products and services might be accepted by the community of users, or what we might call the marketplace. A classic example is the videophone, which the telephone industry has tried to promote over the years but has met stiff resistance from users apparently seeing it as a potential invasion of privacy.

More specific to transportation, there are already serious questions being raised as to how many information gadgets in a vehicle can be dealt with by drivers without compromising their ability to keep their attention on driving. And some new technological devices for collecting and transmitting traffic information are finding widespread apathy as travelers continue to draw the traffic reports from "old technology" traffic radio stations.

A third important uncertainty facing efforts to develop a 10 year national program plan for ITS in the USA is the divergent policy priorities of different jurisdictions throughout the country, seen in the following divergent ITS projects adopted in different jurisdictions:

- Los Angeles City's "Smart Corridor," running some 30 km from downtown Los Angeles to the western suburbs, facilitates drivers shifting from the Santa Monica freeway to parallel arterials when the freeway is congested, and is uniquely suited to the area's strong grid network of highways.
- Boston's "Smart Route" system collects traffic data from television cameras placed atop downtown skyscrapers that are then distributed throughout the region. This system reflects the region's strong satellite television orientation.
- Suburban New York City's TRANSCOM links some 20 transportation jurisdictions – state, county, and city – through wireless communication, and is designed to promote cooperation among these jurisdictions in the management of traffic flow into Manhattan. This operation reflects the complexity of New York City's suburban environment.

Of course, the US national program plan was not meant to be a dictate to be rigidly imposed throughout the country, but it is intended to give strong guidance to state and local transportation jurisdictions in the country, and it is intended to help shape the systems architecture and standards program discussed below. The obvious limitations of this plan just discussed, then, are a serious problem with national ITS policy in the USA.

2.2. *Systems architecture and national standards*

The 1998 US ITS Act states:

> The Secretary [of Transportation] shall develop, implement, and maintain a national architecture and supporting standards and protocols to promote the widespread use and

evaluation of intelligent transportation system technology as a component of the surface transportation systems of the USA.

(2) Interoperability and efficiency. – To the maximum extent practicable, the national architecture shall promote interoperability among, and efficiency of, intelligent transportation system technologies implemented throughout the USA.[a]

The US national system architecture is stated in such general terms that it has little bite, and has therefore occasioned minimal controversy. On the other hand, the US national standards program, flowing from this systems architecture, is the subject of intense discussion. ITS standards include the data content and organization of message sets, communications protocols, and hardware and software configurations.

The overall ITS standards setting strategy adopted by the US federal government has been to rely primarily on collegially developed standards, using existing standards development organizations (SDOs), e.g. the American Society for Testing and Materials, while holding in reserve two federal coercive powers – the authority to hold rule-makings to impose standards, and to withhold federal funds from jurisdictions failing to develop a systems architecture incorporating federally established standards.

On paper, the record of standards setting looks rather impressive. As of June 2004, the US Department of Transportation Web site reports that 73 ITS standards have been published and another nine are approved and awaiting publication.[b]

But these numbers are misleading. Just because a standard has been adopted by a national standards organization does not mean it has been adopted in the marketplace, and, while no detailed and comprehensive review of marketplace adoption of standards has been completed, anecdotal evidence suggests that few of these formally promulgated standards are actually in widespread use throughout the USA.

This failure of adoption reflects fundamental problems with the standards program in the USA. First, many standards were promulgated without adequate technical testing, or if there had been technical testing, they were not tested in a working environment. As a key study by the Transportation Research Board (2004) notes:

formal testing, observation of in-service applications, or other verification of functioning may or may not be considered a routine part of a particular standard's development process.

[a]Transportation Equity Act for the 21st Century. Title V: Transportation Research. Section 5206: National Architecture and Standards.

[b]http://www.standards.its.dot.gov/status4.htm.

The report therefore recommends not only

> testing to ensure that the proposed standards can be used in field applications and will perform as expected

but also an

> assessment of readiness for deployment, which would likely consider the number of applications that might realistically be expected in the near term (i.e., 3 to 5 years), the number of manufacturers and system integrators capable of delivering ITS installations meeting the standard, and the availability of information and materials to facilitate the standards' application.

A second problem is that the level of detail in ITS standards has been inconsistent, leaving numerous opportunities for confusion. For example, it is not unusual for manufacturers whose basic device conforms to the National Transportation Communications for ITS Protocol (NTCIP), to add on a feature offering competitive advantage whose adherence to NTCIP is at best unclear. To take another example cited by Transportation Research Board, two jurisdictions in Washington, DC, planned to purchase systems that were presumed to comply with the national data dictionary, but in fact were incompatible with each other.

These problems of inconsistent inadequately tested and inadequately vetted standards do not result from project mismanagement. Indeed, able and dedicated professionals are doing their best to carry out a congressional mandate. Instead, these problems raise questions about the ability of a single, top-down organization such as the federal government to set a realistic national agenda for the development and promulgation of technical standards in a rapidly evolving industry. As argued by Cargill (1989):

> In a market that is growing explosively and is dynamic to the point of instability, internal relationships are unclear, and there are no guarantees that tomorrow's market will be based on today's activities. In this atmosphere, planning is impossible; it is a speculator's market. No standards can be written for this type of environment: standards are postulated on a recognized growth path and certain continuity of progress; in a high-growth, highly dynamic market, there is nothing but discontinuity. If standards, voluntary or otherwise, are introduced into this environment, they will fail, since standards act to stabilize a market.

Even when standards can realistically be set in a dynamic environment, it is as likely that they will emerge from the struggle for market dominance as from a national professional collegial standards-setting organization. To cite just one obvious example, many critically important software standards for the personal computer have been set through the market dominance of Microsoft.

Indeed, the collegial approach typically employed by the SDOs may by its nature be unsuited to a technologically dynamic environment. As the Transportation Research Board (2004) study says:

SDO procedures share at least one common characteristic: they are deliberate and therefore relatively slow, particularly compared with the rate of change in telecommunications, computing, and other technologies underlying ITS. If the examples of other telecommunications and information processing products are relevant, ITS infrastructure standards will require periodic revision to avoid obsolescence and may face the problem even before their initial publication.

Another fundamental problem is that many ITS technologies are better suited to standardization at the regional than the national level. As the Transportation Research Board (2004) notes:

> many ITS functions do not require interactions among system components above a local or regional level. There is no compelling reason, the committee observed, why a traffic information and management center in Boston, for example, should be able to communicate directly with its counterpart in Phoenix. Why, then, should agencies in either region be expected to compromise on the functional requirements they seek to purchase or the cost they must pay for their systems so that their hardware and software will meet national standards?

For all these reasons – the dynamic nature of the technologies, the success of market-driven standards in some cases, and the greater relevance of ITS standards to the regional than national level – it is appropriate to examine what should be the federal role in standards setting for ITS in the USA. And, indeed, the Transportation Research Board (2004) study suggests that the federal role should be carefully circumscribed:

> The bases for allocating ITS Standards Program resources could be explained more clearly and the allocations could be more tightly targeted than they are now. ... Because resource limitations certainly will influence decisions to provide federal support for particular ITS standards and because the range of development and deployment activities that could warrant such support is large, the committee anticipates that pursuing their recommendation may limit the number of standards being developed with federal support.

The report then recommends that the US federal government should limit its standards setting based on consideration of the standards contribution to implementation of priority services in the US national systems architecture, whether the standard is cost-effective, and whether federal support is essential to the development of the standard.

3. Mainstreaming ITS

It is now 14 years since the 1991 federal IVHS Act began the national ITS program in the USA, and an important policy question for transportation planners is the degree to which ITS technologies can now stand on their own and compete directly with alternative transportation practices, and the degree to which ITS still merits protective nurturing.

The former approach is called "mainstreaming," and commits ITS to the direct support of the full range of traffic operations – management of traffic signals, commute traffic, incident response, highway work zones, special events traffic, and regulation of commercial vehicle movement. ITS can play a major role in all of these services, with advanced traffic control devices such as advanced programming of traffic signals (including "adaptive" traffic signals, whose cycle breaks adjust continually to accommodate changing traffic flow patterns), ramp metering, electronic toll collection, and advanced traffic information systems.

The argument for mainstreaming is that improved traffic operation is vitally important – for example, at least half of all metropolitan area congestion in the USA is due to operational rather than infrastructure problems[a] – and by supporting traffic operations ITS can find its best utilization.

The alternative perspective is that ITS still includes a large number of technologies that are not yet fully proven and accepted throughout the transportation profession but nonetheless hold great promise. This is particularly the case with highly advanced concepts, such as "hands-off-the-wheel" automated highways in which vehicles are guided by roadside control systems. From this perspective, ITS is still an infant industry meriting special treatment.

4. Legal issues

In the early days of ITS there was considerable concern that a complex of legal problems could be show stoppers, or at least present formidable barriers to the development of ITS. In the event, a sensible and conservative approach to deployment of ITS in the USA all but eliminated liability and privacy problems, and procurement and intellectual property rights, although matters of concern, have proven tractable.

4.1. Liability

In the early days of ITS, there was widespread concern in the USA that use of new information technologies might spawn a host of devastating lawsuits; for example, should alternative route recommendations lead motorists into unfamiliar territory

[a]According to the US Federal Highway Administration, on average 25% of metropolitan area congestion is due to crashes and breakdowns, 10% to construction, 5% to poor signalization, 5% to special events, and 15% to weather (http://www.fhwa.dot.gov/congestion/congest2.htm).

resulting in accidents, or should hands-off-the-wheel automated highways result in massive chain collisions. In the event, ITS backed away from particularly risky procedures such as recommending specific alternative routes, and also backed away from futuristic research that might have produced risky technologies. In summary, according to a US Federal Highway Administration (2000) study, "liability has not been a major issue to date."

4.2. Privacy

In the early days of ITS there was widespread concern in the USA that surveillance technologies and motorist databases envisioned for some applications would generate a privacy backlash. In the event it has been found that usually there are simple solutions to protect individual privacy. For example, data on a driver's speed used to calculate traffic flow is destroyed as soon as its immediate usefulness is over, typically after 15 minutes. One ITS surveillance technology, use of cameras to enforce red-light running laws, is controversial in the USA, but the safety benefits are seen in many jurisdictions as outweighing the threat to privacy.

Again citing the aforementioned US Federal Highway Administration (2000) study:

> To date, ITS professionals have addressed most of the privacy concerns surrounding ITS, including concerns that traffic surveillance cameras would be sued for purpose other than monitoring traffic and that information on trucking firms collected by preclearance systems would be released to competitors.

4.3. Procurement

Two mainstays of US government procurement – presenting detailed specification requirements, and awarding to the low-cost bidder – have proved problematic for ITS. In rapidly evolving technologies such as ITS, it is often in the best interests of the procuring government agency to have a flexible agreement with the contractor, permitting specifications to evolve. It is also often preferable to choose the most qualified rather than the cheapest vendor. For example, former Virginia Department of Transport ITS chief James Robinson called the two freeway management centers in Virginia an initial "disaster" because of low bid procurement proceeds that followed US Federal Highway Administration guidelines (*ITS International*, 1999).

However, US state departments of transportation have found ways around this problem. Virginia split off its civil engineering from its software procurement, permitting flexible contracting for software. The Arizona Department of Transport, constricted by traditional regulations, gave the legal authority to manage a key ITS procurement to the county of Maricopa (location of Phoenix), which has more flexible statutes. And Minnesota passed legislation permitting its department of transport to enter into flexible partnerships with the private sector.

4.4. Intellectual property rights

One particular contracting issue at the state department of transport level has been intellectual property rights, as many state laws are unduly protective of the state's interests, thereby driving away capable vendors possessing useful patents. But solutions to this problem have been found. In an Arizona agreement, the vendor was permitted to keep rights to the software it brought to the project, while the state received a 5 year license to use this software, and the intellectual property rights in any improvements to the software were given to the state. In Texas the problem was solved by tightening up overly broad and ambiguous legislative protection of the state's intellectual property rights (US Department of Transportation, 2000).

5. Road pricing

The enormous success of deploying electronic toll collection (using on-vehicle transponders and roadside readers) in the USA has encouraged transportation planners and analysts to consider additional uses of electronic toll collection (ETC). The two most frequently mentioned possibilities are in congestion pricing, and in the funding of new road construction.

Congestion pricing (sometimes called "value pricing") readjusts the price that is charged for use of a toll road to take account of traffic levels. As levels rise, the toll rises, so as to discourage use and thereby maintain a smooth traffic flow. Such adjustment may be automatic at certain times of the day, to accommodate standard commute patterns, or it may be continually adjusted in "real time" to reflect changing traffic patterns. Congestion pricing using ETC has been successfully applied to four freeways in the USA – SR 91 in Orange County, California; I-15 in San Diego, California; Katy Freeway in Houston, Texas; and Lee County, Florida – and in Toronto, Singapore, Paris, and three locations in Norway.

Congestion pricing has generally not proven controversial when applied to a newly constructed road, but, when applied to an existing road, it is often considered to be a new layer of taxation, and resented as such. In an effort to soften such resistance, pricing advocates have proposed various schemes for distributing funds collected through pricing – reducing transportation taxes such as fuel taxes, improving existing transportation facilities, and distributing funds to low-income travelers – but no such plans have appreciably reduced political opposition to pricing on already constructed roads.

A second possible application of electronic toll collection is the financing of new road construction, where added capacity is needed but cannot be financed by government. The privately owned 20 mile (32 km) Greenway toll road in Northern Virginia outside Washington, DC, is one such example of ETC-financed construction. Two other such projects in Virginia are under active consideration: additional lanes on the Washington, DC, Capital Beltway and along the rural I-81. The Washington, DC, Beltway project envisions using road pricing, so as to combine the benefits of congestion pricing with toll financing.

6. Cell phone usage

Cell phone usage by drivers is the most controversial safety issue facing ITS in the USA, partly as a result of a few highly publicized fatal traffic accidents apparently due to drivers preoccupied with their cell phones, and partly because of the enormous number of cell phone calls made by US drivers, roughly 1 billion minutes a day, or 40% of all cell phone calls, according to one estimate. Virginia Polytechnic, University of Virginia, Center for Transportation Research director Tom Dingus reports that a study at his institution found that "the biggest distracting agent was cell phones – much bigger than eating or grooming or tuning the radio" (*Wall Street Journal*, 2004).

Two states, New York and New Jersey, plus the District of Columbia have now banned the use of cell phones in cars without a headset or other hands-free device, in an effort to address this problem, yet some argue that such laws miss the critical point – that even hands-free cellphone usage distracts a driver's attention. As National Highway Traffic Safety Administration administrator Jeffrey Runge said, "We have states and local municipalities making rules that basically give hands-free phones a free pass as being safe. … That's not good policy" (*Wall Street Journal*, 2004).

7. Education and training

As with any new industry, ITS has faced a severe shortage of personnel trained in the emerging technologies, a particular problem in government agencies, with typically modest salary levels and rigid personnel procedures. Universities, professional societies, and the US Department of Transportation have responded with a wide range of ITS courses, including on-line instruction offered by the Consortium for Intelligent Transportation Education (CITE), housed at the University of Maryland. In a survey of ITS offerings at US colleges and universities, it was found that half of the 26 responding engineering programs offered at least one full semester course in ITS, and the other half offered an ITS short course or included ITS in a transportation course (Benson, 1996).

Such efforts are not controversial, but there is a dialogue on the question of educating the so-called "new transportation professional." MIT professor and long-time ITS advocate Joseph Sussman argues (Sussman, 1995),

> We will need to be especially assertive in developing an interdisciplinary environment in which we can educate the New Transportation Professional. ... We need to re-emphasize the fundamentally interdisciplinary nature of our field and the roles of new technologies and concepts within it.

On the other hand, ITS professional Donald Orne (unpublished work) counters that this may be asking too much of our educational institutions:

> Multi-disciplines may never be encompassed in single individuals. It may take too long or change may be too rapid to ever expect to see the paragon we keep trying to describe. The systems process could well end up describing two or three, or a dozen, types that we can seek simultaneously. Or, there may be a time series during which different evolving mixes of individuals are produced.

The question, then, is not about the ideal of the broadly educated ITS transportation professional, but about its practicability.

8. Conclusion

This chapter has discussed the major long-term policy issues in the US national ITS program. Of course, there are many other controversies which have flared up at one time or another, for example, whether video cameras should be used to enforce high-occupancy vehicle lane restrictions, whether traffic signal pre-emption for behind-schedule transit buses is desirable, and whether freeway ramp metering is worthwhile or counter-productive. Important as these questions are, they generally concern deployment at the local level, and have been resolved at the local level, so this chapter has not focused on them.

Also, the questions raised in the early days of ITS by environmentalists concerned that improved traffic flow promised by ITS would lead to more vehicle-kilometers traveled and therefore more pollution have been skipped over. This concern has dissipated, as evidence has accumulated that ITS does not increase the distance driven, and that the smoother traffic flow ITS can in fact produce results in diminished emissions.

In conclusion, then, the principal policy issue in the US ITS program is the appropriate role of the federal government; the other key issue is how rapidly ITS should be mainstreamed. Legal hurdles once thought formidable – liability, privacy, intellectual property rights, and procurement – have proven tractable. And there is general consensus on the importance of education and training, with some difference of opinion on the practicality of fully educating the new transportation professional.

References

Benson, B.G. (1996) "ITS education in American universities," *ITS Quarterly*, fall/winter:17.

Benson, B.G. and A.P. Sage (1995) "Case studies of systems management for emerging technology development," in: W.B. Rouse, ed., *Human/technology interaction in complex systems*. Greenwich: JAI Press.

Cargill, C.F. (1989) *Information technology standardization: theory, process, and organizations*. Palo Alto: Digital Press.

ITS International (1999) "Are low bids killing ITS?" *ITS International*, May/June:32.

Sussman, J. (1995) "Educating the new transportation professional," *ITS Quarterly*, summer:8.

Taylor, M.A.P. (2001) "Intelligent transport systems," in K.J. Button and D.A. Hensher, eds, *Handbook of transport systems and traffic control*. Oxford: Pergamon.

Transportation Research Board (2004) *Development and deployment of standards for intelligent transportation systems: review of the federal program*. Washington, DC: Transportation Research Board.

US Department of Transportation (2000) *What's yours, mine, and ours: overcoming intellectual property rights issues: a cross-cutting study*. Washington, DC: US Department of Transportation, Intelligent Transportation Systems.

US Federal Highway Administration (2000) *What have we learned about intelligent transportation systems?* Washington, DC: US Federal Highway Administration.

Wall Street Journal (2004) "As industry pushes headsets in cars, US agency sees danger," *Wall Street Journal*, July 19:A7.

This Page Intentionally Left Blank

Chapter 39

OPTIMIZATION OF TRANSPORT STRATEGIES[a]

ANTHONY D. MAY, SIMON SHEPHERD and GUENTER EMBERGER
University of Leeds

1. Introduction

Optimization implies identifying the best strategy to meet a given policy objective. As noted in Chapter 14 on integrated transport strategies, any strategy is likely to involve a set of policy instruments, and to draw on the interaction between those instruments. The process therefore involves specifying one or more objectives, combining them into an objective function, selecting a wide range of possible policy instruments, and varying their levels, in combination, until the maximum value of the objective function is found. This set of levels of policy instruments will represent the theoretical best strategy. In practice, decision-makers will usually not be able to implement such a strategy, either because elements of it are unacceptable or because of constraints such as availability of finance. Practical optimization therefore typically involves identifying the combination of policy instruments and levels that best meets a given set of policy objectives, subject to satisfying a set of policy constraints.

Traditionally, decision-makers and their consultants attempt to determine the best strategy through a process of identifying one or more possible solutions, appraising them, and then seeking improvements. This process can be inefficient and under-productive; time will be spent appraising inappropriate solutions, and there will be no guarantee that the best strategy has been found. Two benefits of optimization are thus the ability to develop more effective strategies and the speed with which this can be done. In an early example in Edinburgh in the UK, an initial study used some 70 runs of a strategic transport model to develop a "best" strategy for a given set of objectives; a subsequent study using optimization methods found a combination of policy instruments, after 25 model runs, which

[a]The research on which this chapter is based was funded by the UK Engineering and Physical Sciences Research Council and the European Commission, and was conducted in collaboration with several partners. Their support is gratefully acknowledged. The responsibility for the conclusions rests, however, with the authors.

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

was 20% more effective when measured against the same objectives (Fowkes et al., 1998). While decision-makers are unlikely to want to have their decisions automated in this way, optimization provides a very elegant way of suggesting strategies for consideration. Moreover, it can often suggest policy combinations that would not otherwise have been considered.

Formal optimization is a relatively new concept in the analysis of transport strategies, and there are still very few examples of its use. The authors' work, which is presented here, developed from the kernel of an idea that emerged in the preparation of a transport strategy for Edinburgh (May et al., 1992). This led to the initial specification of the method (Fowkes et al., 1998). The method has subsequently been demonstrated and enhanced in applications in several European cities (May et al., 2000, 2001; Emberger et al., 2003), and in a Europe-wide transport strategy (SAMI, 2000). A recent European project has demonstrated the method in six cities, and developed guidance on its application (Minken et al., 2003). That guidance has been drawn on in this chapter. The authors know of few other attempts to apply similar concepts, though the work of Proost et al. (2002) on national and conurbation strategies offers some parallels.

The process involves the following key stages:

(1) Specifying one or more scenarios within which the optimal strategies are developed.
(2) Developing an objective function, as a weighted combination of the objectives of concern to the decision-maker. If necessary, optimization can be conducted with a range of weights to test the robustness of the strategy.
(3) Introducing into the objective function any constraints to be observed; these could be related, for example, to targets for particular objectives, such as carbon dioxide (CO_2) emissions or maximum levels of financial support. It can be helpful to run the optimization unconstrained as a way of determining what these constraints should be.
(4) Specifying an appropriate range of policy instruments to be tested, and representing each as a policy variable. Infrastructure projects are typically included as dummy variables, while management and pricing instruments can be reflected through continuous variables such as fare or frequency.
(5) Determining whether these policy instruments should vary by location, time of day, or over elapsed time, and how these variations should best be represented.
(6) Assessing whether any limits should be placed on the range within which each policy instrument can be tested; these can be determined by practical concerns such as the capability of traffic management measures, acceptability concerns such as the maximum level of a fares increase, or analytical concerns related to the capability of the model.

(7) Using a transport model to test an initial set of combinations of these policy variables, and determine the level of the objective function for each.
(8) Using an optimization procedure to suggest other policy combinations to be tested.
(9) Repeating steps 7 and 8 until an optimum has been identified.

In the next section this process will be described analytically. Two case studies will then be provided, both of which have used a strategic land use–transport interaction model, MARS (Pfaffenbichler, 2003). Finally, some conclusions will be drawn.

2. Optimization

This section outlines the general optimization problem including the definition of an objective function to reflect issues of sustainability. It discusses how to model the policy instruments and simplify the possible search space, and presents a summary of alternative optimization approaches. It is illustrated with the approach adopted in the examples that follow. It should be stressed that this approach only represents one possible objective function; the reader is encouraged to specify an objective function appropriate for the problem to be solved.

2.1. *The general optimization problem*

The general problem relates to maximizing a quantified objective function (either some form of cost–benefit analysis (CBA) or quantified multi-criteria appraisal) within a given exogenous scenario using a given range of land use and transport policy instruments. The appraisal period is taken to be 30 years, though sustainability issues relate to an even longer term. A recent European project, PROSPECTS (Minken et al., 2003), defined an overall objective in terms of sustainability and its seven sub-objectives as follows:

- economic efficiency;
- protection of the environment;
- livable streets and neighborhoods;
- safety;
- equity and social inclusion;
- contribution to economic growth;
- inter-generational equity.

The first sub-objective is dealt with by standard cost–benefit analysis. For the next five sub-objectives, indicators have been defined. There is no one-to-one relationship between indicators and sub-objectives, though, since some of the sub-objectives are of a composite nature. Inter-generational equity is dealt with by weighting the annual benefits in the objective function.

Let M be the set of indicators that we want to apply in a particular study. In principle, we should be able to compute them for any of the years in the appraisal period. For all or a sub-set of the indicators, their annual levels are included linearly in the objective function. For some of the years, there may be targets on some of the indicators. It is possible to include these targets, and constraints on, for example, availability of finance, in the objective function. To appraise all strategies against the same objective function, it is important that the objective function be the same throughout a particular study.

Let $X_t = (X_{1t}, \ldots, X_{nt})$ be the vector of the levels of the n policy instruments in the year t, and $X = (X_1, \ldots, X_{30})$ be the vector of policy instrument vectors from the first to the last year of the appraisal period (here, from year 1 to year 30). Any specific X describes a strategy. We also recognize that there may be limits on the range of application of individual policy instruments. This is represented by the requirement that the strategies we consider belong to a set K.

For the set of objectives listed above, the general form of the optimization problem was written, in PROSPECTS, as follows:

$$\begin{aligned}
&\underset{X \in K}{\text{Maximize}}\ \mathrm{OF}(X) = \sum_{t=0}^{30} \alpha_t [b_t(X) - c_t(X) - I_t(X) - \gamma_t g_t(X)] + \sum_{i \in M} \sum_{t=0}^{30} \mu_{it} y_{it}(X), \\
&\text{s.t.} \\
&\sum_{t=0}^{30} \omega_{it} y_{it} \le C_i \qquad \forall\, i \in M, \\
&y_{it} \le C_{it} \qquad \forall\, i \in M, \forall\, t \in (0, 1, \ldots, 30).
\end{aligned} \tag{1}$$

Equation (1) can be explained as follows. OF is the objective function. As can be seen, all of its elements are functions of the whole strategy X, and not necessarily only of the policies in place in any particular year, X_t. The first term represents economic efficiency, where b_t is the sum of all benefits in year t; c_t is the sum of all costs in year t; I_t is the sum of capital investments in year t; γ_t is the shadow cost of CO_2 emissions, reflecting national CO_2 targets for year t; and g_t is the amount of CO_2 emissions in year t.

The annual cost and benefit terms are weighted by α_t. Then,

$$\alpha_t = \alpha \frac{1}{(1+r)^t} \tag{2}$$

for all years between 0 and 29. Here, r is a (country-specific) discount rate and α, the intergenerational equity constant, is a constant between 0 and 1, reflecting the relative importance of welfare at present as opposed to the welfare of future generations. So, for these years, α_t is an ordinary discount factor. For year 30,

$$\alpha_{30} = \alpha \frac{1}{(1+r)^{30}} + (1-\alpha). \tag{3}$$

In case not all 30 years are modeled, we may use the last modeled year instead of year 30 as the year with the exceptional term.

The second term reflects the other objectives, where i represents the remaining indicators ($i \in M$); y_{it} is the level of indicator i in year t; μ_{it} is the weight for indicator i in year t; C_{it} is the constraint/target for indicator i in year t; C_i is the overall constraint/target for indicator i (e.g. a financial constraint); and ω_{it} is a weight, possibly but not necessarily a discount factor.

Note that there are two kinds of constraints in the optimization problem (equation (1)). The first says that a weighted sum of the annual values of indicator i should keep within a certain bound C_i, while the second is a constraint on the indicator value at particular points in time. The first type of constraint can be used if we require the annual average of some indicator to reach a certain level, or if we are interested in the cumulative effects. The second type will often be used to set environmental goals for the end of the appraisal period. Not all of these constraints will actually be used in a particular case – if there is no constraint, C_i or C_{it} can be set to a high number.

It may be desirable to add specific financial constraints within the objective function. It may be that financial constraints can be dealt with prior to any optimization process as part of the initial screening of available instruments.

It is assumed that the policy instruments can in the most general case be applied at any level in any one year (t = 0, ..., 30). Thus, for a single instrument there could in theory be 30 different levels in the optimal solution. In practice we have not attempted to solve this theoretical problem for a number of reasons:

- the optimal policy should be easily understood and easy to present to the public and other decision-makers;
- optimization processes become harder to solve as the number of variables is increased, with increased likelihood of finding local optima rather than a global optimum;
- furthermore, each optimization requires more computing time as the number of variables is increased;
- some software packages used cannot represent instruments varying over time to such a fine degree, and many more runs would be required, which would be computer resource intensive.

Given the nature of the objective function, i.e. a very complex "black box" function, the fact that it is a constrained optimization would suggest that the maximum number of variables should be limited as much as possible. Previous experience in SAMI (2000) suggests a limit of around 30 variables would be reasonable before the problem should be broken down further and a decentralized approach utilized. Experience in OPTIMA and FATIMA (May et al., 2000) would suggest that around 10–15 variables is a very complex problem to solve using one of the authors' optimization algorithms, the regression approach. More variables can be used with automated procedures.

While some policy options such as discrete measures being considered in only one year can help cut down the problem, the most efficient and practical method for trimming the problem down is to limit the variation of all the instruments over the evaluation period.

The approach adopted in PROSPECTS (Shepherd et al., 2003) is to specify a piece-wise linear policy profile where policy instrument levels are optimized for two points in time, t_A the implementation year and t_L the long run year. Thus, only the year of implementation t_A and the number of years until a long run value is to be expected will need to be specified.

2.2. *Optimization approaches*

This section presents an overview of various optimization approaches. A non-linear optimization algorithm is needed in order to obtain a solution to the maximization problem (equation (1)). Non-linear optimization algorithms are based on differing principles. An important difference is that some algorithms require the gradient, d(OF)/dx, whereas so-called DUD algorithms ("doesn't use derivatives") do not. In general the former have higher order rates of convergence, whereas the latter are more robust and easier to apply.

The choice of optimization algorithm for the maximization of a given objective function depends on whether the objective function is easily differentiable, and the practicalities of modeling procedures. These practicalities include the required number of model runs and the computer time needed per model run. Some optimization algorithms require the value of the derivative of the objective function for arbitrary values of function arguments. The derivatives of simple functions can often be expressed as analytical functions. For other functions, finite differences can be used to approximate the derivatives. Although algorithms that use values of the derivative are often efficient in terms of function evaluations, it is sometimes cumbersome to establish the routine that calculates the values of the derivatives.

The authors have experience in applying three optimization approaches to both time-marching and equilibrium-based models within the PROSPECTS case studies (Shepherd et al., 2003). The first two approaches treat the model as a "black box," and merely require the model to output the value of the objective function and of the indicators that appear in the constraints at each model run, whereas the third approach additionally requires that the model outputs the value of the derivatives of the objective function. The authors are currently experimenting with a fourth approach using neural networks.

The downhill simplex approach (AMOEBA)

The downhill simplex method, due to Nelder and Mead (1965), was applied via the AMOEBA routine (Press et al., 1990) in projects SAMI, AFFORD, and PROSPECTS. It is a robust and easy to use DUD method in multi-dimensions. It can deal with a set of continuous policy variables, and can be applied with "hard" and "soft" constraints within the objective function.

The simplex method is well suited for optimization of objective functions both with and without constraints on independent variables (policy instruments) and on performance indicators (target constraints). However, the objective functions must be modified to include penalty functions if target constraints are introduced.

The AMOEBA routine provides a robust automated routine, and is easy to link to models where the input of policy instruments and output of the objective function can be easily automated. It is most suitable for fast-running models. Generally, some programming of input and output procedures will be required to link to the AMOEBA algorithm.

A regression-based approach

Another approach to optimization is based on a regression analysis (Fowkes et al., 1998). The basis of the technique is to perform an initial set of orthogonal (in terms of policy instruments) model runs, calculating the objective function for each run. Next, the user has to create a regression analysis of the outputs, and simple calculus is used to predict where the optimum should lie. The user then adds a few more model runs based on this prediction and their own judgement and updates the regression analysis. The process is repeated until the regression model predicts the optimum reasonably well. This method has the advantage of being able to cope with continuous variables and a limited number of discrete variables. The method is not automated, and requires the user to create regression models after adding more model runs to the data set. Previous experience shows that the number of variables should be limited to around 12–15. The method does not require any interface to the transport models, and can allow users to input their own preferred strategies as part of the process, thus accounting for prior belief.

The regression-based approach is well suited to slow models or for those where it would be difficult to automate input and output procedures. No programming is required to implement the regression-based approach.

Constrained optimization algorithms

A constrained optimization algorithm that does not require the user to specify penalties was used in the Oslo case study in PROSPECTS. Such algorithms are available for very general constrained optimization problems, where the constraints can be linear or non-linear equalities or inequalities, or a combination of both. They may be applied equally well to constraints on the ranges of policy instruments or constraints on the indicators (i.e. both model input and model output). Associated with each constraint is a Lagrangian multiplier that can be interpreted as a shadow price. The multipliers are automatically computed by the algorithm.

Such algorithms all apply the gradient of the objective function, either as analytical expressions or in terms of finite differences. Though analytical derivatives of the objective function are usually not available, there is the option to approximate the derivatives in terms of finite differences. Implementations of constrained optimization algorithms are available as part of many commercial software packages and as free FORTRAN and C library routines, available on the Internet, which can be linked into conventional programs.

One problem related to optimization of this kind is the long computation time per function evaluation. This may cause one optimization to take several days to complete. Moreover, a completely new optimization must be run if it is decided to alter the constraints.

One way to circumvent such problems is to generate a grid for the instruments that are available for optimization, making sure that the optimal value of the objective function is likely to be found somewhere within the range of the grid. Polynomial interpolation or regression is then applied to the grid values to approximate the objective function and the indicator responses. Polynomials are well-defined functions that are easily represented directly in any library routine or software package for constrained optimization. Any other constrained optimization algorithm may then be applied to the polynomials instead of the true objective function. With polynomial representation, constraints are easily altered so that new optimizations can quickly be run.

3. Application of formal optimization

This section demonstrates two applications of the AMOEBA optimization approach in a decision-making context. In the first case, formal optimization is used to identify and compare optimal transport strategies among four European cities. In the second case study the method is used to show how constraints affect optimal policy combinations. The latter application is based on only one city (Edinburgh). In both cases the objective function in Section 2.1 was applied to rank the tested transport policy combinations. In all case studies the strategic land use transport interaction model MARS was used (Pfaffenbichler and Shepherd, 2003).

3.1. The strategic land use transport interaction model (MARS)

MARS (Metropolitan Activity Relocation Simulator) is a strategic, interactive land use and transport interaction (LUTI) model. It was developed as a time-saving alternative to traditional four-step transport models. MARS can model the transport and behavioral responses to several demand and supply-side instruments. These impacts can then be measured against targets of sustainability. MARS assumes that land use is not a constant but is rather part of a dynamic system that is influenced by transport infrastructure. The interaction process is modeled using time-lagged feedback loops between the transport and the land use sub-models over a period of 30 years.

Two person groups, with and without access to a car, are considered in the transport model. The transport model is broken down by commuting and non-commuting trips, including travel by non-motorized modes. The land use model considers residential and workplace location preferences based on accessibility, available land, average rents, and amount of green space available. A rather high level of spatial aggregation is used in MARS. In most case studies this means that the wards/districts are chosen as travel analysis zones. The outputs of the transport model are accessibility measures for each zone while the land use model yields workplace and residential location preferences per zone. The interaction between land use and transport modeling components is influenced through a set of policy instruments. For example, new road infrastructure will change the location of housing and workplaces in the long term.

3.2. Identification of optimal transport strategies in four European cities

The objective of the study was to identify city-specific transport policy combinations, and compare them. This should then help in understanding the transferability of policy combinations, and in indicating whether there are instruments which are generically more effective in achieving objectives.

The four European cities under scrutiny were Oslo, Vienna, Leeds, and Edinburgh. The cities vary in size and population and in modal split shares, as shown in Table 1. The Edinburgh model covers the travel to work area beyond the city itself, which has a population of about 450 000.

As can be seen from Table 1, the modal split shares vary significantly. In Oslo, for example, just 17% of all trips are made by slow modes, whereas in Vienna more than 27% of all trips are. The two cities from the UK have similar modal split figures; in both cities more than 50% of all trips are made by private car. In Vienna there is a very high share of environmentally friendly transport means (55% of all trips by either slow mode or public transport) despite the high car ownership rate. One reason for the high share in Vienna is the public transport system: it is mainly based on tramway (233 km) and metro (62 km). In the other cities the public transport system is mainly bus based.

Optimal strategies – CBA-based appraisal

The formal optimization was used to optimize a common set of transport policies to enable direct comparisons between these four cities. The set of instruments was chosen in collaboration with the city authorities. The optimization process applied was an unconstrained optimization in that there are no financial constraints or performance targets. The lower and upper bounds of the common set of instruments were defined as follows:

- Fares: –50 to +100% for peak and off-peak.
- Frequencies: –50 to +200% for peak and off-peak.
- Road charging: 0 to +€10 for a specified cordon around the center.

The instruments were introduced in year 2006 (implementation year) and then linearly increased/decreased until 2016 (long-run year). The results from the optimization process are shown in Table 2.

As can be seen in Table 2, an optimal strategy generally consists of a reduction of public transport fares. In all cities the optimization process suggested a significant decrease in fares in peak and off-peak periods. Sensitivity tests for Leeds showed that a decrease in off-peak fares in 2006 results in slightly higher objective function value, too.

For frequency changes there is not such a clear picture. Here the optimization process suggests a minor reduction of frequency for Vienna in the peak period. This seems to be plausible, since in Vienna the peak headway is 3 minutes for the metro and 7 minutes for tramways. A slight reduction will lead to significant cost savings for operators in combination with marginal disbenefits for the public transport users. In the other cities the optimal combination includes a medium

Table 1
Overview of case study cities

City (state)	Population covered by the model	Area covered by the model (km^2)	Modal split (%) Pedestrian	Public transport	Private car	Car ownership per 1000 population
Oslo (Norway)	396 974	454	17.0	22.7	60.3	400
Vienna (Austria)	1 550 123	415	27.6	27.2	45.3	354
Leeds (UK)	727 700	559	23.4	23.9	52.7	307
Edinburgh (UK)	1 071 768	2305	22.1	24.5	53.4	371

Table 2
Optimal policy combinations

	Fares (% change from do minimum)				Frequency (% change from do minimum)				Road charging (€)			
	Peak		Off peak		Peak		Off peak		Peak		Off peak	
City	2006	2016	2006	2016	2006	2016	2006	2016	2006	2016	2006	2016
Oslo CBA	–29	–32	–23	68	76	93	97	178	4.64	2.24	1.45	0.47
Vienna CBA	–35	–44	–49	–50	–3	–5	198	198	1.89	0.95	0.34	0.00
Leeds CBA	–45	–49	0	–42	190	199	180	200	3.31	5.20	0.02	0.36
Edinburgh CBA	–50	–50	–50	–50	64	33	42	56	2.88	2.46	0.50	0.00

Table 3
Economic evaluation (units: € millions)

	User benefits			Operators					
City	NM	PT	PC	PT	PC	CO_2 benefits	Local pollutants benefits	PVF	OF CBA
Oslo CBA	–611	3 891	–1 715	–1 125	1 480				
Vienna CBA	–197	14 104	–1 091	–6 487	144	14.7	385	–7 750	873
Leeds CBA	–348	7 465	31	–1 931	155	10.2	245	–2 026	890
Edinburgh CBA	–249	4 905	–374	–3 318	238	17.8	229	–3 100	219

Key: NM, non-motorized; PT, public transport; PC, private car; PVF, present value of finance; OF CBA, objective function CBA.

increase of frequency in Edinburgh (64 and 33% for the peak period, and 33–42% for the off-peak period), an intermediate level of increase in Oslo, and a major increase in frequency in Leeds (to the upper bound of +200%).

For the road-charging measure, differing optimal values are derived. In all cities the optimal road-charging value is higher for the peak period compared with the off-peak period. The lower road-pricing values in the off-peak period are a feature of MARS, which assumes no congestion in the off-peak period and therefore no time savings for highways. In all cities (except Leeds) the optimal solution suggests starting with a higher value for road charging in the implementation year 2006 and then reducing it in the long-run year 2016.

Table 3 shows the detailed results for all optimal solutions separated by different transport system user groups.

There are two main groups within the transport system: on one side the users and on the other side the operators and providers, both groups disaggregated by mode. The user benefit (if the sign is positive, disbenefit if the sign is negative) consists of two main sources. The first source is related to direct money savings/cost induced through the suggested policy. For example, a reduction in fares will lead on the transport user side to money savings while on the other hand the operator will lose money. The second source of user benefits/disbenefits is related to travel time savings/losses caused by the optimal policy package. As can be clearly seen, the slow modes are the losers in all of the suggested strategies.

The major gainers are the public transport users, caused on one hand by significant fare reductions and on the other hand by the increase in frequency and the related time savings (mainly waiting times). The situation for car users must be seen in a more differentiated way: on one hand some drivers lose money because of the implementation of a road-pricing scheme (peak and off peak); on the other hand, they gain some time benefits due to higher travel speeds caused by a reduced demand for the car. But, as can be seen in Table 3, the gain in time savings is not enough to outweigh the drivers' losses in money terms.

On the operator side the public transport operator loses in all cities because of the fare reduction and frequency increase (less revenue and higher operating costs), but the road operators gain some money, owing to the introduction of road pricing. More detailed results show a loss of fuel duty to the government from reduced fuel consumption.

The column headed "CO_2 benefits" displays the relative discounted benefits related to CO_2 emissions. In all cases the suggested strategies reduce the negative impacts of CO_2.

The column headed "local externalities" shows the impacts of the optimal policy combination on accidents and local pollutants (volatile organic compounds, nitrogen oxides, etc.). As can be seen, the suggested optimal solutions reduce the negative impacts of these local externalities in all cities, and thus generate benefits for society.

Under the column headed "PVF" the present value of finance is displayed. This value summarizes the net impact of the revenues and costs for operators/providers and the government, and indicates whether the suggested policy combination needs additional funding (if negative) or produces a surplus. The negative PVF values imply that the suggested combination is expensive and therefore may not be feasible.

The column headed "OF CBA" displays the final result, the value of the objective function as described in Section 2.1. It is difficult to spot a pattern here; the objective function value seems to depend on the population and the resulting policy combination. It appears that the higher the population and the higher the frequency changes, the higher the objective function value.

In summary, the optimal strategies are similar in that in all tested cities an optimal policy consists of a significant reduction of fares in peak and off-peak periods, a medium to major increase of frequency in peak and off-peak periods (depending on the existing level of frequency), and a medium level cordon road-pricing scheme in the peak period. Cordon pricing in the off-peak period plays only a minor role. Additionally, the study revealed that there is a system-wide trend toward motorized transport means in the do-minimum strategies over time, and that none of the optimal solutions found so far can reverse this trend. However, implementation of these optimal strategies can ease the burden from car use by reducing the modal share of private car.

3.3. The Edinburgh case study

For the purposes of the Edinburgh case study, $\alpha = 1.0$ has been chosen, i.e. there is no inter-generational equity factor, and a discount rate of 3.5%, so that policy results are in line with current UK practice in terms of discounting procedures. The following sections discuss the model, the scenarios and results with and without financial constraints.

The case study results

The MARS model for Edinburgh is made up from 25 zones, with 14 zones representing the urban area and nine larger zones representing the surrounding regions, as shown in Figure 1.

Optimization of combinations of policy instruments

Two scenarios were examined within this case study, with and without financial constraints. The first scenario allows the variation of public transport frequencies and road capacity across the study area, and road pricing charges based on a cordon around the central area (zones 1, 2 and 12). The second scenario adds peak and off-peak fare changes, which are currently outside local authority control, to the set of instruments, which may be varied. In all cases the levels may be varied by peak and off-peak period, and implemented with a profile over time. As before, this profile is determined by the levels applied in years 5 and 15. Thus, the level of each instrument is optimized by period (peak or off-peak) and by short- and long-run values.

Two financial constraints were considered for each scenario, defined as follows:

(1) the whole strategy should be self-financing – this requires the present value of finance to be positive over the evaluation period (PVF ≥ 0);
(2) the public transport operator should at least break even at the evaluation discount rate.

The first of these constraints allows for cross-subsidies between sectors, whereas the second ensures that the public transport sector is self-financing.

Scenario 1: optimization of frequencies, cordon charges and road capacity changes

The upper and lower bounds for the policy instruments are as follows:

- Road capacity: –20 to +5%.
- Public transport frequency: –50 to +100%.
- Road-pricing cordon: 0 to +€10.

Table 4 shows the percentage changes in peak and off-peak frequencies in years 5 and 15, the peak and off-peak cordon charge in euros in years 5 and 15, and the percentage change in road capacity for all periods and all years – note that the upper bound of 5% is always met so the presentation is simplified to one column. The final three columns show the objective function value, the change in the present value of finance, and the change in the value of finance for the public transport operator.

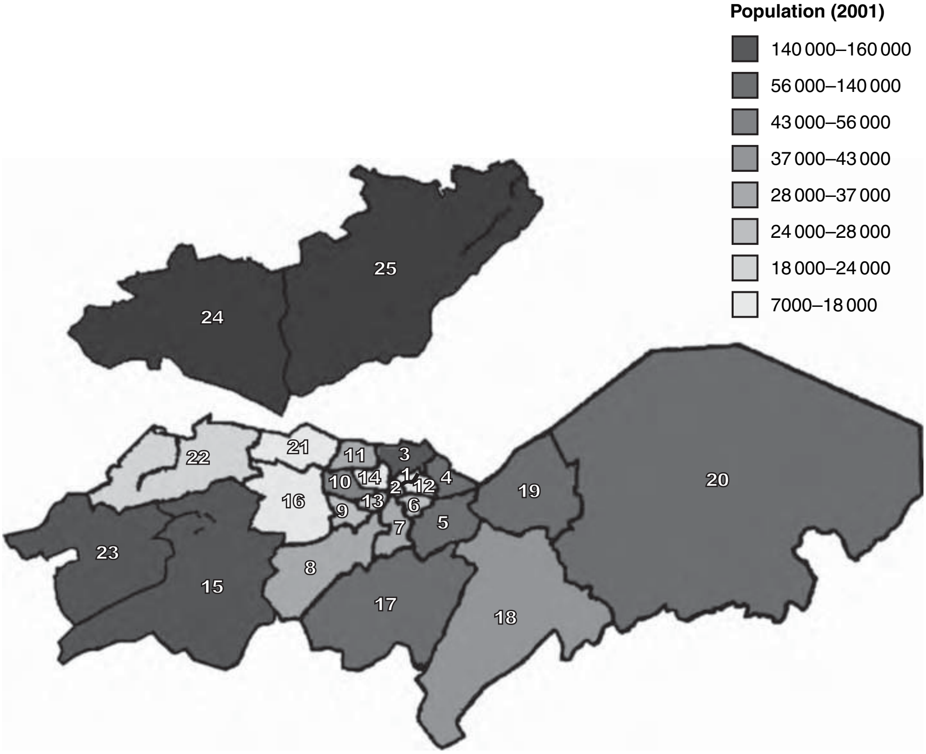


Figure 1. Population and zone numbers for the Edinburgh study area (MARS).

Table 4
Scenario 1 optimization results with and without finance constraints

Run No./ code	Frequency (%)				Road price (€)				Road capacity: all periods and years (%)	OF (€ m)	PVF (€ m)	PT PVF (€ m)
	P05	P15	OP5	OP15	P05	P15	OP05	OP15				
S1	25	50	30	40	3.2	5.75	0.0	0.0	5	**2067**	798	–222
S1-2	25	50	30	40	3.2	5.75	0.0	0.0	**0**	569	551	–379
S1b	20	23	20	23	3.2	5.75	0.0	0.0	5	2013	1073	4.4
S1-2b	10	11	10	11	3.2	5.75	0.0	0.0	**0**	467	1002	3.0

Key: P05, year 5 peak period; P15, year 15 peak period; OP05, year 5 off-peak period; OP15, year 15 off-peak period; PVF, present value of finance; PT, public transport.

The optimal unconstrained solution S1 consists of a 5% increase in road capacity across the whole study area, which is the upper bound for this instrument; increases in public transport frequencies in both periods, which increase over time; and the introduction of peak period cordon charges, which also increase over time. Note that there are no charges in the off-peak period as the model assumes that there is no congestion in the off-peak period – hence the optimal charge is zero.

As the road capacity change is on the upper bound, test S1-2 was conducted to show the effect of removing the additional area-wide road capacity. The objective function value drops by 72%, which shows the important contribution that road capacity improvements make. Other tests have shown that applying a 5% increase in capacity solely to radial movements contributes only 3.5% of the area-wide impact.

Scenario 1: finance-constrained solutions

For the unconstrained optimum solution there is no problem with the first financial constraint, i.e. the revenues collected from the cordon charge more than pay for the capital and operating costs associated with the frequency changes and for the low-cost road capacity changes.

However, for the optimal strategy S1 the public transport operator loses in the region of €222 million over the evaluation period and in the case with no change in road capacities (S1-2) the operator loses around €379 million. Thus, there is a significant subsidy requirement to support the increased frequencies in both cases.

The obvious way to reduce the cost to the operator is to reduce the increase in frequencies. A number of sensitivity tests were conducted with road cordon charges and capacity changes as before to find where the break-even point occurred for the public transport operator. Thus, the objective function has not been optimized with a finance constraint, though this is possible. Table 4 shows the highest scoring combinations that just break-even, coded S1b and S1-2b for with and without capacity changes.

Note that in the first case with capacity increases the long-run change in frequency is +23%, whereas with no capacity increases the long-run change is only +11%. This is because the public transport users benefit from the increased speeds due to increases in road capacity, which bring a greater mode shift to public transport from slow modes which in turn pays for additional services.

Scenario 2: area-wide policies

Scenario 2 adds fare changes as another instrument, with a lower bound of –50% and an upper bound of +200%.

Table 5 shows the optimization results for scenario 2. For the optimal unconstrained solution, S2, the fare changes are optimal at the lower bound of –50% for both periods and all years. The addition of fare changes increases the objective function value by 75% over that for S1.

S2 has increases in frequencies of 60%, which are higher than for S1. This can be explained by the fact that the fare reductions attract more users, who then benefit from the reduced wait times and hence justify greater increases in service levels. Note, however, that the objective function surface is insensitive to changes in frequencies and that the fare and road capacity changes contribute over 80% of the final value. This confirms that fare reductions and capacity changes dominate the solution and are, in this case, on their lower and upper bounds, respectively.

Scenario 2: finance-constrained solutions

The unconstrained optimum resulted in fare reductions of 50% and increases in frequency of 60%. The large fare reductions and increases in frequency mean that both financial constraints are broken this time. Various sensitivity tests were conducted by varying fares and frequencies around the S2 solution to lower costs to operators. Once again the optimization procedure was not used, though this would have been possible.

S2b is the solution that ensures that the public transport operator breaks even. Notice that the fare reductions are now only 5% and the frequencies changes are only +5%. The costs of the frequency changes are balanced by increased fare revenues (despite the 5% reduction in fares). Since the road cordon charges were not revised downward, this solution results in a large present value of finance

Table 5
Scenario 2 optimization results with and without finance constraints

Run number/ code	Fare change				Frequency (%)				Road price (€)				Road capacity: all periods and years (%)	OF (€ millions)	PVF (€ millions)	PT operator change in PVF (€ millions)
	P05	P15	OP05	OP15	P05	P15	OP5	OP15	P05	P15	OP05	OP15				
S2	–50	–50	–50	–50	60	60	60	60	5.0	6.0	0.0	0.0	5	3604	–2556	–3297
S2b	–5	–5	–5	–5	5	5	5	5	5.0	6.0	0.0	0.0	5	2038	1178	0
S2pvf	–24	–24	–24	–24	5	5	5	5	5.0	6.0	0.0	0.0	5	2642	–1.3	–1043

Key: See Table 4.

overall. Fares are then reduced further to find where the present value of finance constraint is just broken. S2pvf shows that a 24% fare reduction can be afforded if a significant cross-subsidy of over €1000 million were possible from the road cordon charge to the public transport operator.

Comparing this scenario with the unconstrained solution, applying a present value of finance constraint reduced welfare by €962 million or 27%, and that imposing a break-even constraint reduced welfare by €1566 million or 43%, and resulted in a positive present value of finance. Both solutions reduced the increases in frequency significantly from 60 to only 5% – this reflects the marginal costs of increasing frequency compared to the costs of fare reductions.

Note that the above results are not necessarily optimal as they have been found via simple sensitivity tests while holding other variables constant – however, they do demonstrate a potential method for finding more financially acceptable strategies using the sketch plan model MARS.

4. Summary and conclusions

The application of formal optimization in identifying optimal combinations of a set of policy instruments has been demonstrated through the case studies. In doing so, the development of an objective function, and the ways in which financial constraints can be imposed, have been illustrated. It has been shown how policy instruments can be defined as continuous variables, within acceptable ranges, and allowed to vary by time of day or over time. The application has been illustrated with one transport model and one optimization routine, but two other routines that can be applied have also been described. While most of the results presented come directly from the optimization process, it has also been shown how sensitivity tests can be used to provide a richer understanding of the contribution of different instruments. The authors have demonstrated the use of response surfaces as a way of enhancing that understanding elsewhere (May et al., 2001).

The results of such methods can be valuable in a number of ways. For an individual city or region, they can suggest a combination of policy instruments to be pursued in the more detailed design of the strategy. Provided that a full enough list of types of instrument is tested, they can highlight combinations that might well have not emerged simply from professional judgment. Conversely, they can identify policy instruments that appear not to be worthwhile. In the authors' earlier work (May et al., 2000), it was found that infrastructure projects rarely formed part of an optimal strategy because, once the other policy instruments were in place, the added benefit from investment was insufficient to justify the high capital cost. One response to this could be to look for lower cost, and hence more effective, design options.

Again within an individual city or region, optimization against different scenarios, or with differing weights within the objective function, can indicate how robust a strategy is, and may also suggest the most appropriate implementation sequence. The application of constraints can indicate the benefits lost as a result if, for example, insufficient funding is available, or acceptability constraints limit the use of a particular instrument. The application of constraints related to targets for specific objectives can help in checking that those targets are realistic, and not imposing undue burdens on other objectives (Emberger et al., 2003).

A comparison across cities or regions, provided that the same transport model is used, can help in understanding the transferability of policy recommendations, and in indicating whether there are some policy instruments that are generically more effective in achieving objectives. In a similar vein, comparisons of constrained strategies can indicate whether there is a need for government action to remove that constraint, and can indicate the potential benefits of doing so. In earlier work by the authors this approach was used to make tentative recommendations on the importance of public transport fares and frequencies, low-cost road capacity increases, and road pricing as solutions to urban transport problems; the authors also demonstrated the importance of maintaining flexibility in the way in which local government can influence public transport operation.

References

Emberger G., A.D. May and S.P. Shepherd (2003) "Method to identify optimal land use and transport policy packages, introduction of a indicator/target based appraisal approach," in: *Proceedings of the 8th International conference on Computers in Urban Planning and Urban Management*. Japan.

Fowkes, A.S., A.L. Bristow, P.W. Bonsal and A.D. May (1998) "A shortcut method for optimisation of strategic transport models," *Transportation Research A*, 32:149–157.

May, A.D., M. Roberts and P. Mason (1992) "The development of transport strategies for Edinburgh," *Proceedings of the Institution of Civil Engineers: Transport*, 95:55–59.

May, A.D., S.P. Shepherd and P.M. Timms (2000) "Optimum transport strategies for European cities," *Transportation*, 27:285–315.

May, A.D., S.P. Shepherd, H. Minken and G. Emberger (2001) "The role of response surfaces in specifying transport strategies," *Transport Policy*, 8:267–278.

Minken, H., D. Jonsson, S.P. Shepherd, T. Jarvi, A.D. May, M. Page, A. Pearman, P.C. Pfaffenbichler, P. Timms and A. Vold (2003) *A methodological guidebook*. Oslo: Institute of Transport Economics.

Nelder, J.A. and R. Mead (1965) "The downhill simplex method," *Computer Journal*, 7:308–313.

Pfaffenbichler, P. (2003) "The strategic, dynamic and integrated urban land use and transport model MARS (Metropolitan Activity Relocation Simulator) – development, testing and application," unpublished dissertation. Vienna: Vienna University of Technology.

Pfaffenbichler, P. and S.P. Shepherd (2003) "A dynamic model to appraise strategic land-use and transport policies," *European Journal of Transport and Infrastructure Research*, 2:255–283.

Press, W.H., B.P. Flannery, S.A. Teulosky and W.T. Vetterling (1990) *Numerical recipes*. Cambridge: University Of Cambridge.

Proost, S., K. Van Dender, C. Courcelle, B. De Borger, J. Peirson, D. Sharp, R. Vickerman, E. Gibbons, M. O'Mahony, Q. Heaney, J. Van den Bergh and E. Verhoef (2002) "How large is the gap between present and efficient transport prices in Europe?" *Transport Policy*, 9:41–57.

SAMI (2000) *Strategic assessment methodology for the interaction of CTP-instruments, Deliverable 3. Optimisation of CTP instruments*. Luxembourg: DG7 Transport Research (Strategic), European Commission.

Shepherd, S.P., T. Järvi, A. May, M. Page, A. Pfoffenbichler, P. Jimms and A. Vold (2003) *PROSPECTS deliverable 8. Forecasting and analysis*. Luxembourg: EESD, European Commission.

Part 6

NATIONAL STUDIES

This Page Intentionally Left Blank

Chapter 40

TRANSPORT POLICY WITHIN APEC

CHRISTOPHER FINDLAY
Australian National University, Canberra

CHRISTOPHER KISSLING
Lincoln University, Canterbury

1. Introduction

This chapter reviews the work program on transport policy that has been undertaken by Asia Pacific Economic Cooperation (APEC), and which is still in progress.

APEC's membership includes 21 economies on both sides of the Pacific. The current membership is Australia, Brunei Darussalam, Canada, Chile, China, Hong Kong, Indonesia, Japan, Korea, Malaysia, Mexico, New Zealand, Papua New Guinea, Peru, the Philippines, the Russian Federation, Singapore, Chinese Taipei, Thailand, the USA, and Vietnam.

The significance of transport for APEC's mission is the starting point of the chapter. The ways in which the members work cooperatively within APEC on transport policy questions are also reviewed, although the chapter does not provide a detailed review of transport policy in each of these economies. Following a review of some current transport sector challenges, the structure of APEC and the manner in which it operates will be examined, including some comments on some institutional challenges that it now faces. APEC's work on transport policy is the topic of the following section. Projects are grouped according to classifications commonly used by APEC, and comments are offered on the mix of projects, and their relative significance. These points are illustrated by a short review of specific collective projects and some results to date. The chapter concludes with comments on the importance of APEC's work on transport policy, the significance of that work for APEC's own agenda, and on the requirements for greater effectiveness of APEC's work on transport policy.

2. APEC and transport policy

APEC was established in 1989 as "an informal intergovernmental process" (Soesastro, 2003) whose role was to promote economic cooperation for the

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

purpose of building a "prosperous Asia Pacific through free and open trade and investment" (APEC, 2003). Since its creation,

> APEC has worked to reduce tariffs and other trade barriers across the Asia-Pacific region, creating efficient domestic economies and dramatically increasing exports. Key to achieving APEC's vision are what are referred to as the "Bogor Goals" of Free and Open Trade and Investment in the Asia-Pacific by 2010 for Industrialized Economies and 2020 for Developing Economies. These goals were adopted by APEC Economic Leaders at their 1994 meeting in Bogor, Indonesia. APEC also works to create an environment for the secure and efficient movement of goods, services and people across borders in the region through policy alignment and economic and technical cooperation. This cooperation also helps to ensure that the people of the APEC region have access to training and technology to take advantage of more open trade and investment.
>
> (APEC, 2003)

For an organization seeking to free up trade between its members and forge closer cooperative arrangements, it has been imperative that the trade bridges be strengthened in a variety of ways. Without the means of transport, there can be no trade. If the conduct of international transport and logistics is frustrated by antiquated laws, inconsistent rules and regulations, ineffectiveness through lack of skilled human capacity, physical bottlenecks at key points of interchange, and lack of specialist equipment needed for efficient operations, then trade between economies cannot reach its potential. The following paragraphs offer more detail on each of these challenges, and in the following section examines whether the APEC structure and mode of operations are relevant to meeting them.

Reaching sufficient levels of physical capacity in the transport system is one of the challenges facing the APEC membership. The capacity required includes specialized maritime transport for dry and liquid bulk cargoes and, in the case of an increasingly diversified range of products, in standardized containers shipped in large purpose-built container vessels. Whereas air cargo volumes cannot match those transported by sea, the value of air cargo is assuming ever-greater significance in trade balances, and is especially important for time-sensitive and perishable items. The right mix and volume of capacity in various modes depends on investment decisions in the private sector, which in turn depends on the extent of openness of markets for transport services.

All trade movements are in response to demands that arise from entrepreneurs and organizations seeking to realize profits from international business. They expect the performance of all providers of transport services, mobile and static, to be aligned with global best practice in the industry. Clearly, when APEC was formed, huge disparities in the quality of service in transport existed. There remain disparities, but they are certainly less gross than at the outset, and the steady pressure to move to best practice remains.

Not just physical but also human capacity is critical. Training seamen, officers, aircrew, stevedores, and airport ground staff are vital if the three goals of safety, security, and efficiency are to be achieved. All modes of transport need skilled employees. While substantial expertise exists within APEC, it is not uniformly distributed, nor is it likely to be sufficient for the magnitude of the task ahead. Encouraging education and training for the transport sector cannot be paid just lip service: it needs persistent promotion, along with a more vigorous approach to recruiting able women into the sector.

Just as the ships and aircraft have become larger and more sophisticated in their design and economics of operation, so too the supporting infrastructure has had to be improved to ensure the smooth transfer of products through gateway ports and between all modes of transport involved in the domestic carriage of goods. Containers need to penetrate sites well removed from harbors and airports. Only those ports that are strategically located with respect to the corridors of international commerce and that have sufficient depth of water to cope with the size of vessels entering service can hope to remain important nodes in the global system. Major airports have required upgrades, and even relocation, in order to provide ongoing capacity to avoid congestion.

Standards in transport can vary within and between member and non-member economies. Inconsistent standards can add to the costs of doing business and inhibit trade flows. It is important from the customer point of view, and from the perspective of transport companies seeking to operate worldwide, to move to global standards, and not to create a new set of regional standards. There is a plethora of documentation needed to satisfy not just commercial transactions but also the requirements of governments seeking to control international trade for revenue gathering, protection of domestic production, bio-security, and for statistical purposes. If this documentation is not facilitated through the adoption of electronic data exchange and remains tied to hand scribing, it will remain inefficient and error prone. Information systems are a vital adjunct to trade.

3. APEC structure and operations

APEC's work is divided into the areas of trade and investment liberalization, trade and investment facilitation, and economic and technical cooperation. Liberalization removes policy barriers to trade and investment across borders. Facilitation is designed to reduce the costs of cross-border business transactions. Economic and technical cooperation, or capacity building, is

> about the sharing of experiences and resources among APEC members aimed at enhancing the ability of all sections of society to benefit from effective participation in the global economy. It recognizes the reality of the enormous disparities that exist within the Asia Pacific in terms of the capacity to benefit from globalization.

(APEC Business Advisory Council, 2000)

These areas complement each other (APEC Business Advisory Council, 2003). They also match the broad categories of challenges just described in the transport sector, including the role of liberalization in providing capacity and moving the sector to world's best practice, the importance of dealing with gaps in human capacity, and action to deal with inconsistencies in standards and other government regulation. APEC has a structure and agenda that is relevant to the challenges in the transport sector.

The work of APEC is completed within each member economy. There is no central bureaucracy to manage the cooperative work. There are no formal agreements between economies or a mechanism for resolving disputes between them over the implementation of APEC initiatives. The organization's success depends on the activity that it stimulates within the members: "the forum succeeds by promoting dialogue and equal respect for the views of all participants, making decisions based on consensus to achieve its free and open trade and investment goals" (APEC, 2003). The effort to liberalize trade and investment without the use of formal and binding agreements is unusual among multilateral economic institutions. The diversity amongst the APEC membership, and their different stages of development, the histories of relationships between them, and, most importantly, the common interest in the preservation of sovereignty all require the use of less formal types of cooperation at this stage of the institution's development.

The APEC leaders meet once a year in the economy of the chair of APEC. That position rotates each year. Meetings of ministers precede that of the leaders. The main ministerial meeting includes both trade and foreign ministers, but there are also sectoral ministerial meetings. A committee of senior officials supervises the work program under the direction of leaders and ministers. Officials from member economies lead its subcommittees, and the work is managed in and from member economies. There are now hundreds of meetings each year at these working levels. An international secretariat in Singapore supervises the flow of information and manages a very small budget for collective work.

The manner in which APEC's goals are implemented is laid out in its Osaka Action Agenda. Members report their progress toward the goals in Individual Action Plans (IAPs), which refer to 15 "issue areas." Other members review these IAPs regularly and collectively. In addition, there is also a plan for collective actions in the same issue areas.

The members adopted a number of principles designed to guide their work. These include a commitment to non-discrimination, meaning that reductions of barriers to trade achieved through APEC are available to both members and non-members of APEC. This commitment reflects the economic links among the members, and the value to each other of liberalization by any one of them, but also their common interest in economic relationships with the rest of the world. Other principles refer to the importance of a comprehensive treatment of all barriers to trade and investment, consistency with the World Trade Organization (WTO), endeavoring to make comparable efforts and with a simultaneous start, but allowing for different timetables, not taking measures to increase levels of protection, and increasing transparency. The Bogor goals and these principles are the core elements of an "APEC vision" of how economic cooperation in the region for the purpose of promoting development can be organized.

4. Challenges for APEC

The scope of work within each of the APEC categories has changed considerably since the late 1980s. Ministers and leaders now face a set of issues that probably were not envisaged by APEC's founders. They also sit on top of a burgeoning work program, yet one, generally, with less profile than a decade ago.

A number of commentators have been concerned that the political support for the APEC vision of regional economic cooperation has waned. The Pacific Economic Cooperation Council, for example, points to new uncertainties in the Asia–Pacific region, including the proliferation of regional and discriminatory trading arrangements and the impact of and the government responses to the greater perception of the risk of terrorism. However, the Pacific Economic Cooperation Council also concludes unequivocally that the APEC vision, as just described, is now even more important, as member economies respond to these new uncertainties (Pacific Economic Cooperation Council, 2003).

One risk in APEC is that it adopts too many projects, each of which has a relatively small impact. Parsons (2003) observes that the APEC system has become cumbersome and is too dependent on the "bubble up" approach to developing new projects. He says that elevating the leadership of presidents and prime ministers would help ministers to initiate and implement work, rather than tick off the work of officials, and help to sift out lower-priority issues for which APEC does not have the time and resources. The management of projects that APEC undertakes can be improved, especially those in relation to economic and technical cooperation. The APEC Business Advisory Council, established by leaders to advise on APEC priorities and review progress, has, for example, been calling for some time for greater focus and accountability in that pillar of APEC.

APEC leaders met in Thailand in October 2003. According to their declaration, they too discussed "the need to make APEC more efficient and responsive to all stakeholders [and they] instructed ministers to explore the issue and report on progress by the time [they] meet in Chile in 2004." It has been proposed to strengthen the secretariat with an executive director who "speaks for APEC," to add more senior people with management responsibilities, and to provide a greater capacity in the Secretariat to do "substantive policy research" (APEC International Assessment Network, 2003). The origins of ideas are evident in the problems just described. But at this stage, their adoption involves some risks. For example, a stronger secretariat creates another set of interests in the APEC process. It may also lead to free-riding on the secretariat by members, whereas in the original model APEC itself was not an implementation agency. There is also a risk that secretariat interests will impose a greater degree of uniformity on APEC projects than the membership can sustain. In this respect, more attention to the specification of a "pathfinder approach," in which some move faster than others, is a priority for further work; this question will be returned to in the discussion on transport sector work, below. Finally, there may not be enough resources available for the secretariat to take on more jobs.

Soesastro (2003) urges further thought on the questions of the longer-term institutional development of APEC. He makes the important suggestion that the work should be linked to an examination of the question of whether APEC should move away from being a voluntary organization to a binding one. Soesastro stresses, however, that success in such a radical change depends on making the current model work better. That point highlights Parsons's suggestion, noted above, of even greater direction from the top within the current structure.

A new development in APEC has been its focus on security questions. About one-third of the declaration of the 2003 Bangkok meeting referred to security, with a focus on trans-national terrorism, but also with mention of weapons of mass destruction. APEC, because of its basis in cooperation between economies rather than states, cannot take on the full role of a formal security institution. But it has been for some time a *de facto* security organization (McKay, 2003). When the leaders come together, they inevitably talk about issues that concern them, including security. Much of the APEC agenda related to economic integration has a security dimension. Proposals for new security arrangements, like the various counter-terrorism proposals, can sensibly be examined in APEC, because of the elevation of security issues into so many economic policy initiatives, including as illustrated below in a more detailed discussion of APEC's transport work program.

5. Transport policy work

APEC economies make commitments to liberalization of trade and investment in transport services in their IAPs. The IAPs refer to four sectors: maritime, air, rail, and road transport. Members are invited to make commitments with respect to four areas: operational requirements, licensing and qualification requirements of service providers, foreign entry, and the application of the principle of non-discrimination in transport policy. The IAPs show current policy, recent changes in policy, and further changes proposed. The IAPs contain a mass of information about the policy of each member – four sectors by four areas by 21 economies, i.e. up to 336 cells of qualitative information, as well as additional material showing the rate and direction of change.

The WTO negotiations and commitments in the General Agreement on Trade in Services (GATS) also refer to transport services. APEC's principle of non-discrimination means that its work program is highly complementary to its members' interests in the WTO, where the same principle applies. That principle is also critical to meeting the challenges in the transport sector, including the pressure to move to global best practice. However, no summary assessment is yet available, for example providing a comparison of commitments in the IAPs with those made in the WTO negotiations. Work in APEC should add, in a manner that is credible to business, to that already achieved in the WTO. Assessment of the IAP commitments by this criterion is a topic for further work.

The Transportation Working Group, which has been operating since 1991, manages APEC's collective transport work program. The current portfolio of 19 projects is reviewed in the Appendix (organized according to the group's three sub-committees), and summarized in Table 1 by mode and type of activity. The projects in Table 1 are grouped by the topics of

- liberalization (two projects);
- facilitation (11 projects) – these can be divided into three groups, one related to mutual recognition work (two), another related to information exchange (six), and the third related to the implementation of new technology (three);
- human resource development projects (six projects).

The group has an ambitious program, managing 19 different projects. Not all economies are necessarily active participants in all projects (details are not available on the extent of participation of economies in each project), and further comments are offered below on the effects of some groups of economies trying to move faster than others. However, the large number of projects relative to the membership highlights the questions raised earlier about the value of further effort to prioritize and concentrate the work program in APEC.

The mix of projects illustrates the capacity of the APEC work program to cover the sorts of challenges now evident in the transport sectors. However, the portfolio of collective projects has a bias toward facilitation work, followed by human resource development. Few of the collective projects listed in Table 1 are designed specifically with liberalization objectives in mind – there is one on air transport and one on non-tariff barriers in transport. Partly this is because liberalization in the transport sector is also covered in the IAP process just described. As border barriers fall, vigilance that other forms of trade barrier do not replace them is important, and the Transportation Working Group has taken on this task. The effort to work on the liberalization of air transport is significant because of the lack of progress on liberalization in that sector in the WTO. However, indirectly nearly all projects, it can be argued, contribute to liberalization. Work on policy transparency (as in the maritime project), for example, is often argued to be a first step in the evaluation of and then change in policy.

Table 1
APEC collective work program on transport (a)

Topic	Sector: Road	Air	Maritime	Cross-sectoral and multimodal
Liberalization		Identifying best practices for promoting more competitive air services (1.3)		Investigating non-tariff measures in transport (1.8)
Facilitation: mutual recognition	Mutual recognition of vehicle standards (1.1)			Mutual recognition of transport professional qualifications (3.1)
Facilitation: information exchange	Exchanging information on road safety strategies (2.4)		Raising the transparency of maritime policies (1.4)	Exchanging information on security policy (2.1)
			Developing proposals for improved port efficiency (1.5)	
			Sharing best practices for vessel arrival, stay, and departure (1.7)	
			Making recommendations on maritime safety (2.2)	

Facilitation: information technology related				Promoting paperless trading in transport (1.2) Identifying standards for Intelligent Transport Systems (2.3) Implementing satellite navigation and communications (2.5)
Human resources		Aviation law training (3.4) Air services negotiations (3.5)	Seafarers training (3.3)	Developing intermodal transportation skills (1.6 and 3.6) Training and development for women in transport (3.2)

Note: (a) Numbers in parentheses refer to the APEC Transportation Working Group programs (see the Appendix).

In the facilitation group, there are two important projects on mutual recognition, one focusing on vehicle safety standards and the second on professional qualifications. These are valuable projects of the type where a regional organization is likely to have a comparative advantage. The agreement to harmonize with the United Nations Economic Commission for Europe vehicle standards and the development of a mutual recognition agreement on automotive certification, are in themselves significant achievements. They are consistent with the target to move to global best practice. Other projects regarded by participating officials as significant include some of the training programs.

The discussion so far refers to projects in the top two rows of Table 1. They refer to matters of market integration, that is, liberalization and mutual recognition. They account for four out of the total of 19 projects. The bulk of the remainder of the work includes projects related to information exchange, promoting technological change, and building human resource skills.

Work on training and human resource development is an important part of the capacity-building program in APEC. Training is a vehicle for sharing policy experience, but it is not the only element. Nearly all projects in the table have a capacity-building component, such as those exchanging information on best practices for regulation or infrastructure management. The use of new technology often complements the facilitation agenda in APEC. However, some attention to the structure of these sorts of projects will be important. Successful capacity building involves three steps, of design, delivery, and follow-up or "after-care." For example, in the development of new institutions relevant to regulation of the transport sector, the value of that institution and key principles in its design must first be established. Then the institution has to be built and the staff trained. APEC has a strong track record in these two steps. Its performance in the third step is generally weaker.

There are two other features of the work program. One is its emphasis on multimodal projects and the other is the recent increase in interest in work on security in transportation in all modes. The work on multimodal transport is recognition that many issues cross the modal boundaries. International transport more often than not requires good integration throughout supply chains. That coordination demands consideration of cross-sectoral matters perhaps more so than within individual economies, wherein modal management can be bound by tradition. International transport operators are now being challenged to reflect more on intermodal cooperation. Work multimodal transport in APEC is more significant because of lack of progress in this field by the WTO.

There has been a noticeable increase in the work on security in transportation for all modes. This topic is likely to dominate the collective work on transport by APEC over the next few years. Ministerial statements show that security is a priority issue, reflecting the leaders' interest in these problems. Clearly, the September 11, 2001 terrorist attacks in the USA have been a driver. Security has become more important in the ongoing work, but as yet has not deflected the resolve to continue progress toward achieving more open trade despite the burden of additional security arrangements and costs.

Work on "track and trace" technology for containers – part of the work on intelligent transport systems – is an example where commercial customer service interests and government concerns for added security complement each other. Industry has been moving to provide greater transparency in supply chains. That will inform security requirements, but the adoption of various competing technologies and imposition of certain procedures has led to increases in cost and a subsequent debate over who should pay. These are topics for further research attention. APEC provides an important forum for the discussion of the results of that work and for the design of cooperative programs to assess and then meet the new security requirements at least cost.

Business participation in APEC is important, because of APEC's ultimate objective of market-driven integration and business' contribution to that process, as well as business' appreciation of and therefore capacity to advise on the key impediments to integration. Business interest in APEC is also a test of the effectiveness of the work program. One mechanism for tapping business information is through the use of competitive bidding by consultants for projects. Nevertheless, APEC is a government-to-government forum. Most of the people who attend meetings are government officials. Some consultants and academics become involved through specific projects that inform the specialist and plenary sessions; but they are not heads of delegations. Nor are business leaders much involved despite the recognition that it is valuable for the commercial sector to be engaged.[a]

There are proposals for sub-groups of APEC to move faster on reaching agreement on some policy changes. There is an example of this approach from within the transport program, where a sub-group of like-minded APEC members (Singapore, Brunei Darussalam, Chile, New Zealand, and the USA) established a Multilateral Agreement in Liberalization of Air Transport (MALIAT). Discussion of the agreement began in APEC, but ultimately was signed only by the sub-group of members plus Peru (2002), which has subsequently given notice of withdrawal. Samoa (2002) and Tonga (2004) – non-members of APEC – have also joined. It is not regarded as an APEC project, in the same manner as other collective activities, and is therefore not listed in Table 1. The agreement (reviewed by Findlay, 2003) has some desirable features, including an approach to capacity and price regulation, to network design and aircraft management, and to resolving business issues. Findlay concludes that its designation clause (or "rule of origin") is more liberal than those in other "open skies" agreements, although it does not meet a more common GATS standard. Cabotage is included as a protocol that has been signed by Chile, Brunei Darussalm, New Zealand, and Singapore.

[a]Kissling (1999) noted a perspective from the private commercial sector.

The main problem with MALIAT, according to Findlay's assessment, is its accession clause, which runs the risk of adding new layers of discrimination, and which fails to support an agreement of this type becoming a stepping stone to a more widespread liberal regime. The momentum for adding new members can stall under the rules applied here. Further, there is also little evidence of enthusiasm by others to join, since the agreement appears to offer little compared to the existing bilateral agreements. An ability to join the spokes in an existing set of US hub-and-spokes open skies agreements does not appear to be a sufficient foundation for interest among the non-members.

6. Conclusion

The transport sectors of the APEC members face a number of challenges if they are to support the members' ambitions to participate in the global economy. These challenges can be divided into three types, related to matters of liberalization of trade and investment in transport services, to the value of adopting global best practice in the transport sector in order to facilitate the growth of trade, and to avoid the bottlenecks that might be associated with the lack of human or technical capacity in the sector. This set of challenges matches the structure of the work agenda in APEC, which runs across work on liberalization, facilitation, and economic and technical cooperation. A globally competitive transport sector is critical to the ambitions of the APEC membership, and the work program that APEC members have designed offers the coverage required to reach that target.

A distinguishing feature of the APEC work program is the focus it puts on economic and technical cooperation. APEC works to support economic integration driven by markets, which depends on the confidence among policy-makers to promote trade and investment. That confidence comes from knowing that policy-makers in other economies are also making changes, and that those changes will provide benefits for each other member of the group. It also comes from learning of the experience of other economies, and how they responded to the various impediments to reform. Many APEC activities, particularly in the capacity-building programs, are designed for these purposes. These contributions of APEC are not obvious and are hard to measure, but they are a critical part of its operation. They are also vital in an environment where pressure is applied through market forces, especially through competition for investment, and by other governments in negotiating processes for accelerated policy change.

The value of this sort of work is evident in APEC's transportation program. Collective work in that sector includes work on human resource development, which is an important part of capacity building. But also the work on sharing of information on best policy practices and the promotion of the adoption of new technology complements the efforts by individual economies to make commitments to open markets in transport services. Evaluating the extent of commitments to trade and investment liberalization and facilitation made at the individual economy level by APEC members, and the interaction of those commitments with the collective work program, is a topic for further work.

APEC work has complemented that in global bodies. For example, the work on air transport liberalization and impediments to the development of multimodal systems offers the chance to its members to go beyond efforts in those fields by the WTO. The complementarity of the liberalization program in APEC in transport services with negotiations in the WTO is a topic for further work. APEC's work on moving to global safety standards in automobiles is also significant.

However, there are some risks in the APEC structure, in terms of the mix and levels of activity. The first is that there will not be enough attention to projects that contribute directly to market integration, and which can be more challenging to adopt for their domestic political effects. Second, there is a risk that a cooperative process like that which operates in APEC will carry too many projects, each with relatively thin participation by the membership and therefore with an overall lesser degree of impact. On the face of it, these risks are evident in the transport work program, with relatively few projects focusing directly on market integration and with a relatively large portfolio. It is important for the transport sector to respond to the challenge posed by APEC leaders for officials to establish new processes for establishing and implementing priorities in APEC.

Another long-run issue for APEC is the manner in which it implements its work. APEC members can endorse activities or propose to adopt particular arrangements or policy structures. At this stage, however, they are not bound to do so. The use of a non-binding approach was essential for APEC's establishment, as a direct movement to complete harmonization was not possible in all areas. But the role of the principle of non-binding commitments in APEC continues to be debated. Some members seek to move faster than others to reach binding commitments, and there are examples of smaller groups of economies adopting such agreements among themselves. One example of a pathfinder approach in the transport sector was reviewed, and its lessons are that, if such an approach is to lead to genuine liberalization rather than the addition of new layers of discrimination, then all members should endorse the principles of the agreement and all should eventually be able to join on the same terms as the foundation members. These are important principles for the operation of sub-group strategies in APEC.

Appendix. APEC work on transportation

The APEC Transportation Working Group organizes its work programs under three steering committees involving collective actions and joint activities. The following is extracted from APEC documentation related to the 22nd APEC Transportation Working Group Meeting held in Busan, Korea, September 1–5, 2003 (TPT-WG22/PLEN/8). The material is presented under the headings of the three sub-committees that manage the work.

1. Steering Committee on More Competitive Transportation Industry (including Infrastructure)

1.1. *Road Transport Harmonisation*. The aim of this project is to establish mutual recognition for automotive products among between APEC members, who have agreed to employ the safety and environmental standards adopted by the United Nations.

1.2. *Electronic Commerce*. The purpose here is to promote paperless trading and the adoption of electronic commerce in both business and government transactions throughout the transport sector in the region. The target is to eliminate paper requirements for key messages relevant to international transport and trade by 2005.

1.3. *Air Services*. The Auckland APEC Leaders Declaration refers to eight steps for more competitive air services, including work on air carrier ownership and control, tariff regulation, business matters, airfreight, multiple designation, charter services, airline cooperative agreements, and market access. The purpose of this project is to monitor their implementation.

1.4. *Maritime Initiative*. The aim of this work is an efficient, safe and competitive operating environment for maritime transport in the APEC region. The main activity is to improve the transparency of maritime policies.

1.5. *Port Experts Group*. The exchange of information and expertise among port experts and programs is expected to improve port efficiency. Ongoing projects include the operation of a port database and work on container throughput forecasting.

1.6. *Intermodal Transportation*. A three-phase Intermodal Transportation Skills Workshop has been implemented to achieve a more efficient and secure transportation system, increased trade, an environmentally sustainable transportation system, a more streamlined regulatory and institutional environment including harmonized standards, rules and regulations, and a more educated transportation workforce. Projects include the identification of competency standards for perishable goods handling.

1.7. *Efficiency in the Facilitation of International Seaborne Trade.* This project seeks to increase the efficiency of processes and procedures related to vessel arrival, stay, and departure in the South-East Asia region, through streamlined and simplified customs, immigration, quarantine, and port authority procedures. Best practice models are sought that will ensure cost savings and more competitive shipping.

1.8. *Investigation of Non-tariff Measures in the Transport Sector.* The goal of this work is to establish a best practice guide to the elimination or reduction of non-tariff barriers.

2. Steering Committee on Safe and Environmentally Friendly Transportation (including New Technologies)

2.1. *Transportation Security Experts Group (TSEG).* This group aims to provide a focus on transportation security and a forum for exchange of information regarding security policies, practices, and new technologies. This involves all modes of transport. It will work on maritime piracy in the region. It looks for procedures that will promote violence-free shipping. It has examined research on detecting explosives. It is identifying security training needs and the means to meet them. It also looks to assist economies increase the capacity of their communities to survive and recover from terrorist attacks through risk management decision support tools and procedures. It has a focus on all areas of aviation security, particularly the implementation of relevant technology. There is an accent on training, information sharing, and liaison with other bodies such as the International Civil Aviation Organization (ICAO).

2.2. *Experts Group on Maritime Safety.* This group recommends improvements in maritime safety. It has produced a booklet, *Maritime Safety – Charting a Course towards Safer Shipping in the APEC Region*, that seeks to eliminate substandard shipping, and it promotes harmonized implementation of international maritime conventions. Accreditation of Seafarer Manning Agencies in the APEC region looks to improving hiring procedures of seafarers, and consequently higher standards of ship crews benefiting safety. It is reviewing existing maritime training in the region, and will include consideration of maritime environmental issues.

2.3. *Intelligent Transport Systems (ITS) Experts Group.* This group identifies ITS standards requirements that are APEC priorities, facilitates the establishment of such ITS standards by ISO TC204, promotes the universal use of those ITS standards, and shares information regarding ITS developments. The benefits expected accrue to the public from faster, safer transportation of exports and imports with decreased pollution, congestion, and possible delay of roadway infrastructure requirements. Track and trace technologies are being investigated, to seek best practice that can be applied to intermodal transport operators.

2.4. *Road Safety Experts Group (RSEG).* This group looks for strategies that can improve road safety in the region, including initiatives leading to standards and measures that constitute best practice.

2.5. *Satellite Navigation and Communications.* The aim of this group is to develop a plan of action to facilitate the implementation of satellite navigation and communications systems in the region, working with the ICAO and International Maritime Organization as appropriate.

3. Steering Committee on Human Resources Development (including Training, Research and Education)

3.1. *Mutual Recognition of Transport Professional Qualifications.* This group considers measures that would promote transparency in regulations, resolve differences in conformity assessment, and facilitate the mobility of transport personnel by encouraging the mutual recognition of professional qualifications.

3.2. *Gender Project Team.* The team aims to identify and remove barriers to women's full participation in the transportation sector. This includes identification of training and development initiatives for women in the transportation sector.

3.3. *Seafarers Training.* Projects by this group are designed to promote the provision of high-quality instruction for seafarers consistent with the requirements of the revised provisions of relevant international conventions. Standardization of training will assist mutual recognition of certificates and professional competence of seafarers.

3.4. *Aviation Law Training in APEC Economies.* This group seeks greater understanding of the role played by aviation law in regulating civil aviation in the international arena.

3.5. *Air Services Negotiations.* Aviation-rights negotiators are the focus of this project, the aim of which is to provide participants with a greater appreciation of the tools and techniques used in negotiating air service agreements.

3.6. *Identification of Needed Intermodal Skills and Developments of Required Training Programs.* Some APEC economies lack methods for developing skills in intermodal transport. This project aims to fill that gap.

References

APEC (2003) *Asia Pacific economic cooperation*. Singapore: APEC Secretariat.
APEC Business Advisory Council (2000) *Report to APEC leaders*. Manila: ABAC.
APEC Business Advisory Council (2003) *The first decade since Bogor: a business assessment of APEC's progress*. Manila: ABAC.
APEC International Assessment Network (2003) "Remaking APEC as an institution," in: R. Feinberg, ed., *APEC as an institution, multilateral governance in the Asia–Pacific*. Singapore: Institute of South East Asian Studies.
Findlay, C. (2003) "Plurilateral agreements on trade in air transport services: the US model," *Journal of Air Transport Management*, 9:211–220.
Kissling, C.C. (1999) *The TPT-WG means business* (http://www.apec-tptwg.org.tw/TPT/tpt-main/shepherd-page/tpt-wg-16-final-papers/presentation-by-prof-kissling.htm).
McKay, J. (2003) "APEC's role in political and security issues," in: R. Feinberg, ed., *APEC as an institution, multilateral governance in the Asia–Pacific*. Singapore: Institute of South East Asian Studies.
Pacific Economic Cooperation Council (2003) *The Brunei Declaration*, PECC, Singapore (available from www.pecc.org).
Parsons, D. (2003) "Will it be everyone for themselves after Bangkok?" *APEC Economies Newsletter*, 7, No. 10 (http://apseg.anu.edu.au/pdf/newsl/APEC7_10.pdf).
Soesastro, H. (2003) "APEC's overall goals and objectives, evolution and current status," in: R. Feinberg, ed., *APEC as an institution, multilateral governance in the Asia–Pacific*. Singapore: Institute of South East Asian Studies.

This Page Intentionally Left Blank

Chapter 41

TRANSPORT POLICY IN THE EUROPEAN UNION

PERAN VAN REEVEN
Erasmus University, Rotterdam

1. Introduction

This chapter provides an overview of transport policy in the EU. In order to facilitate an understanding of transport policy in this group of European countries, Section 2 provides a brief outline of the EU and its institutions. The main issues leading to the appearance of a common transport policy are presented in Section 3. The liberalization legislation is presented in somewhat more detail for air transport, maritime transport, inland shipping, rail transport, and road transport in Section 4. Section 5 covers the more general issues of antitrust and state aid, while the last section concentrates on the European infrastructure policy and the charging for infrastructure usage. More details on the background and politics of the European transport policy can be found in Stevens (2004).

2. European union

The Treaty of Rome established the European Economic Community (EEC), and came into force on January 1, 1958. The Treaty of Rome was concluded "for an unlimited period," which gives it an almost constitutional character. The six founding countries were Belgium, France, Germany, Italy, Luxembourg, and the Netherlands. The community has since grown larger. Denmark, Ireland, and the UK joined in 1973, Greece in 1981, Spain and Portugal in 1986, and Austria, Finland, and Sweden in 1995. In 2004 the biggest ever enlargement took place, with ten further, mostly eastern European, countries joining.

The aim of the EEC was to establish a common market, to allow the free movement of goods, and the mobility of factors of production. The mobility of factors of production involves the free movement of workers and enterprises, the freedom to provide services, and the free movement of capital. The Treaty of Rome provided for completion of a common market over a transitional period of 12 years in three stages, to be concluded by the end of 1969. Its first aim, the

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

customs union, was completed more quickly than expected: the transitional period for enlarging quotas and the phasing out of internal customs ended in the middle of 1968. By the same date, the EEC had adopted a common external tariff for trade with countries outside the community. However, at the end of the transitional period the internal market was not complete.

The Single European Act of 1986 was the first substantial change to the Treaty of Rome, and boosted an objective that was already set out in the treaty. The Single European Act stated that the internal market had to be realized by January 1, 1993. In order to make this possible, it improved the decision-making capacity of the Council of Ministers by the more frequent use of qualified majority voting rather than unanimity. Subsequently, the Maastricht Treaty of 1992 and the Amsterdam Treaty of 1997 marked significant progress toward the creation of an ever-closer union among the peoples of Europe. The Maastricht Treaty established the European Union (EU), which extended European cooperation to a common foreign and security policy, and to the fields of justice and home affairs. The treaty also led to the EEC being renamed the European Community (EC).

The EC is in a sense an evolution of the EEC. Its task is to promote, throughout the member states, a harmonious, balanced, and sustainable development of economic activities, a high level of employment and of social protection, equality between men and women, sustainable and non-inflationary growth, a high degree of competitiveness and convergence of economic performance, a high level of environmental protection, the raising of the quality of life, economic and social cohesion, and solidarity among member states. The community has pursued these objectives by establishing a common market, related measures, and an economic and monetary policy.

There are five EU institutions, each playing a specific role:

- the European Parliament (elected by the peoples of the member states);
- the Council of the EU (represents the governments of the member states);
- the European Commission (the driving force and executive body);
- the Court of Justice (ensures compliance with EU law);
- the Court of Auditors (controls sound and lawful management of the EU budget).

These are flanked by five other important bodies:

- the European Economic and Social Committee (expresses the opinions of organized civil society on economic and social issues);
- the Committee of the Regions (expresses the opinions of regional and local authorities);
- the European Central Bank (responsible for monetary policy and managing the euro);

- the European Ombudsman (deals with citizens' complaints about maladministration by any EU institution or body);
- the European Investment Bank (helps achieve EU objectives by financing investment projects).

3. Transport policy in the Treaty of Rome

Agriculture and transport are the only two economic sectors that received separate sections in the Treaty of Rome. A common policy in the sphere of transport is one of the instruments for realizing the objectives of the EC. Transport is important for the realization of a single internal market; efficient transport services significantly contribute to the free movement of goods, persons, and services. The close link between the establishment of the common market and transport was again emphasized in the amendments introduced by the Single European Act.

The special status of transport in the Treaty of Rome is in part due to the fact that the transport sector was highly regulated in all the original member states. Moreover, the national intervention systems in transport were highly divergent.

3.1. Strong regulation of the transport sector in the member states

The strong regulation of transport was caused by a number of special features of the transport sector itself (Kapteyn and Verloren van Themaat, 1998). These features are the problem of the allocation of infrastructure costs, the usage of the transport sector to realize specific political objectives located outside the transport sector, and the possible existence of destructive competition in the sector. These three features will be detailed below.

Allocation of infrastructure costs

Transport requires expensive, long-lasting infrastructure, for which government is mostly responsible. The associated allocation of infrastructure costs was a problem for the community: member states did not allocate these costs in the same way to the different modes of transport. Recovery of infrastructure costs differed widely among transport modes, so that measures were needed to restore a balanced competition. By trial and error, member states managed to create this balance between modes for their own countries. However, this pragmatic method created vast differences in subsidies and taxation between modes and countries. The lack of agreement between member states on how to allocate infrastructure costs blocked progress toward a common transport policy for years. Germany in particular had diverging views from the other member states, having a policy that favored rail transport; tariffs for road transport were fixed in such a way that rail transport faced no real competition. In addition, low rail tariffs were used as an instrument for promoting the economic development of the northern seaports in Germany.

Realization of political objectives

Over the years, the objectives of transport policy were merged with other societal objectives such as cheap transport to underdeveloped regions or for certain social groups. For example,

- Italy operated unprofitable railway lines in order to stimulate employment in less economically developed regions in the country.
- Belgium had low railway tariffs for employees, traveling from home to workplace, with the aim of equal economic development in all regions within the country.
- Germany promoted rail transport for environmental reasons, and to relieve the highway network.

Different groups of people, such as the elderly, children, and large families, received special discounts in member states for different social reasons. Many modes of transport had so-called "public service obligations." National transport policies that were focused on many different political and societal objectives resulted in as many different intervention systems. Some countries, such as Germany and France, saw transport as a public service or as an integral part of the social structure, affecting the distribution of population and shaping the community's social life; they regulated their transport sector in a very detailed way, and considered that government intervention was the only way to realize the social objectives of their transport policy. Others, such as the Netherlands and the UK, took a largely commercial view of transport; they considered that the application of market-economic principles was in the best interests of both customers and society.

Destructive competition

Transport services cannot be stocked. Consequently, capacity tends to be adapted to peaks in demand, which leads to overcapacity in off-peak periods. Because a large proportion of costs are fixed, nowadays also including wages, it was feared that this would lead to very keen price competition. Overcapacity exists especially on return journeys, particularly on return journeys from economically less developed regions. When the costs of transport are already covered on the outward journey, the return journey will yield profit even if the price is very low. Inelastic supply of transport services in periods of low demand combined with a low price–elasticity of demand, and the sensitivity of transport demand to economic trends, makes this problem even worse. This sensitivity to economic up- and down-swings led, for example, to the introduction of the *tour-de-role* system in inland shipping in the 1930s. Ships that arrived at a freight exchange received cargoes based on a "first in, first out" principle, and were paid at a fixed rate. This system survived until recently. Some modes of transport, for example road transport and inland shipping, required only small investments, with labor costs comprising a high percentage of the total costs. This allowed for a fast adaptation of supply to peak demands. These latter modes were characterized by being operated as one-man undertakings. By means of long driving hours, these small businesses were able to survive in economic down-swings by reducing costs even further at the expense of road safety. Consequently, social as well as safety considerations necessitated regulation.

3.2. The development of the common transport policy

Those features of transport discussed in the previous section necessitated a common transport policy, but their nature made it very difficult to find the required agreement between the member states. It appeared impossible to achieve an internal market through an umbrella philosophy on intervention in prices and capacity. This was primarily because the national systems were based on specific policy objectives reflecting national political considerations. Decision-making on the common transport policy was paralyzed for years due to the required unanimity for most decisions in the Council of Ministers. This was changed after the European Parliament summoned the council to appear before the Court of Justice for failure to establish a common transport policy. The court decided in 1985 in favor of the European Commission, and set with its judgment the realization of the common transport policy in motion. The realization of the common transport policy gained further momentum in 1986 with the passage of the Single European Act, which allowed more frequent use of majority voting in decision-making; the judgment of the court in the cases of *Asjes* (1986) and *Ahmed Saeed Flugreisen* (1989) that the general Treaty of Rome provisions (on competition) also apply to transport; and with publication of the European Commission's White Paper on completing the internal market.

Price and capacity rules from the national regulatory systems disappeared, and the common transport policy adopted was one with a market economy orientation, for a number of reasons. First, it appeared impossible to achieve an internal market through harmonizing national political considerations. Moreover, in the course of the 1980s, intervention in transportation was abandoned in various member states. National market control schemes appeared to increase the costs for the users of transport services. Road transport was already largely deregulated, and in air transport the effects from deregulation in the USA, which created strongly competitive airline companies, was felt. The philosophy of deregulation gradually conquered the community as well. The experience in countries that decided to deregulate transport markets was that service improved and prices dropped, without disturbing the market. In addition, the progressive development of the internal market played a role. If the objectives of the Treaty of Rome were to be realized, national government interventions had to be partly harmonized and partly abolished. Consequently, liberalization of the European transport market was considered the most plausible way forward, and in line with the EU regime for other economic sectors.

Partial harmonization was necessary for the market regime to work. For example, the qualitative (minimum) conditions for access to a profession are regulated for each sector. However, these requirements have only an indirect effect on market access. Furthermore, there are specific harmonized rules for each sector, such as driving and rest hours provisions for road transport. The condition of non-discrimination has put an end to the practice of a member state promoting export and its own nationally based operators and discouraging imports and foreign operators. Interference in prices is now restricted to only a very few precisely defined exceptional situations. In road transport a crisis mechanism is provided for emergency cases, and public service obligations may be imposed in relation to road, rail, inland waterway, and air transport.

4. The internal market for the provision of transport services

The freedom to provide services in international transport follows from the Treaty of Rome provisions on the realization of a common transport policy and, in particular, from the decision of the Court of Justice that in breach of the Treaty of Rome the Council of Ministers had failed to ensure this freedom and had to lay down the conditions under which non-resident carriers may operate transport services in a member state.

The liberalization legislation is to a large degree sector-specific, but has a similar structure for all modes. The legislation consists of rules governing:

- access to professions, to become qualified for carrying out transport services;
- access to the market, to be allowed to operate as qualified transporter-specific transport services;
- access to transport infrastructure;
- liberalization of complementary services that are necessary to operate transport services.

This structure of the liberalization legislation is presented below for air transport, maritime transport, inland shipping, rail transport, and road transport.[a]

4.1. Air transport

The air transport sector in the EU was liberalized in three successive stages. The first package of measures, adopted in 1987, started to relax the established rules. The last stage of the liberalization of air transport in the EU was the subject of the third package of measures from 1992. This package gradually introduced freedom to provide services within the EU and led in 1997 to the freedom to provide cabotage, i.e. the right for an airline of one member state to operate a route within another member state.

The market is open to all airlines that hold an EU air carrier's license. For a company to obtain this license, most of its capital must be held by member states or nationals of the EU. The latter must also exercise effective control over the company at all times. The technical capabilities and financial capacity of the companies concerned are still sanctioned by means of national certificates based on national regulations. All international air routes in the EU are open to all companies holding a EU license, without any restrictions, which means that unconditional access to all domestic markets has been granted to all airlines in the EU.

The most important safeguard measure on access to the market concerns public service obligations, which enable governments to maintain services considered essential for harmonious development within their territory. A member state may impose a public service obligation in respect of scheduled air services to an airport serving a peripheral or development region in its territory or on a minor route to any regional airport in its territory. This may be done to the extent necessary to ensure the adequate provision of scheduled air services on that route, satisfying fixed standards of continuity, regularity, capacity, and pricing – standards that air carriers would not adopt if they were solely considering their commercial interest. First, member states have to publish the public service obligations that will be imposed. Then, if no carrier is prepared to provide the service, the member state may restrict access to the route concerned to a single carrier and decide to grant that carrier financial compensation in exchange for compliance with the obligations.

There are two major restrictions to free access to the market. First, member states may regulate – without discrimination on grounds of nationality or identity of the air carrier – the distribution of traffic between airports within an airport system. Examples of airport systems are, for example, London with the airports

[a]More information on EU legislation that applies to these sectors can be found on the Internet (http://www.europa.eu.int/comm/transport/index_en.html).

Heathrow, Stansted, Gatwick, Luton, and London City, or Paris, with Orly, Charles-de-Gaulle, and Le Bourget. Secondly, when serious congestion and/or environmental problems exist, a member state may impose conditions on, limit, or refuse the exercise of traffic rights, notably when other modes of transport can provide satisfactory levels of service

Freedom with regard to fares and rates was an essential part of the freedom of access to the community market. Airlines are no longer required to submit their fares to national authorities for approval. All they have to do is to inform the relevant bodies 48 hours before applying the new fare. However, a regulation does provide for control mechanisms to be reintroduced in exceptional circumstances, although it has not been necessary so far to apply such measures. The Commission has noted, however, that the International Air Transport Association (IATA)-approved "fully flexible" fares (which are not subject to any kind of restriction as regards changing the reservation, the length of stay at the destination, etc.) are still excessively high on some routes.

The continuous growth in air transport during the last decade has increased pressure on the capacity available at airports for aircraft movements. There was therefore a need for a regulation on slots, which are defined as the scheduled times of arrival allocated at an airport. Basically, the rules maintain grandfather rights; a slot that has been operated by an air carrier entitles that air carrier to claim the same slot in the next equivalent scheduling period. At congested airports where slot allocation takes place – so-called coordinated airports – there exists a pool that contains newly created slots, unused slots, and slots that have been given up by a carrier during or by the end of the season, or which otherwise become available. Slots placed in the pools are distributed among applicant carriers. Of these slots, 50% are allocated to new entrants, unless requests by new entrants are less than 50%.

The European Commission is studying changes that should make the slot system more flexible in terms of both allocation and use, and this should strengthen the coordinator's role and the monitoring of compliance. Options for optimizing slot allocation by identifying market mechanisms that could be used for the allocation process are being analyzed. The prime objective is to encourage the mobility of slots and to make the most efficient use of scarce airport capacity. Other important objectives are to maintain effective competition at EU airports and to ensure that the slot allocation scheme fits in with the overall EU air transport policy and matches other slot allocation procedures used worldwide.

Ground handling services make an essential contribution to the efficient use of air transport infrastructure. A distinction can be made between airside and landside services, the latter being passenger-related services such as ticketing and baggage handling at the check-in desks. Airside services comprise services such as ramp handling, fueling operations, aircraft maintenance, and provision of catering.

The market in ground handling services is gradually being opened up to competition. The provision of ground handling services at EU airports was largely monopoly controlled by a limited number of service suppliers. This led to relatively high prices for the services provided and sub-optimal efficiency. Currently, access to the market by suppliers of landside ground handling services is free. For airside ground handling services at the larger airports in the EU, the number of suppliers may be no fewer than two for each category of service. Moreover, at least one of these suppliers should be entirely independent of the airport or the dominant carrier. Similar provisions exist with regard to self-handling, which means that airlines provide the services in question for themselves: for these services there should at least be two airlines admitted, and they should be chosen on the basis of objective and non-discriminatory criteria.

A number of airports used an exemption to fully open up the market and/or implement self-handling on the basis of constraints with regard to available space or capacity, arising in particular from congestion and area utilization rate. While the process of liberalizing ground handling services has not always been smooth, the positive impact can now been seen throughout the EU: the number of suppliers has increased, along with the growth in air transport in general, and the quality of services has improved, while costs for service users has, in many cases, fallen.

4.2. Maritime transport

The process of liberalization and opening up of national markets to competition is almost complete in maritime transport in the EU. Freedom to provide maritime transport services between member states and between member states and third countries applies to nationals of member states who are established in a member state and to nationals or shipping companies of member states established outside the EU in so far as their vessels are registered in a member state in accordance with its legislation. In addition, maritime cabotage, i.e. an operator's right to provide a service between two ports in a country other than the one in which it is established, has been allowed, provided that ships comply with all conditions for carrying out cabotage in that member state. This implies that ships registered in "second" registers, such as the Danish international ship register (DIS) and the Madeira register (MAR), are excluded from cabotage. Second registers are more attractive with respect to taxation, social legislation and safety or environmental standards. Ships registered in those registers operate under more favorable conditions than ships in national "first" registers and are, consequently, excluded from cabotage in order not to distort competition in national markets. By offering favorable and competitive conditions such as those from other international registries, member states with second registers aim to stop the "flagging out" process to flags of convenience as well as to attract new ship owners and vessels.

Attempts to introduce free access to the provision of port services have not been successful so far. The aim is to open up competition in port services by establishing clear rules and setting up an open and transparent procedure for access to these services. The most contested element was the allowance for "self-handling," where a port user provides for itself one or more categories of port services, for example ferry operators carrying out their own loading operations. The European Parliament considered that self-handling would degrade safety conditions and that competent workers would risk losing their jobs to cheaper and less qualified people.

4.3. Rail transport

Liberalization of rail transport is still in its infancy compared with other modes. The first piece of major legislation goes back to 1991. It required separate accounting for railway infrastructure (track and related equipment) and the operation of transport services. The aim was a greater transparency in the use of public funds, but also the ability to measure the actual performance of railway activities. It is with this requirement in mind that a number of member states set up bodies to manage the railway infrastructure separate from the railway companies that manage the carriage of passengers and freight. Member states were required to make railway undertakings independent by giving them a budget and a system of accounts separate from those of the state, and to take steps to reduce the indebtedness of railway undertakings.

The separation of railway infrastructure from the operation of transport services allowed for an open market for rail transport services. Rights of access were created for rail transport operators that operate international combined transport services. The possibility of access was also created for international freight and passenger services under certain conditions.

In 1995, the Council of the EU set common criteria for the licensing of railway undertakings established in the EU. To obtain an operating license the railway undertaking must meet a number of specific conditions (requirements in respect of good repute, financial standing, and professional competence plus civil liability). The operating license that is issued by a member state is valid for the whole EU territory.

Member states have to designate capacity allocation bodies that ensure that railway infrastructure capacity is allocated on a fair and non-discriminatory basis and that the allocation procedure allows optimum effective use of the infrastructure. The infrastructure may only be used by railway undertakings that hold an operating license. Capacity priorities may be set in the allocation of infrastructure. In order to guarantee the development of adequate transport services, in particular to comply with public service requirements or to promote the development of rail freight, member states may take any measures necessary – under non-discriminatory conditions – to ensure that such services are given priority when infrastructure capacity is allocated. Member states lay down the procedures for the allocation of railway infrastructure capacity in a network statement.

The body responsible for managing railway infrastructure also charges for the use of infrastructure. Member states must ensure that the prices charged are market prices, and must not charge fees that are unfair or discriminatory. The charges for the minimum access package and track access to service facilities have to be set at the costs that are directly incurred as a result of operating the train service. They may include a charge that reflects the scarcity of capacity of the identifiable segment of the infrastructure during periods of congestion and may also take into account of cost of the environmental effects caused by the operation of the train. Charges for the use of railway infrastructure are paid to the infrastructure manager, and used to fund its business. The accounts of the infrastructure manager have – under normal business conditions and over a reasonable time period – to at least balance income from infrastructure fees plus state contributions on the one hand and infrastructure expenditures on the other.

It should be noted that the railway rules do not cover railway undertakings whose activity is confined to urban, suburban, and regional transport or the road vehicle shuttle service through the Channel Tunnel.

The European Commission has announced its intention to amend existing legislation in order to improve access to the railway network for freight transport and to improve the interoperability of conventional rail systems and high-speed rail systems.

4.4. Inland shipping

EU individuals or undertakings wishing to pursue the occupation of carrier of goods by waterway must satisfy conditions of professional competence. Member states issue licenses on the basis of knowledge with respect to civil law and commercial, social, and tax legislation necessary for the pursuit of the occupation, the commercial and financial management of an undertaking, access to the market, technical standards and technical aspects of operation, and safety. The necessary knowledge can be acquired either by attending courses (diplomas) or by practical experience in a waterway transport undertaking, or by a combination of the two.

Member states have to recognize certificates issued by another member state as sufficient proof of professional competence. This also applies to requirements with respect to good repute, absence of bankruptcy, and financial standing as demonstrated by certificates issued by banks in other member states.

Any operator is allowed to transport goods or passengers by inland waterway in the EU provided that it is established in a member state, is licensed in that member state to carry out international transport of goods or passengers by inland waterway, and uses inland waterways vessels that are registered in a member state. The so-called *tour-de-role* system, also called "turn-by-turnabout" and "chartering by rotation," in which cargoes of customers are allocated based on the order in which vessels become available after unloading at fixed prices has been abolished. Currently, contracts can be freely concluded between the parties concerned, and the prices negotiated freely, where appropriate within charter clearing houses.

4.5. Road transport

Undertakings wishing to engage in the occupation of international road transport operator have to be of good repute, of appropriate financial standing, and have professional competence. Member states have to recognize for the purpose of admission to the occupation of road transport operator extracts or certificates that are legitimately issued in other member states. International carriage is restricted to EU authorization that is issued by a member state to any operator carrying goods by road for hire or reward who is established in a member state and who fulfils the requirements concerning admission to the occupation of road haulage operator. Such operators are also entitled to operate national road haulage services for hire and reward in another member state, i.e. cabotage, without having a registered office or other establishment therein.

The European Commission has proposed strategies to strengthen the conditions of fair competition that minimize the adverse impacts of increasing competition due to the advent of the internal market. Ever-increasing safety requirements mean that measures must be taken to protect the safety of workers and road users. In addition, in order to deal with the social disquiet caused by the use of illegally employed drivers, working conditions must be improved, monitoring strengthened, and the image of the profession enhanced. The measures proposed relate to the working time, conditions of employment, and professional training of drivers, enforced through improved monitoring of road transport.

5. Antitrust and state aid in transport

The general Treaty of Rome provisions on antitrust and state aid also apply to transport, but there are a number of exemptions.

5.1. Antitrust

The Treaty of Rome prohibits all agreements between undertakings, decisions by associations of undertakings, and concerted practices that may affect trade between member states and which have as their object or effect the prevention, restriction, or distortion of competition within the common market. Also, any abuse by one or more undertakings of a dominant position is prohibited as being incompatible with the common market in so far as it may affect trade between member states.

In air transport, there are a limited number of restricted exemptions. The general prohibition to agreements between undertakings does not apply to joint planning and coordination of the schedule of an air service between EU airports, especially in so far as it concerns low-density routes. In a similar vein, the joint operation of a scheduled air service on a new or on a low-density route between EU airports is exempted. More surprisingly, consultations on tariffs for scheduled air services between EU airports are also exempted. However, these consultations only concern the fully flexible tickets that are discussed in the context of IATA and that allow the passenger to use at any time flexibility in terms of departure time or airline. Also, slot allocation and airport scheduling negotiations are exempted in order to create workable sets of departure and arrival times for airlines within the EU. Finally, the Treaty of Rome does not apply to agreements between undertakings for the purpose of purchasing, developing, or managing a computer reservation system in common. This last exemption is conditional upon the possibility that any air carrier is allowed to participate on an equal and non-discriminatory basis.

In maritime transport, there exists an exemption with regard to cooperation between liner shipping companies in so far governed by the United Nations Code of Conduct of Liner Conferences. Without the code, cooperation is only allowed if all EU nationals have fair, free, and non-discriminatory access to the cargo shares of the member states concerned. This cooperation can concern the coordination of shipping timetables, sailing dates or dates of calls, the determination of the frequency of sailings or calls, the coordination or allocation of sailings or calls among members of the conference, the regulation of the carrying capacity offered by each member, or the allocation of cargo or revenue among members. Liner conferences are considered to have a stabilizing effect that assures shippers and users of adequate efficient services. The mobility of fleets, a characteristic of shipping, subjects conferences to constant competition. The prohibition does not apply to agreements between liner shipping companies for the purpose of rationalization of their operations by means of technical, operational and/or commercial arrangements, with the exception of price fixing (consortia). Liner shipping is a capital-intensive industry, and containerization has increased pressures for cooperation and rationalization. The main reason for this exemption is that the EU shipping industry needs to reach the necessary economies of scale in order to compete successfully on the world liner shipping market

5.2. State aid

The Treaty of Rome also states that any aid granted by a member state or through state resources in any form whatsoever, and which distorts or threatens to distort competition by favoring certain undertakings or the production of certain goods, is, in so far as it affects trade between member states, incompatible with the common market. The principle used to assess the existence of state aid is the "market economy investor principle," also called the "prudent private investor principle"; public aid is considered state aid if no private investor would have invested the same amount considering the market prospects, the position of the company, and any measures taken to improve it.

There are a number of exemptions to the prohibition of state aid. One relates to services of general interest that increase the quality of life for all citizens and that overcome social exclusion and isolation. A minimum guaranteed level of mobility is such a service of general interest. In the case of urban, suburban, and regional services (road, inland waterway, and rail), it is possible to impose and financially compensate for public service obligations. Public service obligations means obligations that the transport undertaking in question, if it were considering its own commercial interests, would not undertake these obligations, or would not undertake them to the same extent or under the same conditions. Public service obligations consist of the obligation to operate, the obligation to carry, and tariff obligations. The obligation to carry means the provision of a transport service that satisfies fixed standards of continuity, regularity, and capacity. The obligation to carry implies the obligation to accept and carry passengers or goods at specified rates and subject to specified conditions. Finally, tariff obligations are obligations to apply for certain categories of goods or passengers, or on certain routes, with rates fixed or approved by a public authority. The policy is to move away from public service obligation and use competitive tendering for such services instead.

In maritime shipping, the opening up of the market has given the consumer a wide choice of competitive shipping services, but the proportion of ships entered in the member states' registers and the number of EU seafarers employed have both declined significantly, especially over the last decade. The European shipping industry faces stiff international competition, and the size of the EU-registered fleet in total worldwide maritime transport has been decreasing steadily over the last three decades. There has also been a correspondingly steady decrease in the number of EC seafarers employed on board ships. The problem is that the registration of a vessel in a member state offers some disadvantages, such as rather strict manning conditions and less favorable fiscal and social arrangements for companies and their employees. This means that in most cases it is relatively expensive to operate EU registered ships with EU seafarers on board. Since the maritime shipping industry is rather footloose, many shipping companies have moved to flags of convenience, countries that are far more attractive to ship owners than Europe in terms of taxation, social legislation, and safety or environmental standards.

In order to counter this tendency, member states are allowed to give financial support to the maritime sector that reduces fiscal and other costs and burdens borne by EU ship owners and EU seafarers. This implies reduction of taxation and social security contributions to world norms, conditional on the requirements that such reductions directly stimulate the development of the sector and employment, rather than provide general financial assistance. Many member states have already taken measures to improve the fiscal climate for ship-owning companies, including, for instance, accelerated depreciation on investments in ships or the right to reserve profits made on the sale of ships for a number of years on a tax-free basis, provided that these profits are reinvested in ships. Certain member states and non-EU countries have also replaced the normal corporate tax system by a tonnage tax. Tonnage tax means that the ship owner pays an amount of tax linked directly to the tonnage operated. The tonnage tax will be payable irrespective of the company's actual earnings, or profits or losses made.

In air transport, the crisis in the early 1990s forced many European airlines to restructure. The dramatic decline in demand suddenly exposed their handicaps: overcapacity, low productivity, high costs, under-capitalization, etc. To make restructuring a success, it was necessary, among other things, to grant public funds. The European Commission mapped out a strict approach to make sure that this aid genuinely met a need for restructuring, without distorting competition on the market in the process. A major objective was to prepare airlines, with inherited liabilities and burdens, for liberalization and to make a fresh start. Since 1991, seven airlines, mostly southern European, have received public funding for restructuring: Sabena, Iberia, Aer Lingus, TAP, Air France, Olympic Airways, and Alitalia. The granting of aid was an exceptional measure to support the restructuring of the airlines in preparation for liberalization of the European market. The airlines that needed such restructuring have now completed the process. The European Commission therefore considers that, in principle, there is no longer a reason for state aid in air transport.

6. Infrastructure policy and charging

6.1. Infrastructure policy

In 1996, the European Parliament and Council took a decision toward the realization of a trans-European transport network (TEN-T). Infrastructure projects to complete the TEN-T scheme comprise roads, railways, inland waterways, airports, seaports, inland ports, and traffic management systems that serve the entire continent, carry the bulk of long-distance traffic and bring the geographical and economic areas of the EU closer together. The development of trans-European networks is a key element for the creation of the internal market and the reinforcement of economic and social cohesion. This development includes the interconnection and interoperability of national networks as well as access to such networks. The time horizon to complete the network is 2010.

TEN-T consists of 14 priority projects. Three projects have been completed and are already operational: the Cork–Dublin–Belfast–Larne–Stranraer conventional rail link in Ireland, the Malpensa Airport in Milan, Italy, and the fixed rail/road link between Denmark and Sweden (the Øresund bridge).

Eleven of these projects are still in the construction stage, requiring financial support:

- High-speed railway projects will receive the biggest support. Their completion is expected by the end of 2006: the Paris–Brussels–Cologne/Frankfurt–Amsterdam–London project, the "TGV East" linking France and Germany (with a branch line to Luxembourg), sections of the north–south axis in Germany and Austria, the France–Italy link, the southern high-speed rail link "TGV South" (France, Spain), the Betuwe railway freight line in the Netherlands, sections of the Nordic Triangle in Finland and Sweden, and the west coast main line in the UK.
- Three trans-border railway projects through mountains: TransAlpine tunnels under the Brenner and Mont Cenis passes in Austria and Italy, and, as part of the southern high-speed railway project, the TransPyrenees project linking Figueras (Spain) and Perpignan (France).
- Studies for the southern high-speed rail link in Spain and Portugal, road construction in the Nordic Triangle, and the Ireland-UK-Benelux road link, and final technical studies on the Egnatia motorway in Greece.

In 2001 six new projects were added. The selected projects concern the global navigation and positioning satellite system Galileo, a high-capacity rail link across the Pyrenees between France and Spain, an eastern European combined transport/high-speed train, the Danube river improvement between Vilshofen and Straubing in Germany, high-speed rail interoperability in Spain and Portugal, and a fixed link (bridge) between Germany and Denmark over the Fehmarn belt.

The financial resources needed to complete the network in 2010 are estimated to be €400 billion. A number of financial instruments have been set up at the EU level, each with its own legal basis, in order to conduct the development of TEN-T and to support member states financially in specific cases. The EU will fund not more than 50% of the cost of preliminary studies (feasibility studies), and 10–20% of the cost of the work. At least 55% of funds for TEN-T will be given to railway projects and not more than 25% to roads. The remaining resources will come out of public or private funds.

New outline plans for 2020 will be drawn up with the aim of efficiently channeling the trans-European traffic of tomorrow in an enlarged EU. This will include sections of pan-European corridors situated on the territory of the ten new EU countries and candidate countries, including those that will still not be members of the EU at that time.

The priority projects are part of a much larger program for infrastructural improvements. A particular focus of attention has been the freight sector, where incoherent infrastructure standards and congested nodes are still hampering traffic flows on major trans-European routes. Another priority has been the promotion of intermodal links, such as the connection of airports to the railway network, to facilitate efficient long-distance passenger traffic. The selected measures, which cover projects such as infrastructure works, installation of track equipment, upgrading of signaling systems, improvements in nodes, or action to ensure interoperability, aim at achieving measurable benefits within a limited number of years so as to make a significant contribution toward more efficient and attractive railway services in the short and medium term.

However, there are some criticisms with regard to the selection of priority projects (Sichelschmidt, 1999). Member states' governments may have submitted projects for the TEN-T program not primarily because of their importance in an EU-wide context but simply because they wished to attract from the common treasury as many financial resources as possible. The problems of biased incentives are likely to be exacerbated by the redistributive objective of the TEN-T program to improve connections to insular and peripheral regions. For example, the 1994 outline plans contained many TEN-T projects, e.g. in Central France or in the Iberian Peninsula, which may improve the accessibility of these regions but can hardly be regarded as international public goods in the form of transit corridors or important cross-border connections. It has also been important to have all member states participating, this being an important selling point for the acceptance of the TEN-T program.

The major part of the TEN-T program – high-speed train (HST) routes – is most likely to benefit mainly the metropolitan areas to be served by HSTs. Rural areas that will just be passed through are likely to benefit much less. The effects of HST routes on regional development will presumably, like those of airports, concentrate around stations, thus differing markedly from the effects of conventional railways. It might well be that such projects support centralization rather than decentralization because, as Vickerman (1996) has stated, "improvements directed at specific, e.g. peripheral or lagging regions, also improve the access to markets in these regions of firms in already more accessible, central regions," and decreases the importance to maintain branch factories or offices in peripheral areas.

The main problem as concluded by Sichelschmidt is that the TEN-T program, in its long history since its origins and approval by the Council of Ministers and the European Parliament, has been overloaded with primarily non-economic goals, especially distributional or environmental ones, which come at the expense of efficiency. The program is little more than a continuation of existing national policies. The TEN-T scheme does not provide for a more rational planning of infrastructure, although the European Commission has commissioned research studies on the economic appraisal of TEN-T projects (e.g. Van Exel et al., 2002).

6.2. *Charging for infrastructure use*

A main principle of the EU's infrastructure charging policy is that transport taxes and charges, in every mode of transport, should be varied to reflect the cost of different pollution levels, traveling times, and damage costs as well as infrastructure costs. The application of the "polluter pays" principle and clear fiscal incentives help to achieve the goals with regard to reducing transport congestion, pollution, rebalancing the modal split, and decoupling transport growth from economic growth. Appropriate charging helps make better use of the existing infrastructure capacity.

As a start, in 2003, national systems of tolls and road use charges for heavy goods vehicles were aligned and based on common principles. This should prevent isolated initiatives on the part of member states that exacerbate the fragmentation of transport taxes and charges in the EU with possible unequal treatment of operators on the various road networks. The revenues from infrastructure charges are used for the benefit of the transport sector. In certain cases, there is scope for cross-financing of infrastructure that provides an alternative to road transport (e.g. rail).

A study by Proost et al. (2002) shows that peak private passenger and peak private freight transport have marginal external costs in the EU that are considerably larger than the current tax levels, and that parking was an important unpaid resource cost in urban areas. In the off-peak period, taxes on private vehicles are sometimes too high and sometimes too low. Efficient prices can be computed such that the price level corresponds to the marginal social cost at the most efficient level of traffic. This would involve an increase in the monetary price of peak private transport between 35 and 233%, which would lead to decreases in transport volumes between 10 and 33%.

7. The way forward

The European transport policy has largely come about by the liberalization of transport markets. This liberalization is to a large degree complete. Rail passenger transport and port services in maritime transport are two exceptions that the EU is still working on. In general, the focus in recent years has shifted from liberalization toward the conditions for letting the internal market for transport function. Competition has to create efficient and effective transport services. First, this requires the interoperability of infrastructures (rail gauges, safety systems, etc.) and the completion of missing infrastructure (e.g. TEN-T). Second, equal conditions for all operators need to be established. One of the instruments for realizing this equality is fair and efficient pricing and taxation based on the user pays principle. A second instrument is the harmonization of working and operating conditions. These conditions are harmonized through legislation that covers areas such as driving times, professional qualifications, vocational training, enforcement measures, and technical requirements. Third, measures have to be introduced to increase the available capacity of existing infrastructure. One example is to take away national borders in air traffic control in order to create a single European sky. This would allow for a more efficient use of air space.

It is also increasingly recognized that increased competition caused by liberalization may also jeopardize the interests of society as a whole. This concerns the environment in general, and more specifically safety. Environmental protection and safety measures are predominantly cost drivers for operators that in principle do not generate revenue. Hence, operators may be tempted to economize wherever they can. Stead (2001) argues that the current rate of increase in transport volumes outstrips the rate of improvement in environmental technology for transport, resulting in increasing environmental problems in the transport sector. Consequently, there is an increasingly strong environmental argument to decrease transport intensity in order to reduce pollution, resource use, and waste. The success of road and air transport has created a growing imbalance between modes of transport, with associated problems of congestion. On the other hand, the potential of rail and waterborne transport has been barely exploited. The European Commission (2001) wants to curb this trend through different economic measures. A vast and growing body of regulations also involve technical requirements to safeguard the environment and safety. A straightforward example concerns restrictions on vehicle emissions. In a similar way, consumer protection has increasingly become a priority. Some service characteristics are not visible to consumers, with the risk that their interests are not sufficiently safeguarded. To counter this, the EU establishes consumer rights (e.g. compensation when denied boarding an aircraft) and informs consumers of these rights (e.g. by means of posters at airports).

In 2001 the European Commission published its transport policy for a 10 year period in its White Paper *European Transport Policy for 2010: Time to Decide.* This policy paper provides a detailed guide to the ambitions of the EU with regard to transportation.

References

European Commission (2001) *European transport policy for 2010: time to decide*. Luxembourg: Office for Official Publications of the European Communities.

Kapteyn, P.J.G. and P. Verloren van Themaat (1998) *Introduction to the law of the European Communities: from Maastricht to Amsterdam*. London: Kluwer.

Proost, S., K. Van Dender, C. Courcelle, B. De Borger, J. Peirson, D. Sharp, R. Vickerman, E. Gibbons, M. O'Mahony, Q. Heaney, J. Van den Bergh and E. Verhoef (2002) "How large is the gap between present and efficient transport prices in Europe?" *Transport Policy*, 9:41–57.

Sichelschmidt, H. (1999) "The EU programme "trans-European networks" – a critical assessment," *Transport Policy*, 6:169–181.

Stead, D. (2001) "Transport intensity in Europe – indicators and trends," *Transport Policy*, 8:29–46.

Stevens, H. (2004) *Transport policy in the European Union*. Houndsmills: Palgrave Macmillan.

Van Exel, J., S. Rienstra, M. Gommers, A. Pearman and D. Tsmaboulas (2002) "EU involvement in TEN development: network effects and European value added," *Transport Policy*, 9:299–311.

Vickerman, R.W. (1996) "Restructuring of transport networks," *European Journal of Regional Development*, 3:16–26.

Chapter 42

TRANSPORT POLICY IN POST-COMMUNIST EUROPE

JOHN PUCHER and RALPH BUEHLER
Rutgers University, New Brunswick, NJ

1. Introduction

The formerly socialist countries of central and eastern Europe have experienced profound political and economic changes since the demise of Communism in the late 1980s and early 1990s. Each country has its own particular history of transformation to a freer, more democratic, more market-based society. The timing and specific circumstances of the revolutions in each country vary. Even today, there are considerable differences among countries in the extent to which their political systems are fully democratic and how market based their economies are. Thus, it is somewhat risky to generalize about this group of diverse countries.

Without exception, however, every formerly socialist country in central and eastern Europe has at least moved toward greater democracy and greater market orientation. In every country, that political economic shift has produced a corresponding transport revolution. The most obvious indicator of that revolution is the dramatic growth in levels of private car ownership and use, and a corresponding decline in public transport use. The modal shift in passenger transport is mirrored in most countries by similar changes in goods transport, with substantial shifts from publicly owned and operated rail transport to privately owned and operated trucking firms. While the increasing reliance on roadway transport had already started during the later years of the socialist era, the movement toward market-based capitalism greatly accelerated it, prompted by striking changes in government transport policies. Indeed, a key thesis of this overview is that policy changes were responsible for virtually all of the enormous changes observed in central and eastern Europe from 1988 through the 1990s, demonstrating how crucially policies affect every aspect of our transport systems.

This review focuses on three central European countries for detailed analysis: the Czech Republic, Hungary, and Poland. We also include the former East Germany, whose political, economic, social, and transport systems dramatically changed after German reunification in 1990. Those four formerly socialist countries have the most reliable long-term series of transport statistics, enabling better

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

Table 1
Auto ownership trends in formerly socialist countries of central Europe (cars per 1000 inhabitants)

Country	Year					GNI per capita in 2001 (US $)
	1976	1980	1990	1996	2001	
Belarus	12	31	59	101	142	7 630
Bulgaria	51	56	146	204	262	6 740
Croatia	80	108	147	175	257	8 930
Czech Republic	112	148	228	325	369	14 320
East Germany	122	151	296	440	486	17 668
Estonia	12	31	154	277	347	9 650
Hungary	69	85	188	239	244	11 990
Latvia	12	31	106	153	250	7 760
Lithuania	12	31	133	212	340	8 350
Macedonia	80	108	121	139	151	6 040
Poland	37	67	138	209	272	9 370
Romania	9	11	56	107	144	5 780
Russian Federation	12	31	65	92	140	6 880
Slovak Republic	112	148	163	198	240	11 780
Slovenia	80	108	289	365	433	17 060
Ukraine	12	31	63	93	106	4 270

Note: GNI is gross national income. GNI per capita, as reported by the World Bank (2003), reflects the purchasing power parity of incomes in each country instead of simply exchange rates, which can be misleading and understate real incomes in poorer countries.
Sources: The World Bank (1999, 1998, 2003); American Automobile Manufacturers Association (1978, 1982); German Federal Statistical Office (2003); Institute of Transport Engineering (2003).

analysis of their transport systems, travel behavior, and policies. Moreover, they are typical of developments in other central and eastern European countries as well, with most transport trends being in the same direction even if the magnitudes vary from one country to another. This overview is limited mainly to urban passenger transport, although developments in long-distance passenger travel and goods transport will be briefly noted as well.

2. Trends in transport systems and travel

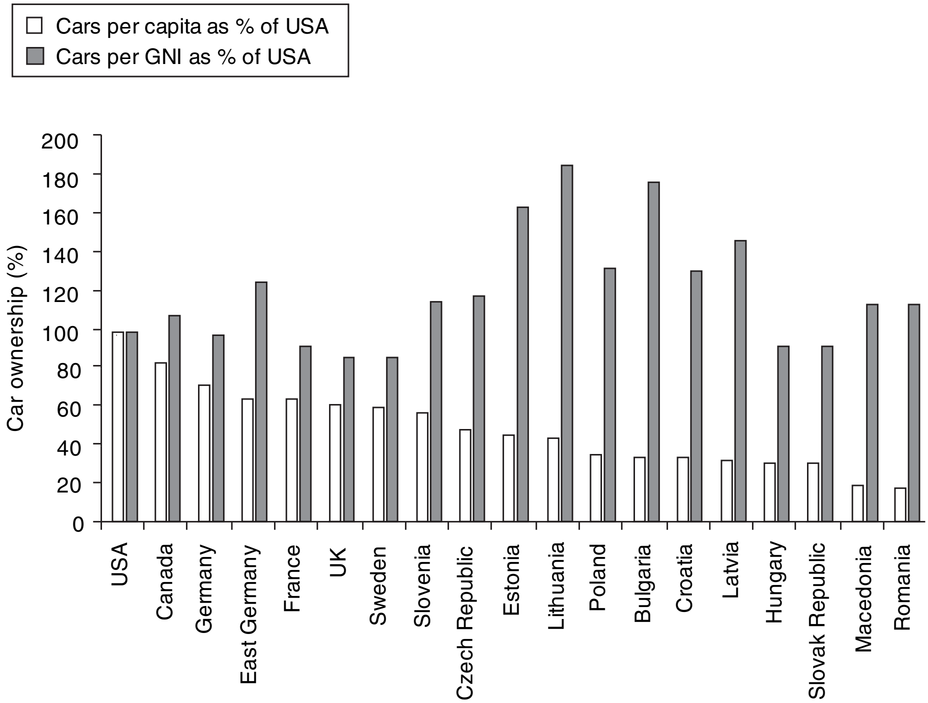
One of the most useful indicators of overall transport orientation in a country is the level of motorization, usually measured by the number of private vehicles per 1000 population. As shown in Table 1, private car ownership has increased rapidly over the entire period from 1976 to 2001, but the largest jump came from 1990 to 2001 during the first full decade after the overthrow of Communism. Most of the countries roughly doubled their levels of car ownership per capita in only a decade. Moreover, the table hides the especially rapid growth in car ownership in some countries in the late 1980s. As part of their liberalization attempts, governments in Hungary and Poland had already begun expanding offerings of consumer goods such as cars well before socialism's complete overthrow. Likewise, car purchases in the former East Germany and Czechoslovakia boomed almost immediately after their peaceful revolutions in 1989.

As one would expect, those countries with higher per capita incomes in Table 1 also have higher levels of car ownership. Thus, East Germany and Slovenia, which have by far the highest per capita incomes among these formerly socialist countries, also have the highest levels of motorization. Countries such as Ukraine, Russia, Macedonia, and Romania have the lowest per capita incomes as well as the lowest levels of car ownership. There is much variation, however, and data irregularities might help explain some anomalies such as Bulgaria, which reports a much higher level of motorization than would be expected from its very low per capita income.

Figure 1 provides two alternative measures of car ownership to compare levels of car ownership in various countries of Europe and North America. The darkly shaded bars represent the traditional statistic of cars per capita, here expressed relative to the USA, which has the highest rate (748 cars per 1000 persons). The lightly shaded bars show levels of car ownership per unit of real income or purchasing power, using the gross national income per capita of each country (purchasing power parity), as reported by the World Bank (2003), but also relative to the USA. While Figure 1 shows the expected decline in car ownership per capita from West to East, formerly socialist countries have much higher levels of car ownership than one would predict on the basis of per capita income.

There are several possible reasons for the weaker than expected relationship between car ownership and income. First, the statistic only considers the quantity and not the quality of cars. Thus, passenger cars in formerly socialist countries generally are older, often purchased as used cars, and of considerably lower quality than those in western Europe and North America. Second, it has been argued that the private car is an extremely important symbol of economic and social status as well as personal freedom. Many residents of formerly socialist countries have purchased cars mainly to possess this crucial symbol of success and independence, often far beyond their economic means or actual transport needs (Pucher, 1995, 1999; Suchorzewski, 2002; Komornicki, 2003). Third, as car ownership reaches high levels, as in the USA, there is a saturation effect, so that higher incomes result mainly in more expensive cars instead of more cars. Conversely, at the lower end of the spectrum in eastern Europe, increases are mainly in quantity, since overall motorization rates are still comparatively low. Finally, cars have become increasingly necessary in some eastern European countries, especially in smaller cities and rural areas, as public transport systems have sharply deteriorated and fares skyrocketed.

In addition, however, data problems surely underlie some of the strange patterns seen in Figure 1. Data on car ownership can be notoriously unreliable, based on



Both motorization statistics are expressed here relative to the USA (index = 100). Cars per 1000 persons is the statistic most frequently used to compare car ownership among countries. Cars per GNI is a useful supplemental measure, however, since it reflects the number of cars per capita relative to average incomes in each country, using estimates for each country by the World Bank (2003).

Figure 1. Passenger car ownership in central and eastern Europe compared with western Europe and North America, 2001.

different collection techniques and statistical definitions. In many countries, for example, discarded vehicles are not properly de-registered and removed from the vehicle stock reported. National income statistics in some countries may be inaccurate as well, since black markets and hidden economies play a significant role in eastern Europe, and are not reported in official statistics, thus understating real incomes and purchasing power.

Figure 2 reveals in greater detail the early and particularly rapid growth in car ownership in East Germany, with a 65% increase in only 4 years (from 1988 to 1992). In the same four years, car ownership rose by 42% in Poland, 33% in Hungary, but only 16% in the Czech Republic. Over the much longer period from 1992 to 2001, car ownership continued to grow, but at lower annual rates. Indeed, the 31% further increase in East Germany over those 9 years was less than half as large as the percentage increase from 1988 to 1992. Likewise, the increase in

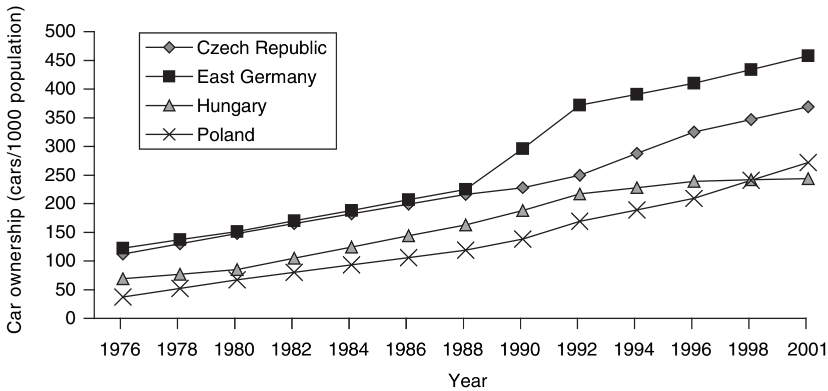


Figure 2. Trends in auto ownership for the Czech Republic, Hungary, and Poland (cars per 1000 population).

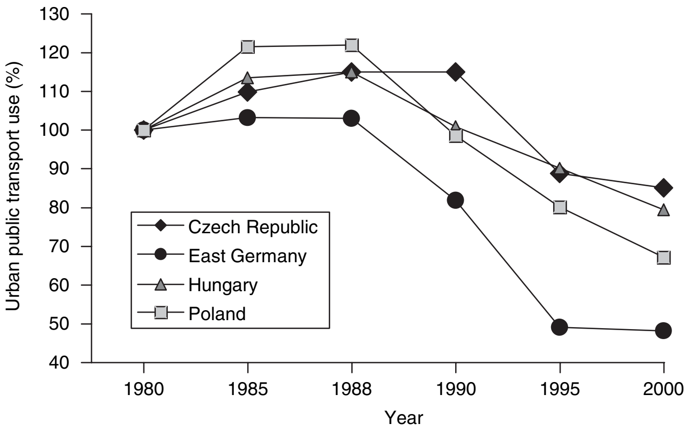


Figure 3. Trends in urban public transport use, 1980–2000 (annual trips relative to the base year 1980).

Hungary was much smaller (12 versus 33%). In sharp contrast, growth accelerated in the Czech Republic and Poland, with larger increases in the second period. As noted later, differences in the timing of policy changes explain much of this variation among these countries in the timing of motorization growth.

While car ownership and use were increasing, public transport use plummeted in the late 1980s and early 1990s. Mirroring the sudden and dramatic jump in car ownership in East Germany, public transport systems there lost almost half their riders in only 3 years after the fall of socialism (from 1988 to 1991). As seen in Figure 3, urban public transport use fell in all four of these countries, albeit with

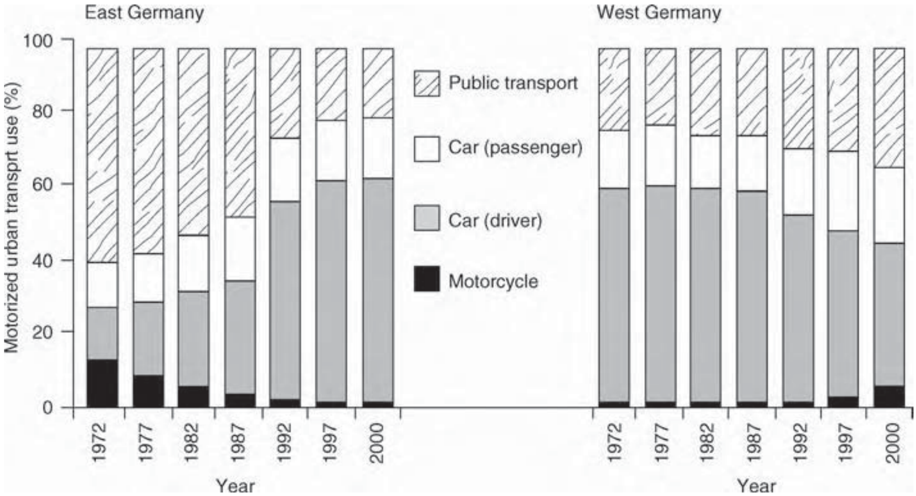


Figure 4. Convergence of travel behavior in East and West Germany, 1972–2000 (percentage of motorized urban trips by means of transport).

different timing. For example, it declined later and less in the Czech Republic than in Hungary and Poland.

The situation has stabilized in recent years, with much slower declines, but the overall loss of passengers from 1988 to 2000 was stunning. Just as urban public transport use declined, long-distance rail travel declined as well – by 36% in the Czech Republic, 26% in Hungary, and 54% in Poland. Intercity and rural bus services suffered even larger losses of passengers – 68% in the Czech Republic and 58% in Poland (Hungarian Central Statistical Office, 2002; Czech Statistical Office, 2003; Polish Central Statistical Office, 2003).

The obvious result of the rise in car use and fall in public transport use has been a dramatic change in modal shares of travel. From the mid-1980s to 2000, public transport's share of total motorized trips fell from about 75–85% to only 50–60% in large Polish, Hungarian, and Czech cities. Public transport has lost even more market share in small cities and villages, many of which now have virtually no public transport at all (Pucher, 1998; Suchorzewski, 1999, 2002; Institute of Transportation Engineering, 2003).

Perhaps most striking is the complete transformation of travel behavior in the former East Germany. As shown in Figure 4, the distribution of urban trips by means of transport in East and West Germany rapidly converged after 1987, so that the modal trip distribution is now almost identical. Indeed, public transport now accounts for a slightly higher percentage of total motorized trips in the West than in the East (24 versus 20%) while the reverse was true in 1987 (25 versus 49%) (Broeg and Erl, 2003).

Just as the modal distribution of urban passenger travel shifted toward the private car and away from public transport, the railroad's share of freight transport has fallen sharply, while the transport of goods by roadway has risen. For example, the rail share of total freight (in tonne-kilometers) fell in the Czech Republic from 73% in 1990 to 25% in 2002. In Poland, rail's share of freight transport in tonne-kilometers fell from 67% in 1990 to 39% in 2002. Almost all of the freight traffic lost by rail has been shifted to lorries on roadways. Over the same 12 year period, the total freight carried by roadway tripled in the Czech Republic (from 14 951 million to 45 059 million t-km) and almost doubled in Poland (from 40 293 million to 74 403 million t-km) (Czech Statistical Office, 2003; Polish Central Statistical Office, 2003). Combined with the skyrocketing use of private cars, that rapid increase in lorry traffic has put an enormous strain on the limited capacity of roadway networks in central Europe.

3. Shifts in land use patterns

Corresponding to these dramatic changes in travel behavior, land use patterns have also changed. Socialist cities in central Europe were densely settled around public transport routes. Low-density suburban sprawl was virtually non-existent prior to 1990. Almost all new housing was government-built and government-owned, and it was concentrated in very high-density and appallingly ugly apartment complexes on the periphery of cities, because that was the only land available for such vast projects. Even in these peripheral settlements, there was little need for a car because they were well served by frequent public transport services to the center (Pucher, 1990).

The situation has changed considerably since 1990. Similar to the long-term trends toward decentralization in North America and western Europe, most new construction has been in the suburbs (Pucher, 1999; Suchorewski, 1999, 2002; Sykorova and Sykora, 1999; Sykora, 2002). In order to avoid the congestion and high land prices in large central cities, many firms are now locating on the periphery along key highways. Shopping centers are also emerging far from the center. Already by 2000 there were over 25 shopping centers and megastore complexes in the Warsaw suburbs (Suchorzewski, 2000). Most new housing is also being built at the urban fringe, but unlike the high-density apartment complexes of the Communist era, most housing units are now low-density single-family homes. Public transport services are sparse in these new suburban developments. Especially with the surge in auto ownership and use, suburbs are becoming entirely auto-oriented in their design and travel patterns.

Suburban sprawl is especially pronounced and problematic outside the confining political borders of many large cities. While some central cities themselves have retained strict land use regulations and building codes on their territory, much new suburban and exurban development is in communities beyond their control. Land use regulations there are far more lax than in central cities, and suburban towns are tempted to permit virtually any kind of development in order to attract jobs, tax revenues, and economic development away from the center. By comparison, there is less auto-oriented suburban sprawl in smaller cities, with their lower growth rates, lower incomes, and lower car ownership rates. Nevertheless, the trend toward dispersal is perceptible to some extent in almost every city in formerly socialist countries.

While the very strict land use controls and high-density housing policies under Communism strongly encouraged public transport use, the recent trend toward low-density commercial and residential development at the suburban periphery obviously reinforces the trend toward more auto ownership and use. After decades of being forced to use crowded public transport and to live in monolithic, unattractive, and boring apartment complexes, the shift to the car and the flight to low-density suburbs is not surprising. In addition, the growing cadre of middle- and upper-class professionals and entrepreneurs are obvious customers both for new cars and for single-family homes in the suburbs. Firms are locating in the suburbs for the same reasons they do in North America and western Europe: convenience, lower cost, less regulation, greater land availability, less congestion, cleaner air, and access to the long-distance highway network. Now that firm location decisions are based mainly on profit incentives, the move is definitely to the suburbs, except for those specialized firms and headquarters functions that still need access to the core.

4. Transport policies in the socialist era

It is not difficult to find the causes of public transport's dominance in socialist countries. Partly as a matter of socialist ideology, Communist dictatorships ensured that private car ownership and use would be extremely expensive and difficult, while public transport was widely available and subsidized to such an extent that it was almost free. Socialist governments set the costs of car ownership and operation very high through their system of regulated prices. In addition, they sharply restricted their own car production while prohibiting imports of Western cars, thus keeping the supply limited. In Poland, moreover, gasoline was rationed from 1981 to 1988, leading to a black market in ration coupons that further increased the price of gasoline for anyone wanting to drive more than was possible with the official allotment of 24–45 liters per month. There were long waiting times for purchasing new cars – over a decade in East Germany and Poland. The quality of cars was abysmally low. They often broke down, and it was difficult to obtain spare parts for repairs. Finally, the roadway network was primitive by

Western standards, and there was a severe shortage of gasoline stations, repair shops, and other service facilities for cars (Pucher, 1993, 1994, 1995).

Until about 1970, Communist governments treated the private car as a luxury and a symbol of capitalism, materialism, and consumerism inimical to the very principles of socialism. During the 1970s, however, restrictions on car ownership had to be relaxed in response to growing popular demand for cars and other consumer goods. Most eastern Europeans perceived the private car as a higher-quality mode of transport, and its limited availability made the car an important status symbol. Communist governments throughout eastern Europe rationed the small supply of cars as rewards to the Party faithful. In some countries, the Communist Party used car sales to lure hard Western currency holdings from the general population, which otherwise had to wait over a decade to purchase a car. Even in the last days of socialism, however, car ownership was limited to a small minority, and increases in auto ownership were only grudgingly permitted.

Public transport, by comparison, was seen as being most consistent with aplanned economy, with its limits on consumption, mobility, and locational choice. To some extent, restricted automobility probably helped Communist dictatorships keep their populations under control. Public transport users were literally "captive riders." In contrast to the private car, public transport was treated as a basic necessity of life, to be provided to all at a negligible charge. Central governments in socialist countries provided generous financing for all public transport investments and operations. Yet, with the exceptions of metro systems in large cities such as Prague, Budapest, Moscow, and Leningrad, public transport services in socialist countries generally had much lower quality than public transport in western Europe. In almost all socialist cities, buses, trams (streetcars), and trolley buses were often overcrowded, slow, poorly coordinated, and subject to frequent breakdowns (Pucher, 1990). Since they had no competition, public transport systems were hardly concerned about rider comfort or convenience. As in so many sectors of socialist economies, overstaffing, incompetence, lack of worker motivation, excessive bureaucracy, and extreme inefficiency also characterized public transport.

Nevertheless, public transport services were extensive, frequent, and cheap. Low fares were an especially strong inducement to public transport use because of the low per capita incomes in most socialist countries. That was somewhat less true for East Germany, Hungary, and Czechoslovakia, which had the highest incomes of any of the world's socialist countries, but still much lower than western European countries. Most people simply could not afford to own cars, let alone use them for daily travel. Even as car ownership rose during the 1970s and 1980s, most cars were used for trips to the countryside on weekends and holidays or to garden plots on the outskirts of the city. As late as 1988, for example, only 10% of Czechs used a car for the journey to work (Institute of Transportation Engineering, 1992).

5. Transport policies transformed by fall of Communism

With the overthrow of Communist governments in central and eastern Europe from about 1989 to 1992, most of these transport policies changed. One important change was a sharp reduction in central government subsidy to public transport. Most of the burden of financing capital investment and operating subsidy was quickly shifted to municipal governments. Cities now pay the entire operating subsidy for public transport (except for some short-distance railroad services).

The situation for capital subsidies is more complicated, and has changed over time. In recognition of a desperate need to renew aging rolling stock and improve deteriorated rights of way, some central governments have established special infrastructure funds with varying degrees of modest assistance. In the Czech Republic, for example, the central government offered to cover 30% of vehicle and infrastructure costs for electric trams and trolley buses, and 10% of bus purchase and rehabilitation costs. As in many countries, however, local Czech government was not able to raise the necessary matching funds, and the central government could not afford to offer the promised contribution. Central government subsidy programs in most countries have been completely eliminated, and those remaining are often revised, subject to the vagaries of annual budgets. Metro systems in the large capital cities – Prague, Warsaw, and Budapest – receive some central government subsidies for extensions and modernization, but those special programs have varied from year to year according to annual parliamentary budget agreements. In general, the overall funding contribution of central governments is small and focused on rail projects.

The consequences of this funding cutback have been devastating for public transport, especially since local governments are in terrible financial straits and cannot offset the reduction in central government subsidies. With sharp reductions in subsidy, public transport systems were forced to raise fares drastically, both in absolute terms as well as relative to inflation, wages, and the cost of car ownership and use. In the 6 years between 1988 and 1994, for example, the price of a one-way tram ticket in Warsaw rose 400-fold, and the percentage of average hourly wage required to pay for that ticket rose from only 4% to 26%. While a liter of gasoline cost eight times as much as a tram ticket in 1988, it cost only twice as much in 1994 (Mitric and Suchorzewski, 1994; Pucher, 1995). In East Germany, public transport fares rose ten-fold from 1990 to 1992, while the price of gasoline fell. A liter of gasoline cost nine times more than a one-way bus or tram trip in 1990, but less than a tram or bus trip by 1992 (Pucher, 1994). Those dramatic shifts in relative prices obviously spurred the shift of travel demand from public to private transport. The situation in the Czech Republic was similar, but not quite so extreme. One-way fares in Prague rose seven-fold between 1989 and 1998, while the price of a liter of gasoline and the price of the average car both rose about three-fold (Czech Statistical Office, 2003).

Not only did public transport systems increase fares, but they also curtailed services, especially in smaller cities. Since funds were not available for modernizing or even maintaining the existing infrastructure and vehicles, service also became less frequent, less comfortable, and less dependable. Thus, both the quality and quantity of public transport services fell in most countries.

The percentage of operating costs covered by passenger revenues rose as fares increased and services declined. In Poland, for example, it rose from an average of about 40% in 1988 to about 65% in 1998, but with considerable variation and generally higher cost coverage in smaller cities. In Budapest, cost coverage rose from 35 to 43%. As in Poland, however, it is currently much higher in smaller Hungarian cities (ranging from 84% in Debrecen to 98% in Gyor) (Suchorzewski, 1999, 2002).

Public transport usage also fell in many countries after the end of socialism because of high unemployment rates caused by the widespread bankruptcies and closings of many formerly state-owned, highly subsidized enterprises, especially in the manufacturing sector. With unemployment rates of almost 20% in Poland, East Germany, and Hungary, there was a sharp fall in work trips beginning around 1990 and continuing throughout the decade. The Czech Republic avoided that problem for a while by continuing large subsidies to heavy industry, but as firm closings increased over the decade, unemployment rates rose there as well, from 0.2% in 1989 to 3.5% in 1995, 7.5% in 1998, and 8.8% in 2000 (Czech Statistical Office, 2003). That might help explain the much later fall in public transport usage in the Czech Republic than in East Germany and Poland, as seen in Figure 3.

Just as government policies in central Europe became much less favorable for public transport, they became much more accommodating to private car ownership and use. Virtually all restrictions on car ownership were removed, almost immediately opening up the central European market to foreign car-makers. That greatly increased the quantity and quality of cars that residents of formerly socialist countries could buy. As an economic development strategy, some central governments (such as in Poland and the Czech Republic) have strongly promoted their own car industries through loans and subsidies for expanding and modernizing car production facilities (Pucher, 1999; World Bank, 2002).

Although budgets have been strained at every level, many central and local governments have been devoting considerable expenditures to improving and expanding roadway networks, focusing on high-speed arterials, suburban beltways around cities, bottlenecks at key intersections, and connections to the main intercity and international routes. Thus, the supply of roadway infrastructure is increasing, although much more slowly than the rapid increase in car and lorry use. Similarly, most governments in central Europe still, in effect, set gasoline prices, either directly or by determining the level of gasoline taxation. As noted previously, no central government has raised gasoline prices by as much as local governments have raised public transport fares.

In other respects as well, restrictions on car use were either lifted or not enforced. That was especially true of parking, which became very problematic in large cities. In the first few years, parking regulations in both Polish and Czech cities were largely ignored, leading to what some officials described as "parking chaos." Since then, some cities have set up zonal parking systems enforced by private parking management firms. Nevertheless, the total supply of parking spaces has greatly increased, much to the detriment of historic central cities. In Prague, for example, many of the most scenic squares have been turned into virtual parking lots (Pucher, 1999).

Buses and trams are increasingly stuck in the traffic congestion generated by the skyrocketing car and lorry use. That obviously slows them down and further decreases the quality of public transport relative to car travel. A few cities such as Prague have begun introducing bus lanes and traffic signal priority for buses and trams at key intersections. Most cities, however, have undertaken no traffic priority measures at all to facilitate public transport movement on increasingly congested roads. Similarly, car-free zones and traffic-calmed neighborhoods, which are so common in western Europe, are rare in formerly socialist cities of central Europe, although a few cities such as Prague have restricted car access to parts of their historic cores.

6. Problems of modal shift in central Europe

While the dramatic shift from public transport to the private car generally reflects consumer preference for the convenience, comfort, speed, flexibility, independence, and status of the car, it has generated some serious problems: rising roadway congestion, parking shortages, air pollution, noise, and traffic crashes. In only 3 years between 1988 and 1991, traffic fatalities jumped by 34% in the Czech Republic, 43% in Hungary, 71% in Poland, and 109% in East Germany (Pucher, 1993, 1994, 1999). The sudden increase in car use, especially with faster and more powerful Western cars, overwhelmed the limited and dangerously designed roadway network. In addition, speeding and reckless driving increased, since enforcement of traffic regulations was lax in the first few years after socialism, partly as a reaction to the repressive police states that had existed previously. Driver training was also much less rigorous than in most western European countries. Since the early 1990s, roadway improvements, safer cars, better driver training, and stricter enforcement of traffic laws have all led to improvements in traffic safety in central Europe. For example, traffic fatalities in Hungary fell by more than half from 1992 to 2000, with current levels lower than during the socialist era (Hungarian Statistical Office, 2002). The traffic safety improvements have been less impressive in the Czech Republic and Poland, but even there, fatalities have declined since 1997, by 5 and 14%, respectively (Czech Statistical Office, 2003; Polish Central Statistical Office, 2003).

Likewise, some other acute problems that developed with the sudden jump in car use after the fall of socialism have been mitigated over the years since then, as adjustments to policies were possible. Thus, unleaded gasoline is now available throughout Hungary, the Czech Republic, and Poland, and most cars have catalytic converters. That has reduced the severity of certain kinds of pollution (nitrogen oxides, carbon monoxide, and airborne lead, for example). Yet some problems such as traffic congestion and parking shortages have gotten even worse as traffic volumes continue to increase more rapidly than roadway and parking supply. Traffic management in formerly socialist countries is still primitive or non-existent in most cities, exacerbating whatever problems the limited roadway capacity causes. The modest improvements in roadway infrastructure during the 1990s have helped divert through-traffic away from some city centers. Similarly, completion of some key missing links in the roadway network has also helped. Nevertheless, the lack of funding makes it virtually impossible to keep up with the rapid growth in roadway travel by private cars and lorries.

7. Further adjustments to transport policies

In response to the problems that came with sudden increases in private transport and the equally dramatic decline in public transport use, some countries have recognized the need to adjust their policies. In particular, there is growing recognition that unfettered car and lorry use cause significant social and environmental problems and that certain measures must be undertaken to control the negative impacts of private passenger and goods transport. We have already noted above some increasing restrictions on car use, such as stricter parking regulations, safety and environmental standards, driver training, and enforcement of traffic regulations. Tolls are also being charged on some motorways, forcing motorists to help finance them.

For the most part, however, there is so much political support for accommodating increased car ownership and use that it is difficult to implement policies that would greatly inconvenience motorists or significantly raise the price of driving. Perhaps the most frustrating problem is the refusal of local government officials to give buses and trams the traffic priority they need to insulate them somewhat from the seriously congested streets in many cities. While most western European cities long ago instituted bus lanes and priority traffic signals to ensure fast movement of buses and trams, only a few central European cities have even begun to adopt such crucially needed measures.

Nevertheless, local governments have at least given more attention to public transport as an essential part of the urban transport system. After the initial shock of the sudden transition to capitalism around 1990, public transport systems have been gradually recovering in recent years. As seen in Figure 3, passenger levels have stabilized in many Czech, Hungarian, and Polish cities. Through cooperation with western European experts and counterpart systems, many central European public transport operations have tried to improve the quality of their service, modernize their vehicles and infrastructure, and increase the efficiency of their operations. Prague's system, for example, has a partnership with the Paris public transport system (RATP), which provides frequent expert advice and assistance. The explicit focus has been improving and monitoring service quality.

While many municipal governments have undertaken vigorous measures to improve their public transport systems, their efforts have been largely frustrated by car-friendly central governments, which have provided very little funding, technical support, coordination, planning, or other guidance. Indeed, the ambitious road-building policies of central transport ministries have encouraged more sprawl and car use, thus further worsening the chances of rebuilding public transport's customer base.

In spite of the increasingly adverse environment for public transport, significant improvements have been achieved. Many cities have transformed their systems into publicly owned corporations, with considerable managerial independence for actual operations. While city governments still own the public transport systems and set overall fare and service policies, the corporate management team has more leeway to improve the efficiency of operations. That has increased customer orientation and focus on service quality. Some cities have also been selectively privatizing parts of their operations. Thus, Prague's suburban bus routes are run by private operators under contract to the main public transport system. In spite of longer routes and lower vehicle occupancy in the suburbs, the lower costs of privately run services enable the expansion of bus routes to outlying areas with minimal subsidy.

Even without adequate support from central governments, many local governments have undertaken a range of measures to improve their public transport systems.

Several cities have built new light rail lines (fast trams) or extended metro systems. Many cities have reconstructed tram tracks and track beds, modernized metro stations, and gradually replaced their aging bus, tram, and metro fleets with modern, Western-style vehicles. Some cities have also rationalized fare structures, improved fare collection systems, and introduced real-time information for passengers at tram and metro stops. Funds for public transport are so limited in most cities, however, that only a fraction of the necessary improvements can be implemented. That makes it difficult to keep up with the ever-increasing competition from the extremely popular private car, especially in the face of rampant suburban sprawl, whose low density, polycentric layout, and multi-destinational travel patterns are so adverse to public transport.

8. Impacts of EU policies on the accession countries

For two reasons, social, economic, and transport policies in many formerly socialist countries in central and eastern Europe are becoming increasingly like the policies of western European countries. The differences that developed during more than four decades of socialism after the Second World War arose from Europe's artificial political and economic division by the Iron Curtain. Central European countries, in particular, had been an essential part of Europe for many centuries, so the lifting of the Iron Curtain enabled the return to Europe of countries that had long belonged anyway. Thus, it is only natural that central European countries would quickly gravitate in many ways toward their western European neighbors. For certain countries, however, there was the additional factor of impending membership in the EU, which has explicit transport policies that all members must adhere to. Hungary, Poland, the Czech Republic, Slovakia, Slovenia, Lithuania, Latvia, and Estonia are among the new members that joined the EU in May 2004. Especially since harmonization of transport policies is a top priority of the EU, the impacts on transport systems and travel behavior are likely to be considerable (Harrop, 2000; World Bank, 2002; Peters, 2003). Indeed, the accession countries started adapting their policies to EU requirements in the 1990s as a pre-condition for approval as EU member states.

For the most part, EU transport policies affect long-distance transport, since they are aimed at restoring rail and road connections that had been interrupted during the Communist era. Thus, the Trans-European Transportation Networks (TEN-T) program established crucial north–south and east–west corridors that expressed the particular importance of improving transport links between all parts of the EU. Those TEN-T corridors have been extended through the accession countries and beyond. Fourteen specific projects were identified for central and eastern Europe (Turre, 1999). Most of them involve improvements in long-distance connections between the capital cities of Europe.

The EU assists in funding central and eastern European road and rail projects in these priority corridors both indirectly, through the European Investment Bank (EIB) and the European Bank for Reconstruction and Development (EBRD), and directly, through the EU Structural and Cohesion Funds and the TEN-T budget. Until 2004, however, transport infrastructure projects relied for most of their funding on the national budgets of each country, with some additional funds from private investment capital for a few selected projects. The tight budgets of most governments in central and eastern Europe explain the slow progress of many planned projects within the TEN-T corridors (Peters, 2003).

Starting in 2004, however, EU funding for transport infrastructure improvements in central and eastern Europe greatly increased for the new member countries. Moreover, further increases are planned around 2007, when the EU is scheduled to initiate its new 7 year cycle of European Regional Funding, which will for the first time include the new EU members in central and eastern Europe. The specific EU-approved projects and funding levels for these future years have not yet been determined, but it is certain that the EU will be providing a substantial portion of the funding. The overall increase in funding will facilitate the completion of planned projects that had been delayed by lack of financing prior to joining the EU.

EU transport policies only indirectly influence urban transport policies, since the EU explicitly states that urban transport is a local issue to be determined at a lower level of government. Nevertheless, some long-distance road and rail projects have clearly influenced urban transport. For example, the EU, EIB, and EBRD provided funding for the ring road around Budapest, which has unquestionably affected the urban transport network, generally encouraging more car use and suburban sprawl (Peters, 2003). Likewise, the EIB and EBRD have helped fund a few urban projects such as the Warsaw Metro and Krakow tramway (World Bank, 2002; Suchorzewski, 2002).

EU vehicle emissions and fuel standards will eventually apply to all roadway transport in the new member countries (Europaeische Umweltagentur, 2002; Volkswagen, 2004). New cars and trucks will have to meet the strict Euro III Standard that became effective January 1, 2001. It requires further reductions in tailpipe emissions of carbon monoxide (53%), hydrocarbons (67%), and nitrogen oxides (68%) relative to the Euro I Standard that became effective in 1992 and required catalytic converters on all cars (UK Department for Transport, 2003). The even stricter Euro IV Standard that becomes effective January 1, 2006 will require further tailpipe emissions reductions (relative to Euro III) in carbon monoxide (43%), hydrocarbons (33%), and nitrogen oxides (50%). The EU also sets standards for fuel composition to require successively cleaner fuels. Thus, the EU seeks to reduce transport-related air pollution not only by requiring cleaner and more efficient engines and catalytic converters but also by mandating cleaner fuels.

The EU requirement for open competition in the provision of local transport services will affect the organizational structure and economic performance of local public transport systems. Thus, central and eastern European countries will eventually be forced to change their legislation to conform to EU regulations that promote cross-border competition. There is the possibility that western European firms might eventually operate many central and Eastern public transport systems under contracts won in a competitive tendering process.

9. Conclusions and policy recommendations

After the turbulent decade of the 1990s, the new millenium has already brought more gradual change to the countries of central and eastern Europe. Much of the rapid increase in motorization was simply to catch up to western European levels.

With a much smaller gap now between car ownership rates in western and central Europe, there is less catching up to do. Car ownership and use will surely continue to grow, just as they are continuing to grow throughout Europe, but the growth will be far less explosive than during the 1990s.

Similarly, it seems likely that public transport use will continue to decline, but more slowly than in recent years. It certainly will not return to the artificially high levels of the Communist era. Public transport systems throughout central Europe are making efforts to expand and improve their services. They are fighting a difficult battle against the extremely popular private car, however. Even if they eventually manage to attain a western European standard of service quality, the best they can hope for is to stabilize usage at current levels. With public transport's modal split share falling throughout western Europe – in spite of superb public transport systems – it is certain that public transport will be serving a lower and lower percentage of urban trips in the formerly socialist countries of central Europe as well.

These trends mean that the countries of central Europe will have to deal with the sorts of urban transport problems plaguing western European cities for many years, the same problems that emerged so suddenly in formerly socialist countries during the turbulent 1990s: congestion, air pollution, noise, accidents, parking, and transport finance. Central European policy-makers and researchers are already looking to western Europe for guidance, and they have already begun adopting the same policy measures used in the EU. Technological improvements in car design, for example, should help mitigate the air pollution, noise, energy use, and safety problems. This is already evident in the new cars central Europeans are now purchasing, partly because they were manufactured in western Europe anyway. The most common Czech car, the Skoda, was vastly improved after the Skoda Company was bought by Germany's Volkswagen Corporation.

Congestion and parking problems will probably remain as intractable in central European cities as in the rest of the world. Elaborate pricing schemes (such as in Singapore) are virtually inconceivable, since there is still a strong backlash to the repressive controls under Communism. Expansion and refinement of the current, crude system of differential parking fees would probably help resolve the parking problem, and if well designed, might mitigate congestion in central cities as well. As western Europe and the USA have learned, massive additions to roadway capacity will not solve the congestion problem, since they generally induce more traffic and more suburban sprawl. Nevertheless, it is clear that the current roadway system in the central Europe needs some key new links to fill gaps. Moreover, at least some additional capacity must be provided to meet the huge new demand for car and lorry use that has emerged during the 1990s and which will surely continue to grow in the coming years, even if at a slower rate.

Central governments must take on more responsibility for urban transport. At the very least, they should help cities by supporting research, disseminating information about best practices, and establishing a legal framework for regional intermodal coordination of public transport systems. Moreover, local governments desperately need the financial assistance of central governments for crucially needed capital investment – through direct subsidies as well as loan guarantees. Local governments, for their part, must give buses and trams the traffic priority they need. Surveys indicate that the majority of central European citizens support giving public transport traffic priority even though that requires restrictions on car use (Suchorzewski, 2002). Local politicians should finally implement policies that reflect those preferences.

With their membership in Organisation for Economic Cooperation and Development, NATO, and now the EU, many central European countries will have to conform to all EU regulations, laws, and standards for transport. Moreover, central Europe seems to look toward western Europe as its model anyway. Thus, the transport systems in central Europe will become increasingly similar to those in western Europe. Central European countries could benefit from decades of western European experience. Although all EU countries depend primarily on the automobile for passenger transport, most of them offer excellent public transport systems and attractive environments for walking and bicycling. Achieving such a balanced transport system will not be possible without the vigorous support of central European governments at every level.

References

American Automobile Manufacturers Association (1978) *Motor vehicle facts and figures*. Detroit: Motor Vehicle Manufacturers Association.

American Automobile Manufacturers Association (1982) *Motor vehicle facts and figures*. Detroit: Motor Vehicle Manufacturers Association.

Broeg, W. and E. Erl (2003) *Verkehrsmittelwahl in Deutschland: neue und alte Bundeslaender.* Munich: Socialdata.

Czech Statistical Office (2003) *Transport yearbook (2002)*. Prague: Czech Statistical Office.

Europaeische Umweltagentur (2002) *Den Weg fuer die EU-Erweiterung ebnen: Indikatoren zur Integration von Verkehr und Umwelt, TERM (2002)*. Copenhagen: Scanprint (http://reports.eea.eu.int/environmental_issue_report_2002_24-sum/de/TERM_DE_web.pdf).

German Federal Statistical Office (2003) *International statistics*. Wiesbaden: Statistisches Bundesamt.

Harrop, J. (2000) *The political economy of integration in the European Union*. Cheltenham: Edward Elgar.

Hungarian Central Statistical Office (2002) *Hungarian statistical yearbook (2001)*. Budapest: Central Statistical Office.

Institute of Transportation Engineering (1992) *Yearbook of transportation (1990)*. Prague: Institute of Transportation Engineering.

Institute of Transportation Engineering (2003) *Yearbook of transportation (2002)*. Prague: Institute of Transportation Engineering.

Komornicki, T. (2003) "Factors of development of car ownership in Poland," *Transport Reviews*, 23:413–431.

Mitric, S. and W. Suchorzewski. (1994) "Urban transport in Poland: the challenge of the ascending private car," in: *Proceedings of the conference "Reconciling Transportation, Energy, and Environmental Issues."* Paris: Organisation for Economic Cooperation and Development.

Peters, D. (2003) "Planning for a sustainable Europe: a case study of EU transport infrastructure policy in the context of eastern enlargement," Ph.D. dissertation. New Brunswick: Rutgers University.

Polish Central Statistical Office (2003) *Transport statistics (2002)*. Warsaw: Polish Central Statistical Office.

Pucher, J. (1990) "Capitalism, socialism, and urban transportation," *Journal of the American Planning Association*, 56:278–297.

Pucher, J. (1993) "Transport revolution in Central Europe," *Transportation Quarterly*, 47:97–113.

Pucher, J. (1994) "Modal shift in East Germany: transportation impacts of political change," *Transportation*, 21:1–23.

Pucher, J. (1995) "Road to ruin? Impacts of economic shock therapy on urban transport in Poland," *Transport Policy*, 2:5–13.

Pucher, J. (1998) Urban transport in Germany: providing feasible alternatives to the car. *Transport Reviews* 18:285–310.

Pucher, J. (1999) "The transformation of urban transport in the Czech Republic, 1988–1998," *Transport Policy*, 6:225–236.

Suchorzewski, W. (1999) "The funding of public transport investment in Central Europe," in: *Proceedings of the conference "Financing Urban Public Transport."* Paris: Conference of European Ministers of Transport and Union Internationale des Transports Publics.

Suchorzewski, W. (2000) *Urban transport in Poland*. Warsaw: Ministry of Transport.

Suchorzewski, W. (2002) "Urban public transport in Poland: main issues and perspectives," in: *UITP Conference "Changing European Market and Global Industry."* Brussels: Union Internationale des Transports Public.

Sykorova, I. and L. Sykora (1999) "Prague metropolitan area," in: *Urban land institute market profiles 1998: Europe*. Washington, DC: Urban Land Institute.

Sykora, L. (2002) *Suburbanization: the social, economic, and ecological impacts*. Prague: Institute of Ecopolitics.

Turre, M. (1999) *Going trans-European: planning and financing networks for Europe*. New York: Pergamon.

UK Department for Transport (2003) *European emissions standards*. London: Highways Agency, UK Department for Transport (http://www.haguidetofreight.co.uk/General/id94.htm).

Volkswagen (2004) *Mobilitaet und Nachhaltigkeit: Umweltbericht 2003–2004*. Wolfsburg: Volkswagen (http://www.mobilitaet-und-nachhaltigkeit.de/_download/umweltbericht_2003_2004_deutsch.pdf).

World Bank (1998) *World development indicators (1998)*. Washington, DC: World Bank.

World Bank (1999) *World development indicators (1999)*. Washington, DC: World Bank.

World Bank (2002) *Urban transport in the Europe and central Asia region*. Washington, DC: World Bank.

World Bank (2003) *Quick reference tables*. Washington, DC: World Bank (http://www.worldbank.org/data/databytopic/GNIPC.pdf).

This Page Intentionally Left Blank

Chapter 43

TRANSPORT POLICIES IN ASEAN COUNTRIES

ANTHONY T.H. CHIN
National University of Singapore

1. Introduction

The Association of Southeast Asian Nations (ASEAN) was formed in 1967 by five countries, which by 2000 had expanded to ten (Table 1). Unlike the EU, ASEAN does not have a coordinated transport policy, due, in part, to the insularity of the region and in part to the intense competition between member states. However, the emergence of China and India as economic powers has lead to a degree of urgency in organizing the region's transport. The challenges involve the growth of the Indian and Chinese economies and diversion of foreign direct investments (FDI) away from ASEAN. There is increasing competition within ASEAN, with China becoming increasingly attractive for FDI as supply chains have shifted over to China. This is within the context of significant disparities in gross domestic product (GDP) (Table 1).

There has been rapid technological change in ASEAN since the 1980s. Reductions in the costs of transport have largely been brought about through technological advancements in air travel and light commercial transport, and simplification of government regulations. Rising incomes have led to demand for more personal mobility. However, multilateral cooperation in the form of an ASEAN transport strategy is absent. One reason is the lack of a regional perspective. Most training programs and practices are national in focus. Cooperation in transport falls under the supervision of the Committee on Transportation and Communication, with sub-committees to implement decisions on major projects. But only about 20% of projects have been carried through, and in many cases projects have focused purely on data gathering or technical assessment.

In November 2002, the ASEAN leaders stressed that "the liberalization of intra-ASEAN trade in services required strengthening transport links, interconnecting telecommunications, increasing the use of information and communications technology, and liberalizing investment in these sectors." Earlier in September the ASEAN Memorandum of Understanding on Air Freight Services was signed, as a step toward the full liberalization of air freight services. The understanding allows

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

Table 1
GDP per capita (US $)

Country	Area (km^2)	Population (c)	Year 1996	1998	2000	2002
Singapore	648	4.0	25 127	20 892	22 757	20 515
Brunei	5 765	0.342	17 096	11 961	12 751	12 090
Malaysia	329 758	23.3	4 766	3 257	3 874	3 914
Thailand	513 115	62.3	3 134	1 900	2 026	2 043
Philippines	300 000	78.4	1 184	896	980	974
Indonesia	1 919 317	212.0	1 167	488	731	819
Vietnam	331 700	77.5	337	361	403	439
Cambodia	181 035	11.0	312	268	293	299
Lao PDR	236 800	5.3	393	244	333	329
Myanmar (a)	676 575	49.0	109	144	184	104
ASEAN (b)	4 494 713	523.1	1 504	947	1 125	1 144

Source: ASEAN Surveillance Coordinating Unit database.
Notes:
(a) Fiscal year April to March. Calculated using the average parallel exchange rates as used in the IMF WEO of April 2003.
(b) As an approximation, the combined GDP of ASEAN is calculated as the sum of the GDP of ASEAN members.
(c) Millions in 2000, except for Cambodia and Lao PDR (1999).

the designated airlines of member countries to operate all-cargo (with third/fourth freedom air traffic rights) services up to 100 tonnes weekly without limitations on frequency and aircraft type. A Roadmap for an ASEAN Competitive Air Services Policy is under consideration.

The ASEAN Land Transport Infrastructure Integration Roadmap sets out implementation plans for realizing the ASEAN Highway and Singapore– Kunming Rail Link (SKRL) project. The Transport Facilitation of Goods Roadmap was formulated to provide an efficient system for operationalizating ASEAN transport facilitation agreements covering transit and inter-state transport of goods. It is intended to cover multimodal transport operations, to support the ASEAN Free Trade Agreement and improve transport logistics in the region. The Special Working Group on SKRL agreed to undertake convening of a high-profile seminar/ forum, establishment of a specialized project monitoring body, establishment of a multinational consortium, and development of an incentive package for the private sector.

ASEAN has stepped up its transport cooperation with China and Japan. ASEAN and Chinese ministers agreed that building an integrated transport network to facilitate the smooth, speedy movement of goods and people is important for a free trade area. The terms of reference for ASEAN–China transport cooperation in land transport infrastructure development and facilitation, including implementation of the SKRL project; facilitation of maritime and river transport; and the expansion of air transport. The ASEAN, with China and Japan, are working toward policy dialog, enhancing human resource development, cooperation in the facilitation of cargo distribution and logistics, promotion of safer shipping, and enhancement of air transport safety and efficiency. ASEAN–India joint cooperative activities involve promoting transport infrastructure development and enhancing human resource development in the road, shipping, and railway sectors. Contacts with transport-affiliated private sector organizations are maintained through ASEAN bodies such as the ASEAN Airlines Meeting, the ASEAN Federation of Forwarders Associations, the ASEAN Ports Association, the Federation of ASEAN Shipowners, and the Federation of ASEAN Shippers' Councils.

2. Transport policy, economic growth and development

There is considerable diversity amongst the ASEAN economies. Table 1 highlighted some of the variations, but there are many others. Part of the diversity is in terms of physical geography, allowing, for example, some nations to develop as major maritime trading nations, whereas others are land-locked. Human geography also diverges, with quite dramatic differences in features such as levels of urbanization. In the short term the 2005 tsunami also impacted many national transport networks. Different levels of economic development, however, create the major differences in needs for and long-term priorities in transport policy. The economies of ASEAN member states can be classified into high, medium, and low income in terms of economic development.

2.1. Low-income economies

The transport challenges confronting the lower-income economies have common features but also some local distinctions. The major challenge for these very low-income countries is essentially financial, but there are national differences under this umbrella.

In Cambodia, for example, the greatest priority is to repair the dilapidated sections of the primary road network and to provide essential equipment and material for the operation and maintenance of transport services. The nation's main infrastructure is in a state of disrepair after years of neglect and destruction in the 1970s, with the result that it is underutilized and unable to meet even the modest demands placed upon it.

Numerous rehabilitation projects are being undertaken or planned to meet the most urgent needs. From 1989 the thrust toward the market economy led to government-initiated policy changes in the sector to alleviate some of the deficiencies in the delivery of transport services. Changes include the gradual devolution of financial autonomy to state enterprises and the privatization of some services. Privatization efforts have proceeded rapidly. In April 1989, a party congress adopted a proposal to allow the establishment of privately owned passenger and freight transportation services. The role of the state would be to collect tax, maintain transport infrastructure, set policy, and regulate activities in the transport sector, which includes technical and safety inspection of vehicles, route licensing, and establishment of maximum fare rates. The private sector has already expressed interest in ferry operations and cargo handling at ports and railways.

The Lao PDR has the particular challenge of being landlocked as well as suffering from an extremely low-income level. It occupies a strategic position in the regional transportation network; routes between Thailand and Vietnam and Thailand and Yunan have to cross the country. The government has signed an agreement with Vietnam on land transport, and with China on cargo and passenger transport on the Lancang–Mekong River and there are discussions on joint railway and transit transport with Thailand.

There is evidence from the most recent Lao Expenditure and Consumption Surveys of a correlation between accessibility and poverty; the level of poverty being higher in the less accessible northern and southern regions, and the incidence of poverty in rural areas is much higher for communities without road access. Crucial hurdles before Lao PDR, however, are the human resource and institutional capacity bottlenecks (UN Development Program, 1997). To meet this challenge, the budget allocation for investment in new roads has increased, but not for improvements and maintenance of current roads despite the aim of the Road Maintenance Project to promote the sustainable preservation of road assets through the development and implementation of financing and management maintenance systems (World Bank, 2000). Given the absence of a policy link between road use and its deterioration, the level of road user revenues allocated to maintenance has never been adequate to ensure preservation of the network.

Myanmar's transportation network is not well established, and transport policy is largely focused on infrastructure development rather than no traffic demand management, although there is a tilt toward market reform and the encouragement of private sector involvement. Transportation policies are designed to develop physical transport infrastructure, and on improving the quality of services, and promoting efficiency and safety. Cross-border trade is becoming important, and is affecting the geography of the network.

The development of transport infrastructure in Myanmar, however, lacks funding. In September 1988, the State Law and Restoration Council announced a move to a more open market-oriented economic system, and there is now some private–public cooperation. Involvement of the private sector is evident in the rail sector. Since 1992, Myanmar Railways has permitted private companies to operate freight trains and passenger trains on certain specified sections of the routes.

The Myanmar situation contrasts markedly to that of Indonesia, where rapid urbanization and growth in motorization is putting pressure on transport infrastructure. Vehicle ownership in Indonesia is growing at the rate of 8–12% annually, whereas road construction projects add capacity at only 7% per annum. A major issue is the weak productivity of the system, and especially the low occupancy of cars; in Jakarta only 4% of cars have more than 1–2 passengers during the peak hour (Hook, 1998). As an alternative, Indonesia's public transport system often lacks quantity and quality, with terminals and control systems poorly planned. The present level of public transport services is substandard and unattractive. Public transport is also expensive to provide, in part because of its dependence on expensive imported spare parts, and subsidies are high (World Bank, 2001a).

Turning to the maritime sector, to ensure a smooth flow of freight and passengers, new pioneer quays and passenger terminals have been constructed and old ones renovated, and shipping safety has improved by installing better navigation systems (lighthouses and light signs). To support under-developed Indonesian regions, the government subsidizes a range of ferry operations. Air transport continues to be developed to promote tourism and economic development. Airport facilities, flight safety, and air traffic management are being improved, and isolated areas are being opened up.

The intra-city transportation demands of the Philippines are heavily concentrated on Manila. Car ownership in the Philippines has grown tremendously since the 1980s. Traffic congestion and pollution are now major problems; about 70% of the air pollution in Metro Manila is from automotive sources. Urban sprawl and high population growth, coupled with widespread urban poverty, have put immense pressure on the transport infrastructure.

There are no restrictions on car ownership, and fees for and taxes on cars are minimal, although tolls are levied on intercity routes. Recently, attempts at controlling usage, such as a scheme whereby cars with even-number registration plates are allowed on the road on certain days and cars with odd-number plates on other days, have been initiated. Color-coding schemes similar to the car registration number scheme have also been tried. Unfortunately, the low cost of cars has led to multiple ownership rather than less use of cars. The poor security and safety record of public transport have also impeded modal transfers; even so, the modal share of public transport in Metro Manila is remarkably high.

The existence of squatter settlements, and unplanned residential areas, make the physical planning of transportation for inter- and intra-city transportation investment difficult. While the Philippine government's Medium-term Philippine

Development Plan does address transport matters, its investment program rests on strong assumptions about investment by the private sector and local governments, particularly in road development. Separate agencies regulate the transport subsector in the country, with an interagency body coordinating activities. A separate toll regulatory board has regulated the nation's toll roads since 1977. Private companies own most public transport services, but the government retains ultimate control of fares. More generally, the country's experience in toll roads, water, electricity, and transportation has made the Philippines a test bed for public–private sector partnerships (World Bank, 2001b). The continued evolution of private sector partnerships in developing urban infrastructure will promote capital inflows from the private sector's perspective, which is important in a country where public sector investment is less than 0.25% of GDP in urban areas (Asian Development Bank, 2002).

Privately owned "jeepneys" are a popular form of public transport. The government controls fares and constructs the infrastructure for pick-up points. The fact that they operate over fixed routes with predetermined schedules makes it easier to standardize fares. Buses in the Philippines are insignificant in terms of public transport; they are privately run, and the companies determine the routes and fares within a government ceiling. Collusion, however, often results in overcrowding on certain routes. Government involvement is thus minimal and limited to investments on terminals (Bayan, 2001).

A light rail system has operated in Manila since 1985. This was privately run but reverted to the local transport authority when the company failed to maintain operating standards. Fares on the system are very low, and there is heavy reliance on state finance. Although the system is running at full capacity, the revenue collected is not sufficient to cover growing costs, and thus there is a need for heavy government subsidy. A second system was completed in the late 1990s by a consortium of private companies on a build, operate, and transfer arrangement, and in September 1996 a consortium of private real estate and retail interests began constructing an additional light railway. To maintain low fares, the government has guaranteed an annual lease payment of 5.5 billion pesos to the consortium, effectively subsidizing fixed and operating costs.

The Philippines National Railroad is publicly owned and runs through much of Luzon, although ICTSI, a private consortium, runs a twice-daily service moving modest container traffic out of Manila. The railway relies heavily on government assistance to finance its capital expenditure, debt servicing, and operations; less than half of its operating costs are met from revenue. Compared with other railroad systems in ASEAN, the Philippines National Railroad is underutilized because of competition from coastal shipping, and its low speeds and unreliable service.

The airports in the Philippines are built, managed, operated, and regulated by the Air Transport Office, which is under the direct supervision of the Department of Transport and Communications. The airports generate their own funds, but rely on the annual budget of the department for support. The government acts as a regulatory body for aviation safety and airport security, and takes measures to ensure internationally recognized safety standards are met by the airlines. The government liberalized the domestic airline industry in 1995 in a bid to make the industry efficient, but this coincided with a financial crisis and was hampered by inadequate infrastructure. The Philippines airlines inherited a huge debt problem, and several routes were cut, and the frequency of flights was severely affected.

Despite its inherent technical limitations (almost half of the country's road network of 15 000 km is in poor to very poor condition), transportation in Vietnam has been and remains an important element in its economic growth. The linear nature of the country also means it serves an important cohesive function. Roads remain the dominant transport mode in Vietnam, with other transport modes such as air, sea, and rail still in their infancy. The growth has not all been in inter-urban movements, and mobility has risen in cities such as Hanoi and Ho Chi Minh City, where traffic speeds on the main roads have fallen to 20 km per hour.

Investment has focused on restoring basic transport infrastructure. Within this framework, the government has also concentrated on enhancing accessibility in poverty-stricken areas. The quality of the infrastructure varies across the country, with the majority of new investment taking place in high-growth areas and corridors around major urban centers. Rural areas, especially in the north, have lagged behind. A primary objective is the modernization of assets on the main traffic arteries, with the largest expenditures on the roads and bridges that serve 70% of the country's freight and 80% of its passenger traffic. This involves the rehabilitation of the main north–south trunk road and roads in the northern growth triangle. In addition, about 11 000 km of rural roads are being upgraded or repaired, with the local populations contributing more that 50% of the cost.

A program to rehabilitate for the main corridors in the Mekong Delta and the main waterways in the Red River delta has begun. The two main international airports have been modernized to cope with the rapid growth in international trade. In addition, minor improvements have been made to some seaports, and preparations are underway to build additional deepwater ports in the north and south to serve the larger container and bulk ships.

The Vietnam Ministry of Transport has been reorganized to focus on management functions, involving creating and implementing policy, setting standards, planning, budgeting, and auditing results. This has de-emphasized the daily micro-management of infrastructure and transport services. Separate administrative arms have been introduced for each major sub-sector; each has been given substantial autonomy. Special project units have been formed to deal with the management of large-scale rehabilitation projects. Air transport has been removed from the Ministry of Transport's purview, and is now under the direct control of the prime minister. In the urban transport sector, the Ministry of Transport has been similarly divested of much responsibility, and now plays only an advisory and coordinating role in the large cities.

Vietnam, however, still lacks a systematic approach to road asset management and maintenance. Management and funding responsibilities are complex, highly decentralized, and fragmented. There is little coordination over asset life, since responsibilities for construction, use, and maintenance rest with different agencies. This lack of coordination impacts most on the Vietnam Road Administration, the agency with maintenance and/or assets use responsibilities. Despite the government's efforts in improving transport, maintenance of existing assets has been a low priority: about 85% of all transportation expenditure is allocated to new investments. The World Bank (1999a) estimate that a more balanced program would include at least 20% for maintenance and minor improvements. Maintenance practices are also of low quality, and are mainly executed by a weak form of quasi-force account through state-owned enterprises. More than 60% of the budget for routine maintenance is spent on administration and other indirect costs. The government has narrowed its direct control over transport with market-based policy reforms, but these have been limited.

The government has made progress, however, in introducing market-based management ideas into transport regulations and the management of public transport. It has ended the process of allocating cargo to specific modes of transport, and has liberalized tariffs for road and inland waterway freight transport. In addition, it has relaxed entry and exit controls (notably licenses) for cargo and passenger transport. Other measures include allowing the private sector to invest in and operate transport infrastructure and abolishing direct operating subsidies for state transport enterprises previously bestowed upon them. Technical regulatory issues have also been addressed, e.g. the issue of decrees and regulations on road safety, urban and rail traffic, and inland waterways, including specific penalties for violations and offences.

2.2. *Medium-income economies*

These ASEAN countries, while still not affluent, have moved up the ladder in terms of average income, but still suffer from spatial variations in their economic performance and retain a heavy dependence on primary industries. As with the lower-income group, these nations are fairly heterogeneous in terms of their geography and inherited legacies. They all have more, although not extensive, resources at their disposal for formulating transport policies.

Malaysia is, for example, an intermediate economy that manages transportation rather differently to the poorer ASEAN nations. Rapid rates of urbanization and accelerated growth in vehicle traffic led to severe traffic congestion in Kuala Lumpur and other urban cities and imposed high economic and environmental costs. The focus on the provision of a well-integrated and inter-phased public transport system in the late 1980s was thus intended to complement various demand management measures. The development of urban public transport became a focus of the Third Malaysia Plan. The subsequent Integrated Transport System strategy is designed to alleviate urban traffic congestion, as well as mitigate the negative environmental impacts. Additionally, Malaysia has an interest in improving transport safety; for example RM400 million were in the 2001 budget for the development of an intelligent transport system.

Malaysia's road system of 63 445 km is among the finest in Asia. The interurban North-South Expressway, the New Klang Valley Expressway, and Federal Highway Route 2 are the largest segments of road transport infrastructure in the Malaysian Peninsula, with the 848 km expressway system linking major industrial areas and urban centers in the country. Some RM 21.2 billion was allocated to the investment in transport sector under the Eighth Malaysia Plan (2001–2005) (Government of Malaysia, 2001), with most of the money being concentrated on roads that are seen as important in aiding economic development in rural areas. The development of this network has always been a major priority since the First Malaysia Plan (there are 4100 km of rural roads in west Malaysia). The concept of an all-weather road system with a network of feeder roads linking the main coastal towns to the interior received particular attention in the Fourth Malaysia Plan. The policy on rural road development was initially laid out in the Mid-term Review of the Seventh Malaysia Plan (Government of Malaysia, 1999); namely "To provide a comprehensive range of infrastructure facilities and amenities to facilitate the revitalization of the agricultural sector as well as to enhance the development of less accessible regions in line with balanced regional development." The road development in Sabah and Sarawak has been slower, although development in these agriculture-oriented states is constrained by poor roads.

There are ten major ports in Malaysia, five of which are in east Malaysia. The National Port Authority and 11 regional port authorities handle the operations of over 30 ports. Maritime transport facilities are being expanded and modernized to cope with economic growth and trade; for example, in the Seventh Malaysia Plan, new port capacities were included with a view to promoting multimodalism through the development of a comprehensive system of land-side facilities and services. The privatization of port facilities was accelerated to increase the operational and managerial efficiency of these facilities. On the whole, the various objectives of the previous plans are being pursued in the Seventh and Eighth Malaysia Plans.

Priority is being given to improvements in port efficiency. Port developments will continue to expand capacity and increase equipment and facilities to cope with the greater traffic. In particular, measures have been introduced to improve efficiency in the clearance of cargo and ships at ports. These include the implementation of an electronic data interchange (EDI) system at Port Klang, and the simplification of trade documentation procedures and clearance systems at ports. In view of the rapid privatization of Malaysia's ports, a National Port Authority was established during 1996–1997 to act as a regulator as well as oversee the implementation of port development, policy, and strategy.

Airport policy objectives were more comprehensively developed in the Sixth Malaysia Plan, including maximizing the use of capacity through hub-and-spoke structures, with the main airports being fed by the lighter-traffic facilities; continuing efforts to upgrade rural airstrips in east Malaysia, to increase the accessibility of remote areas; greater efficiency in the use of capital in airport development; the corporatization of airports; and maintaining a liberal and open sky policy to encourage foreign airlines to provide services into Malaysia.

The first phase of Kuala Lumpur International Airport, opened in 1998, came as part of a policy to develop the airport as a regional hub. Located along the Multimedia Super Corridor, three expressways have been built to enhance the link between the facility and Kuala Lumpur, and a free trade zone has been created around it. The Express Rail Link connects Kuala Lumpur's new railway station to the airport. Due to the downturn in the world economy in the late 1990s there was consolidation of the airline industry and cost cutting. Efforts to obtain additional rights for Malaysian carriers and to attract more foreign airlines were stepped up with a liberal approach to air service agreement being pursued.

The privatized Malaysian railway network runs from Johor Bahru in the south to Perlis on the Thai border, along the west coast of Malaysian Peninsula. A second line runs from Gemas to Kota Bahru in the north-east. These 1640 km of railway operation are under the responsibility of the Malaysian Railway Administration, which was privatized in 1996. Sabah also has a railway of 154 km, but Sarawak has no railway. Measures to help the railways compete with road transport included introducing more effective managerial and operating techniques, and improving the facilities and equipment, including using more air-conditioned coaches. Inter-modalism is being promoted through the increased provision of additional capacity. The Seventh Malaysia Plan emphasized inter-state passenger travel with a priority on increasing reliability and speed through selective doubling, strengthening, and electrification of tracks (Government of Malaysia, 2001). In addition, light rail transport (STAR and PUTRA) has been introduced to complement commuter services to alleviate the problem of urban congestion in Kuala Lumpur.

Thailand has been investing in infrastructure projects at an annual average rate of 5–6% of GDP, with the Seventh National Economic and Social Development Plan allocating 58% of the infrastructure budget to the transport sector. The country's roads are under the supervision of the Department of Highways, and the network comprises 51 775 km of national highways, 59 200 km of distributor roads, and 47 450 km of rural roads. The Expressway and Rapid Transit Authority of Thailand operates the 150 km of expressways within the Bangkok Metropolitan Region.

The railway system is extensive, with 4119 km of track. Part of it links Bangkok to the eastern seaboard area, and is seen as promoting the Eastern Seaboard Development Program. The network does not, however, have the same priority as other modes, but efforts are being made to upgrade it. A new inter-regional track is under construction to link Thailand to Laos and China. The State Railway of Thailand is also working on doubling the length of the 249 km of commuter lines in the Bangkok Metropolitan Region. Two urban rail systems, the Bangkok Mass Transit System and the MRTA Blue Line System, currently serve the Bangkok Metropolitan Region. The Bangkok Railroad Improvement Project seeks to alleviate the chronic traffic problems in Bangkok and to provide a fast and more efficient means of transport.

There are 29 commercial airports in Thailand. Bangkok International Airport can accommodate 33 million passengers a year, and five regional airports (Chiang Mai, Chiang Rai, Hat Yai, Phuket, and U-Tapao) can accommodate 12 million passengers. A second Bangkok international airport (Suvarnabhumi), due to open in 2005, is designed to accommodate 30 million passengers and 1.46 million tonnes of air cargo per year. Thailand's five principal ports (Bangkok, Laem Chabang, Map Ta Phut, Songkhla, and Phuket) dominate trans-shipment. There has been an emphasis on increasing port capacity to allow the country to serve as the gateway for the South-East Asian sub-region.

Policies for Thailand's transport sector have developed primarily to promote national economic growth, with a gradual increase in their definition and focus (Bunyasiri, 1994). More attention is given to transportation infrastructure development compared with usage and pricing. The main objectives of the Transportation Master Plan 1999–2006 aim at facilitating:

- the efficiency of the national transport and communication system so that adequate services are provided to a standard quality and at reasonable prices;
- the integration of transport modes into a unified system so that the costs of services and use of energy are reduced while still taking into account safety, impacts on the environment, and efficient utilization of resources;
- the use of transport and communication to promote and support the development of some specific areas, to reduce disparities among regions in Thailand;
- the systematic expansion of transport and communication services to match demand for the period to 2006, so that Thailand's networks link with those of neighboring countries, enabling the Thai economy to be at the forefront in the region.

Transportation infrastructure investments continue to play an important role in determining the overall growth of the transportation sector. The development of infrastructure for the transportation network is in accordance with the Ninth National Economic and Social Development Plan, which aims to develop air transport, to make Thailand a regional transportation center; land transportation, to improve connections with other modes and to assist economic activity in various areas; and domestic water and international maritime transportation to their full potential by coordinating public and private sector activities.

Transport projects have, though, been delayed through lack of funds. Only a small proportion of the national budget was allocated to transport projects until recently. The government has attempted to attract private sector participation in infrastructure investment, with mixed success. One of the prime targets of Ninth National Economic and Social Development Plan (2002–2006) is to reduce the congestion in Bangkok. Part of the strategy has been an acceleration of the construction of infrastructure (rail mass transit, urban and inter-urban expressways, a ring road, the suburban local road network, etc.), with the promotion of private sector participation in implementation and operation of the projects to improve efficiency and flexibility.

Policies to control pollution from transport have been implemented. Some, directed toward reducing vehicle emissions, include fuel reformulation, use of catalytic converters by cars, and emission standards. Leaded gasoline has not been available in Thailand since 1996, and catalytic converters have been mandatory for all cars since 1993 (Sayeg, 2002). Stringent emission standards for new vehicles in Thailand based on EU laws were established in 1995 (Srisurapanon and Wanichapun, 2001). However, inspections have not always proved efficient. Parsons International (2000), for example, estimated that improved inspections and vehicle maintenance could reduce emissions by around 30–50%. One obstacle faced by the region when trying to reduce pollution is the inefficiency of the public bus transport services. Bangkok's bus system is directly or indirectly operated by the Bangkok Mass Transit Authority; a state enterprise monopoly. The World Bank's urban transport sector review for Bangkok (World Bank, 1999b) identified the major problems of the Bangkok Mass Transit Authority as financial stress and shortage of funds to pay for maintenance. Further, while almost 30% of Bangkok's trips are made by bus, bus lane regulations have not been enforced effectively.

2.3. *High-income economies*

The transport sector in Brunei is seen as a catalyst for the growth of the economy. Investments in transport are based on the analysis of interdependencies, complementarities, and potential areas for cooperation in regional development.

The policy initiatives now in place include facilitation of cross-border movements between Brunei and Malaysia; inter-airline collaboration in aircraft maintenance and human resources development; expanding cooperation among small airlines in the provision of rural services; and cooperation between the ports of Miri, Labuan, and Muara. Within this broader framework, specific projects include completion of the Pan Borneo Highway; feasibility studies for new road links between Brunei and Sarawak; and studies for the development of a bulk cargo terminal near Lawas, Sarawak, or within Brunei.

The policies and projects have been divided into three phases: the short term, which runs from 1 to 2 years, the medium term, which runs from 3 to 5 years, and the long term, which runs beyond 5 years. The identified phasing relates to the period of implementation of the projects rather than to the commencement of actions with respect to the projects. These projects and policies are designed to jump-start the whole development process; they can be implemented quickly as they are readily feasible since funds are already identified. No major land acquisition is needed and no bureaucratic delay is expected.

There has been extensive investment in transport and communications infrastructure facilities and equipment in Brunei, in particular on the air transport system, post and cargo handling facilities, and the telecommunications infrastructure, in response to the strategy of connectivity to international networks; but this has resulted in much spare capacity. The road network is relatively comprehensive and in good condition. Most motor vehicles are new and well maintained.

The transportation services industry in Brunei Darussalam involves a range of publicly owned and commercial organizations. The government departments overseeing the transport sector often double up as regulators, operators, suppliers, and contractors of the services. A constant review of policy is undertaken to improve on the efficiency and competitiveness of the transportation services and in accelerating the economic development. This includes the pace of privatization and liberalization of the services.

Brunei has yet to establish a comprehensive legislation that contains regulations and standards pertaining to protection and control of air quality. Currently, internationally accepted standards such as World Health Organization standards and those of ASEAN member countries are used as references. Another ongoing concern is vehicular accidents due to excessive speed, poor driver education, and carelessness.

Transport infrastructure remains a priority, especially the rehabilitation of trunk and feeder roads and bridges to help realize the potential of agriculture, tourism, and trade in rural areas. There is a need, however, a need to develop a comprehensive transport policy framework, addressing issues such as the development of a balanced construction and maintenance program, increased involvement of the private sector over time, and financing of road maintenance and cost recovery mechanisms. With regard to institutional strengthening, the

Ministry of Public Works and Transport intends to formulate strategies to improve its capacity to plan, manage, and implement road operations.

Economic pragmatism and survival necessitated the growth of air and sea transport in Singapore. Service providers such as the Neptune Orient Line (NOL), Singapore Airlines (SIA), and the Port of Singapore Authority (PSA) accelerated the pace of developments further. This proactive stance excluded issues pertaining to land transport, which was constrained by social, economic, and urban developments, and the long-term implications of transport planning and provisions (Urban Redevelopment Authority, 1991).

The 1970s saw rapid growth in the demand for air transportation in South-East Asia, and presented immense opportunities for Singapore. The aviation industry became regarded as a key driver in industrialization. An important milestone in air transport was the development of Changi Airport, the emergence of a competitive national carrier (SIA), and a liberal air policy. Singapore's infrastructure over-provision strategy aimed at the development of a hub-and-spoke airline network in South-East Asia, although not with narrow short-term consistent cost minimization, is seen as a strategic move to maintain its leadership position, and to function as a first-mover deterrent to entry.

A liberal air policy has encouraged airlines to support air links that promote investment, trade, and tourism flow between Singapore and her partners. Since 1997, Singapore has concluded several open skies air service agreements. In addition to providing benefits not available under bilateral arrangements, such as cross-border financing for international airlines and the opportunity to capitalize simultaneously on a network of open skies linkages with clusters of economies, liberalization is expected to generate greater economic growth through closer trade collaborations (Chin, 1997; Bowen, 2000). Foreign airlines are permitted to operate in Singapore even when SIA does not fly similar route; the rationale being open skies generate economic benefits *per se*, while competition helps to sharpen SIA's competitive edge. The air transport regime also allows airlines to determine their prices according to market forces (Chin and Tongzon, 1998).

Singapore is a major maritime hub. The role of the Maritime and Port Authority of Singapore was mapped out in 1996, and it has the responsibility for drawing up the national port master plan (Maritime and Port Authority, 2000). Generic strategies have been proposed (Singapore Ministry of Trade and Industry, 1998) to strengthen Singapore's hub position with an emphasis on the need for an integrated transport and logistics hub philosophy entailing a total approach to air, land, and sea transport.

Leveraging on information technology and electronic infrastructure has been a strategy used by the Port of Singapore to remain competitive. To better control shipping activities in the port, the Ship Identification and Positioning System was completed in 1999, and at the same time an e-commerce portal, MARINET, was launched to facilitate e-transactions in the shipping community. The Singapore

Maritime Portal (SMP) acts as an e-commerce and information gateway for the global maritime community (Maritime and Port Authority, 2000). The establishment of these portals complements existing services such as TRADENET and PORTNET. The Ministry of Transport also continues to liberalize port and marine services to help Singapore provide high-quality and cost-competitive value-added services, and to meet intensifying competition. The Maritime and Port Authority has launched several initiatives to develop Singapore into a center of excellence for maritime research and development, e.g. the Singapore Maritime Academy provides comprehensive and fully integrated simulator training in ship handling, crisis management, and vessel traffic services.

Regarding surface modes, mobility and accessibility together with the twin concerns of unemployment and housing were priorities in Singapore during the 1960s. Land transport reacted to the overall needs of economic development. In the mid-1990s a comprehensive rail network was envisaged, following the release of a government White Paper (Land Transport Authority, 1996). Accessibility and mobility gave way to the desire for comfort, convenience, connectivity, and affordability. However, with the desire for mobility comes automobile ownership and traffic congestion. The following strategies were adopted to alleviate this:

- integrated and coordinated land use and transportation planning;
- increasing the capacity of Singapore's roads;
- improving the public transport system;
- effective travel demand management;
- effective traffic management.

With 12% of the city state's land used for road-related facilities (Land Transport Authority, 2000), technically advanced road designs such as three-tier road junctions, and even underground roads, are being explored. The Land Transport Authority's 5 year rolling plan for road development provides the framework for all planning considerations. The Road Master Plan of Singapore is a guide to the long-term development of road infrastructure in Singapore. It identifies strategic linkages between major centers, and aims to build a more comprehensive road network. The revenue collected from motor-vehicle-related taxes and charges since 1975 has exceeded the cost of infrastructure investment in the transport sector. Singapore entrusts the operation of transport infrastructure facilities such as a port or an urban railway to the private sector, but the responsibility of bearing the costs of construction lies with the government.

The vehicle quota system (VQS) introduced in 1990 is one of two tools used by the government to manage demand for road usage. The VQS complements usage restraint measures such as electronic road pricing. Balancing ownership and usage restraint measures is a priority, with initiatives to shift to more usage-based measures (Government of Singapore, 1999).

Singapore has an efficient public transport system. The main suppliers are the Delgro Corporation, which provides the SBS transit bus and rail services; Citycab Taxi services; the SMRT Corporation, which provides bus, rail, and taxi services; and the Comfort Group, which provides taxi, bus, and courier services. TIBS is a wholly owned subsidiary of SMRT, following the merger between the operators. Its core business interests include the operation of the north–south and east–west MRT lines, totaling 50 MRT stations, and the Bukit Panjang Light Rail Transit Line.

The Public Transport Council, established in 1987 as a statutory board of the Ministry of Transport, approves bus routes and public transport fares, and regulates bus service standards. Similarly, the Land Transport Authority, founded in 1995, is a statutory board, and represents the amalgamation of the Registry of Vehicles, the Mass Rapid Transit Corporation, the Roads and Transportation Division of the Public Works Department, and the Land Transport Division of the then Ministry of Communications. Besides transport policy planning, the authority is involved in expanding the road network, managing traffic, vehicle registration and licensing, and developing the rail infrastructure.

The rapid transit system network, based on the Rapid Transit System Master Plan, is "a hierarchy of strategic network and a series of local network" (Land Transport Authority, 1999). The strategic network provides radial links to the city, while the orbital lines provide connectivity to the radial points. The completion of three new lines will more than double the total length of the rapid transit network, to 240 km (Land Transport Authority, 2000). Rail transport is seen as the main public transport mode, with buses providing feed. The vision for the public transport system remains "to provide convenience, reliability, ease of use, comfort, affordability, and competitive travel time." The financial component of the transport policy is vested on the concept of partnership whereby, "the government provides the infrastructure, commuters pay for the operating cost, while the operators extract efficiency dividends derived from the service standards and fare structures approved by the PTC" (Land Transport Authority 2000).

A new licensing framework for taxi companies paved the way for the further deregulation of the industry. Any person wishing to run a taxi business must apply for an operator license. Applicants are assessed based on a comprehensive set of criteria that include financial resources and plans for providing satisfactory service. Applicants are required to operate a minimum fleet size of 400 taxis, and are given 4 years to build up to this minimum fleet size to ensure that new operators have sufficient economies of scale to provide a good service.

3. Conclusion

The transportation problems of the ASEAN states are varied, making the creation of a unified policy difficult. Projections suggest, however, that transport demand will continue to grow. The recent tsunami has added to the problems of many countries in the region, such as Thailand and Indonesia, not only because of the immediate disruption and damage caused but also because of the inevitable diversion of resources that full recovery will entail. Many of the members of ASEAN suffered severe damage to their coastal transportation systems and to the urban infrastructure of many of their coastal cities. and much remains to be done to restore the basic infrastructure networks for inter-urban and urban transport. Many rural areas lacked basic year-round access to transport infrastructure and services prior to the tidal wave; and this is in the context of the cost of transport services that were already high because of inefficiencies in infrastructure provision and perfunctory efforts at establishing competitive transport markets. International aid has been forthcoming to meet immediate needs, but longer-term assistance is needed to return some parts of the ASEAN region to its pre-2005 state. There will be a need, however, to make more efficient use of both the aid that is forthcoming and local resources than has sometimes been the case in the past in some countries.

References

Asian Development Bank (2002) *Infrastructure*. Manila: Asian Development Bank.

Bayan J.M. (2001) *Cost characteristics of bus and jeepney transport systems in Metro Manila*. Manila: UP-National Center for Transportation Studies.

Bowen, J. (2000) "Airline hubs in Southeast Asia: national economic development and nodal accessibility," *Journal of Transport Geography*, 8:25–41.

Bunyasiri, S. (1994) *Thailand's transport: a key to sustainable growth and competitiveness*. Bangkok: Office of the National Economic and Social Development Board.

Chin, A.T.H. (1997) "Implications of liberalization on airport infrastructure development in the Asia–Pacific," in: *Conference of the Air Transport Research Group of the World Conference on Transportation Research Society*. Antwerp.

Chin, A.T.H and J. Tongzon (1998) "Maintaining Singapore as a major shipping and air transport hub," in: T.M. Heng and T.K. Yam, eds, *Competitiveness of the Singapore economy: a strategic perspective*. Singapore: Singapore University Press.

Government of Malaysia (1999) *Mid-term review of the Seventh Malaysia Plan, 1996–2000*. Kuala Lumpur: Government of Malaysia.

Government of Malaysia (2001) *Eighth Malaysia Plan (2001–2005)*. Kuala Lumpur: Government of Malaysia.

Government of Singapore (1999) *Report of the Vehicle Quota System Review Committee*. Singapore: Singapore National Printers.

Hook W. (1998) *The transport crisis in Asia*. New York: Institute for Transportation and Development Policy.

Maritime and Port Authority (2000) *New Internet-based maritime portal to be developed for the global maritime industry*. Singapore: Maritime and Port Authority.

Parsons International (2000) *Bangkok air quality management project*, Preliminary report. Washington, DC: World Bank/BMA.

Sayeg, P. (2002) *Successful conversion to unleaded gasoline in Thailand*, Technical paper 410. Washington, DC: World Bank.

Singapore Land Transport Authority (1996) *A world class land transport system*. Singapore: Land Transport Authority.

Singapore Land Transport Authority (2000) *Annual report*. Singapore: Land Transport Authority.

Singapore Ministry of Trade and Industry (1998) *Report of the Committee on Singapore's Competitiveness*. Singapore: Singapore Government.
Srisurapanon, V. and C. Wanichapun (2001) *Environmental policies in Thailand and their effects*. New York: Global Initiative on Transport Emissions (http://www.giteweb.org/iandm/viroatpaper.pdf).
UN Development Program (1997) *Lao PDR – adapting to free market principles*. New York: UN.
Urban Redevelopment Authority (1991) *Living the next lap: towards a tropical city of excellence*. Singapore: Urban Redevelopment Authority.
World Bank (1999a) *Vietnam: achievements and challenges in the transport sector*. Washington, DC: East Asia Transport Sector Unit, World Bank.
World Bank (1999b) *Urban transport sector review*. Washington, DC: World Bank.
World Bank (2000) *Road maintenance program project: Lao PDR*. Washington, DC: World Bank.
World Bank (2001a) *Private solutions for infrastructure: opportunities for the Philippines*. Washington, DC: Public Private Infrastructure Advisory Facility and the World Bank Group, World Bank.
World Bank (2001b) *Transport*. Washington, DC: World Bank.

Chapter 44

TRANSPORTATION POLICY IN THE USA

EDWARD WEINER
US Department of Transportation, Washington, DC

1. Introduction

Transportation policy in the USA has evolved over many years in response to national problems and opportunities, shifts in ideology on the role of government, and changes in the nation's transportation systems, networks, and technology (Weiner, 1999a). In 1962, President John F. Kennedy described the US approach to transportation as a "chaotic patchwork of inconsistent and often obsolete legislation and regulation." Much has been done since then to improve how the USA addresses its transportation issues.

The process of implementing effective transportation policies and programs is facilitated when there is agreement and consistency of the policy framework between the various levels of government and across agencies at the same level. Achieving such consistency requires extensive consultation and negotiation (Hazard, 1988).

2. Establishment of the US Department of Transportation

In 1966, President Lyndon B. Johnson wrote that "[t]he country's transportation facilities respond to the needs of an earlier America," and asked Congress to establish the US Department of Transportation (US DOT) to create a "coordinated transportation system" from the range of transportation programs that were developed in a variety of agencies for a variety of reasons. The highway grant programs, for example, originated as US Department of Agriculture programs to improve access for farmers to markets in urban areas. The aviation programs were started in the 1920s to nurture and ensure the safety of the new mode. Maritime programs were created to enforce revenue laws during the eighteenth century and to ensure domestic sealift and shipbuilding capacity for national defense.

Congress created the US DOT in 1966 to bring order and cohesion to transportation policies and agencies and to move toward a comprehensive and coordinated transportation decision-making process. The US Department of

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

Transportation Act of 1966 tied the creation of the department to the economic growth and stability of the nation. The US DOT was established in 1968 to:

- assure the coordinated, effective administration of the federal government's transportation programs;
- facilitate the development and improvement of coordinated transportation service, to be provided by private enterprise to the maximum extent feasible;
- encourage the cooperation of federal, state, and local governments, carriers, labor, and other interested parties toward the achievement of national transportation objectives;
- stimulate technological advances in transportation;
- provide general leadership in the identification of transportation problems;
- develop and recommend to the president and Congress for approval national transportation policies and programs to accomplish these objectives with full and appropriate consideration of the needs of the public, users, carriers, industry, labor, and the national defense.

3. The institutional and decision-making environment

There is little doubt regarding the complex nature of transportation policy-making in the USA (Weiner, 1999b). Contributing to this complexity is the wide range of government bodies at the federal, state and local levels, and private organizations and individuals involved in the decision-making process, and the wide range of needs and interests that must be accommodated.

The US institutional and decision-making environment presents a challenge to analyze, discuss, reach consensus, and implement transportation policies. This challenge is sufficiently large if only one scheme is being implemented – a road plan, a transit proposal, or a pricing scheme. However, there is widespread agreement that policies and strategies need to be implemented as a package. Individual elements are not sufficient to achieve desired results. Programs and policies need to be a mutually reinforcing, integrated packages to have the greatest effect. On one hand, such packages complicate the planning and decision-making process. But, on the other, they offer the opportunity to meet the needs of many constituencies and garner their support.

3.1. The federal government

The federal government sets the national policy framework for transportation decision-making (Table 1). To advance national goals and priorities, the federal government provides substantial funding for transportation projects. Generally, the funding comes with conditions and criteria that must be met by the recipient of the funds. The federal government also uses regulatory authority, research, and technical assistance to advance national transportation goals, particularly in the areas of safety, health, the environment, security, and the economy.

At the same time, national policies and programs are being advanced in other functional areas, including the environment, finance, housing, agriculture, commerce, communications, and energy. Often, there is a lack of consistency in policies and programs between these functional areas and transportation. The myriad of agencies pursuing policies in their areas of responsibility complicates achieving consistency across functional areas.

The US Congress, which establishes the legislative agenda for all of these agencies, seeks to achieve some compromise across the many competing interests. The legislative process is a dynamic one involving Congressional committees, federal agencies, and various stakeholders. Where objectives and needs clash, the resulting legislation may be ambiguous.

3.2. State governments

At the state level, a similar situation exists. State governments set their own transportation goals and priorities, and use similar means of funding and regulation to advance them. States also face the problem of coordinating their policies across functional areas. State legislatures, too, set the legislative agenda for state agencies.

3.3. Local agencies

At the local level there are multiple governmental bodies as well as regional organizations. Local governments compete for resources and an increased tax base while working with other agencies to agree on major projects to implement. They must respond to the requirements of the federal and state governments for funding of transportation projects as well as any regulations promulgated by these higher levels of government.

3.4. Citizens' and community groups

Increasingly, citizens themselves have become vocal, demanding a role in setting transportation policies and programs. Direct public involvement in decision-making is relatively new, and adds a new dimension to the process. Governments at all levels have passed laws and issued rules giving stakeholders and private

Table 1
Recent federal transportation legislation leading to major changes in transportation policy

Legislation		Impact on transportation policy
Deregulation		
1976	Railroad Revitalization and Regulatory Reform Act	Bipartisan legislation enacted in each of the past three decades codified a new regulatory principle that competition is the best regulator of transportation. Federal economic regulation has increasingly been reserved for major market failures or as a tool to pursue broader social purposes. Consequently, the decision-making process covering entry, exit, control, pricing, and quality of service has been relinquished by the federal government and turned over to the carriers and their customers The federal government is more stringently enforcing safety and environmental standards
1977	Air Cargo Deregulation Act	
1978	Airline Deregulation Act	
1980	Motor Carrier Act Staggers Rail Act Household Goods Transportation Act	
1982	Bus Regulatory Reform Act	
1984	Shipping Act	
1986	Surface Freight Forwarders Deregulation Act	
1994	Trucking Industry Regulatory Reform Act Federal Aviation Administration Authorization Act	
1995	Interstate Commerce Commission Termination Act	
1998	The Ocean Shipping Reform Act Maritime Security Program	
Reauthorization		
1978	Surface Transportation Assistance Act (STAA)	STAA 1978 first combined highway, transit, and safety into a single law. It increased highway funding flexibility
1982	Surface Transportation Assistance Act	STAA 1982 created the Transit Account of the Highway Trust Fund
1987	Surface Transportation and Uniform Relocation Assistance Act (STURAA)	STURAA 1987 provided 5 years of funding for the highway, transit, and safety programs with sufficient funding to complete the National System of Interstate and Defense Highways
1991	Intermodal Surface Transportation Efficiency Act (ISTEA)	ISTEA strengthened state and local planning requirements, increased public participation requirements, and provided unprecedented funding flexibility for state and local officials to tailor transportation investments to meet local needs
1995	National Highway System Designation Act (NHSDA)	NHSDA created the National Highway System and State Infrastructure Bank Pilot Program
1996	Federal Aviation Reauthorization Act (FARA)	FARA instituted FAA reform on cost accounting of federal aviation expenditures. It established airport privatization pilot programs

Table 1
Contd

Legislation		Impact on transportation policy
1998	Transportation Equity Act for the 21st Century (TEA-21)	TEA-21 affirmed key priorities: improving safety, protecting public health and the environment, and creating opportunity for all US citizens. It provided record levels of investment to continue rebuilding the US's highways and transit systems, doing so within a balanced budget. TEA-21 expanded public participation in the planning process
2000	Wendell Ford Aviation Investment and Reform Act for the 21st Century (AIR-21)	The 2000 Wendell H. Ford Aviation Investment and Reform Act for the 21st Century (AIR-21) substantially increased funding for airport development both through the Airport Improvement Program and by allowing an increase in the passenger facility charge. The Act also funded the continued redevelopment of the air traffic control infrastructure, providing the most significant change in technology and procedures in fifty years. AIR-21 provided airport infrastructure grants that can result in competitive access for new entrant carriers. AIR-21 emphasized modernization of air traffic system management

citizens a reasonable opportunity to participate in the development of transportation policies and plans. In some instances, legislation requires a formal, proactive, and inclusive public involvement process that provides ample opportunity for participation by the general public.

In the USA, public interest groups are quickly becoming sophisticated in their participation in the urban transportation planning process. They have built a nationwide communication network that provides technical assistance and forms an integrated lobbying group. And they are now developing tools and expertise to conduct their own independent analyses.

Community groups also have been formed, and have become vocal advocates for improving the sustainability and livability of local areas. These public interest and community groups do not necessarily agree with each other, adding further complexity to the decision-making process.

Even though an enhanced public involvement dimension further complicates an already complex process, in the longer term it can improve widespread acceptance of transportation policies. We have learned in the USA that public involvement cannot simply be an add-on at the end of the planning and decision-making process: it must occur from the beginning of the process when the overall goals and vision are shaped, and it must continue throughout the process. Public involvement gives participants a sense of ownership of the plans, and enlists them in promoting implementation.

3.5. *The private sector*

The private sector plays a major role in the development and implementation of transportation policies as well. Private operators provide virtually all freight and many passenger transportation services. Contracting out transportation services with private operators is also becoming more common. Business decisions on where to locate and what services are needed affect the overall policy-making process. More recently, there have been public–private ventures to build, operate, and finance transportation projects.

3.6. *International coordination*

As economic globalization becomes a more dominant fixture of decision-making, coordination with other national governments and international organizations is increasingly necessary. Establishing agreements at the international level provides a consistent policy framework for national governments. Policies and programs by other national governments increasingly affect not only the national policy arena but also state and local transportation decisions.

4. A comprehensive and coordinated decision-making process

The US federal government has promoted a regional approach to addressing transportation problems for several decades. The Federal-Aid Highway Act of 1962 asserted that urban transportation was to be addressed on a cooperative basis among governmental jurisdictions, encompass all modes of transportation, and be integrated with land development activities. It was the first piece of federal legislation to mandate metropolitan transportation planning as a condition for receiving federal funds in urbanized areas.

The metropolitan planning process provides a framework and rational process for setting goals for the areas, identifying problems, analyzing options for addressing these problems, and coming to agreement on strategies and policies to implement. It also provides a basis for the federal government to make metropolitan transportation funding decisions.

The requirements for metropolitan transportation planning have been extended and broadened in succeeding pieces of legislation. Currently, all urbanized areas with a population of 50 000 or more must have a metropolitan planning process to be eligible for federal highway and mass transit funds. In each urbanized area, a single organization is to be designated jointly by the state governor and general-purpose units of government in the region. Membership on this metropolitan planning organization (MPO) is to be by principal elected officials of general purpose units of government. The metropolitan planning process is to be carried out jointly by the MPO, the state, and publicly owned operators of mass transportation service.

Every metropolitan planning process must develop a long-range transportation plan describing policies, strategies, and changes in facilities that are proposed for the next 20 year period, including analysis of transportation system management (TSM) strategies to make more efficient use of the existing transportation system. In addition, a transportation improvement program (TIP) is required, which is a staged multiyear (3–5 year) program of transportation improvement projects that are consistent with the transportation plan. Only transportation projects in the TIP can be implemented with federal funds. The transportation plan, the TIP, and individual transportation projects must conform to the state implementation plan under the Clean Air Act.

5. Major policy issues

5.1. Passenger and freight travel growth

The last quarter of the twentieth century produced growth in most areas of transportation in the USA. By 1997, the US transportation network was supporting 7.4 trillion passenger-km of travel, an 84% increase from the 4.0 trillion passenger-km in 1975. Vehicle-kilometers of travel and aviation passenger enplanements doubled in the 25 years from 1972 to 1997. This growth is attributed in part to an expanding economy, higher consumer income, greater vehicle availability, reduced travel costs, increased population, and greater consumption of goods. Transit and intercity passenger rail traffic grew more modestly.

Freight travel has shown significant gains. In 1997, over US $7.9 trillion worth of freight traveled 6.3 trillion tonne-km. This compares with US $1.6 trillion of freight and 3.7 trillion tonne-km in 1975 (US Department of Transportation, 2001). Domestic air cargo shipments tripled. Many companies increased the efficiency of production, storage, and shipping practices. Such practices, such as the shift from warehousing to "just-in-time" delivery, place a significant burden on the reliability of the transportation network. The emergence of e-business

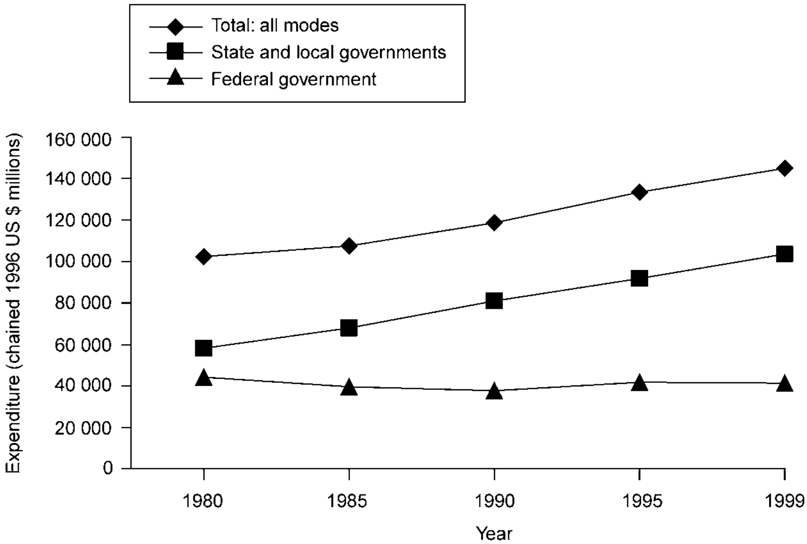


Figure 1. US federal, state and local transportation expenditure, 1980 to 1999.

also increases the pressure on transportation networks because inventories and consumers are more dispersed.

The freight industry has made great strides in improving intermodal connections, developing more efficient rail–truck–maritime networks and aviation–truck systems to accommodate different shipping needs and provide better service. Initiatives to reduce air traffic delays, reduce paperwork for motor freight carriers, streamline cross-border shipments, track freight by the Global Positioning System, and other improvements in domestic and international shipping will continue to improve the productivity of the freight transportation network.

As a consequence of the attacks of September 11, 2001, and the worldwide economic downturn, airline travel dipped in the succeeding months. Airlines suffered losses in passengers and revenue, and went through a period of unprofitability and consolidation. However, over time the industry is expected to gradually improve. In the longer term, growth is expected in aviation passenger and freight traffic.

5.2. *Funding*

Total government expenditure in the USA on transportation – air, highway, transit, waterway and marine, rail, and pipeline – has grown significantly in real terms over the last 20 years (Figure 1). From 1980 to 1999, total government expenditure for transportation, in chained 1996 US dollars, grew from

$102.3 billion in 1980 to $144.9 billion in 1999, a 42% increase (US Department of Transportation, 2001).

Funding for transportation is expected to remain a major challenge. Decisions about additional funding sources will have to address whether to place greater reliance on state, local, and private funding sources; increase use of innovative financing mechanisms such as loans, loan guarantees, and lines of credit; expand or consolidate federal transportation trust funds; or establish new user fees.

Transportation expenditure by level of government

State and local expenditure on transportation has increased at a substantially greater rate than federal expenditure. The state and local share of all government expenditure for building and maintaining the nation's transportation system rose from US $58 billion in 1980 to US $103.6 billion, an increase of 78% in 1996 chained dollars. This increase in transportation spending by state and local governments brought the federal share of all government transportation expenditure down from 43% in 1980 to 29% in 1999 (US Department of Transportation, 2001).

The federal government owns only those transportation facilities that are on federal land. States, local agencies, and private companies own all other transportation facilities and equipment. Therefore, the federal government uses its grant-making and regulatory power to promote the development of a safe and efficient transportation system.

Almost all states have established departments of transportation. States are the recipients of the bulk of federal transportation funds. Under the Intermodal Surface Transportation Efficiency Act of 1991, states are required to develop long-range multimodal transportation plans and short-range transportation improvement programs as a condition for receiving federal funds for transportation projects.

Local governments are responsible for building and maintaining the local street network and for installing and maintaining traffic control devices. Funding for these operations derives from local taxes, and fund transfers from higher levels of government.

There are also numerous agencies at the local level responsible for providing transportation services and managing elements of the transportation system. There are approximately 6000 agencies providing public transportation services. Most are small bus operations that are privately owned. The larger transit agencies are mostly public authorities. Many public transportation operators contract with private providers for some part of their service.

Modal share of federal transportation funding

All transportation modes – maritime, rail, highways, transit, and aviation – received federal support to nurture their development (Figures 2 and 3). As the mode

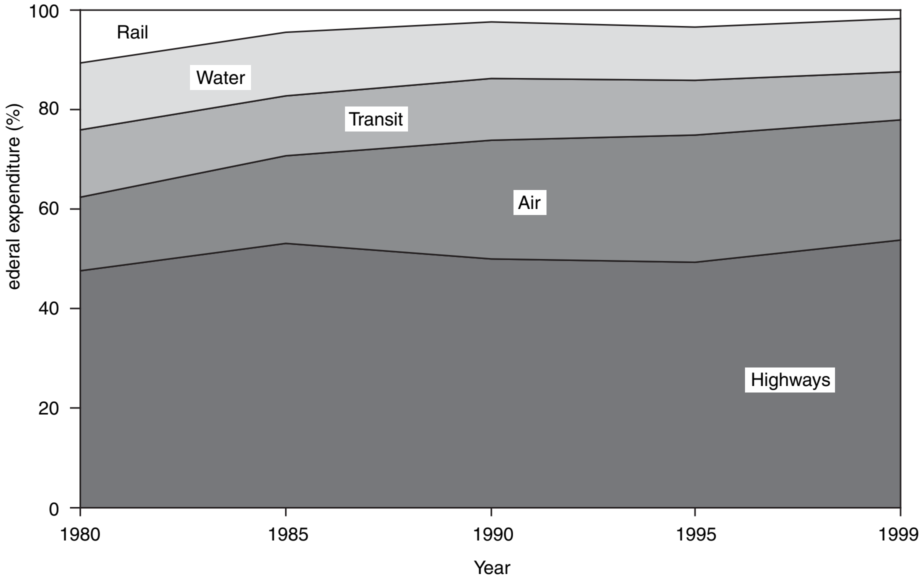


Figure 2. Modal share of US federal transportation expenditure, 1980–1999.

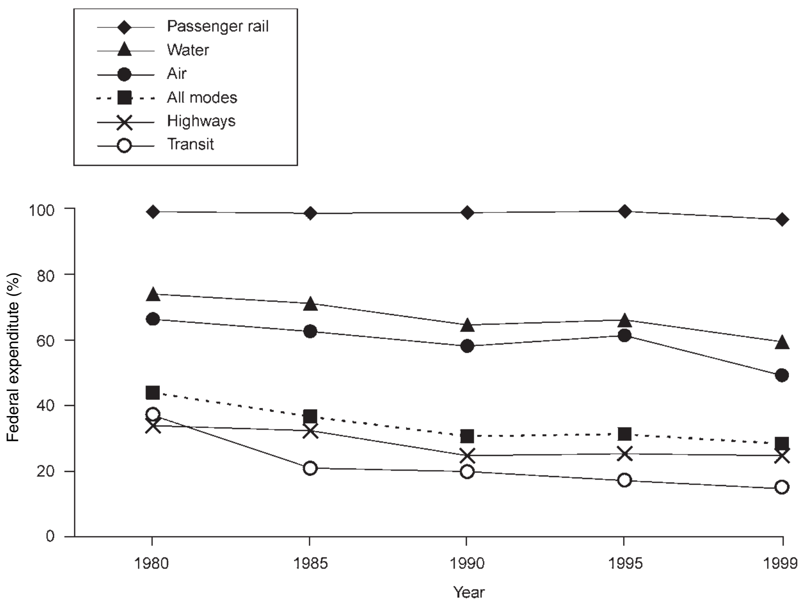


Figure 3. Federal share of US government expenditure on transportation by mode, 1980–1999.

matured, the public sector modified its support to allow the private sector's pursuit of increasing efficiencies within the modal market. As a result of this, the public sector decision-making and regulatory climate has traditionally been separated by individual mode. The public sector's evolution from incubator to facilitator has been accompanied by a broadening of decision-making processes, allowing an increasingly multimodal approach for public and private operators.

With the completion of the National Interstate and Defense Highway System provided for in the Surface Transportation Assistance Act of 1982, the legislative focus shifted to the nature and size of the post-interstate highway program (US Department of Transportation, 2000). The Intermodal Surface Transportation Equity Act of 1991 (ISTEA) and the Transportation Efficiency Act of the 21st Century (TEA-21) authorized funds for highways, mass transit, and safety programs up to 2003. In a major breakthrough, these acts created a surface transportation program with flexible funding that opened the door to new opportunities to address statewide and urban transportation problems.

The purpose of ISTEA was set forth in its statement of policy and continued in TEA-21:

> It is the policy of the United States to develop a National Intermodal Transportation System that is economically efficient and environmentally sound, provides the foundation for the Nation to compete in the global economy, and will move people and goods in an energy efficient manner.

ISTEA established a National Highway System (NHS) consisting of 250 000 km of interstate highways, urban and rural principal arterials, and other strategic highways. The NHS was funded with a 80% federal matching share. ISTEA created, and TEA-21 continued, a new block grant program, the Surface Transportation Program (STP), which makes funds available for a broad range of highway, mass transit, safety, and environmental purposes. Each state was required to set aside 10% of the funds for safety construction activities, and another 10% for transportation enhancements. The remaining 80% was to be allocated statewide. A new Congestion Mitigation and Air Quality Improvement Program was authorized for transportation projects in ozone and carbon monoxide non-attainment areas.

Transit programs were authorized as well through 2003. The Discretionary and Formula Capital Grant transit programs were re-authorized with minor changes. The Discretionary Grant program allows state and local agencies to compete for federal funds for new fixed guideway projects. The federal matching share was increased from 75 to 80%. The Formula Grant program provides transit funds to state and local agencies on a formula for routine transit and transit-related expenditure. Federal expenditure for transportation has increased steadily since 1980, from US $24.7 billion in 1980 to US $44 billion in 1999. During this period, the share of federal expenditure for highways increased from 47 to 54%. Federal expenditure on aviation also increased from 15% of transportations expenditure in 1980 to 24% in 1999. The largest decrease in modal share of federal expenditures was for railroads, declining from 10 to 1%.

The results of transportation expenditure have been significant. The interstate highway system has essentially been completed, numerous transit systems rebuilt and modernized, ports and inland waterways deepened, airport capacity expanded, and rail rights-of-way modernized. But, many transportation facilities are reaching the end of their economic lives and will need to be rehabilitated or replaced, often at a significant increase over their original cost.

5.3. *Safety*

Over the next 25 years, steady growth in travel in the USA will lead to heightened concerns for transportation safety and the need to exert even greater efforts to improve the safety of facilities and vehicles, and promote safer behavior of users. Even though the rate of transportation deaths and injuries have been decreasing in the USA, the number of transportation-related deaths is still over 40 000 per year. This is an unacceptable level of loss that will require a significant, multifaceted effort to reduce.

The aging US population will increase the risk of injury to elderly drivers and pedestrians. Still other safety issues will arise from the growth of air transportation and its increasing use for international travel, stressing an already burdened system. These issues will require greater attention to safety-enhancing technologies and other measures both domestically and abroad. Safety concerns will also require important decisions about traffic law enforcement, vehicle inspections, and driver licensing.

5.4. *Energy needs*

The transportation sector in the USA accounts for two-thirds of the total oil used by the nation. Petroleum use is ubiquitous in transportation, because it has been readily available; easily extracted, shipped, and transformed; has excellent combustion properties; and is low cost. In 1975, the energy costs of driving averaged US \$0.08 per kilometer in the USA. In the intervening years, the drop in fuel costs coupled with dramatic changes in fuel economy and the way petroleum is used lowered that cost to just over US \$0.03 per kilometer. However, even if pump prices reach US \$2.00 a gallon (US \$0.53 a liter), the real fuel costs would be less per kilometer than in 1975, about US \$0.05 per kilometer.

It is possible that the real price of petroleum will rise over the next 25 years compared with the extremely low prices of the 1990s. Market responses are expected to result in more efficient methods to process petroleum as well as incentives to find alternative raw materials. One challenge before the transportation policy-makers is to use the current period of abundance to explore other fuel sources and actions that could be taken to ease any transition to different fuels. Progress is being made in increasing the fuel efficiency of vehicles, and in the development of alternative sources of power – particularly hydrogen.

5.5. *Environmental quality*

In recent years, concern about energy resources and the environment has grown across the USA and around the world. These concerns are likely to increase in importance and continue to have a major impact on transportation decisions. As the population grows and economic development continues, greater efforts will be needed to preserve and enhance the natural and human environment.

Transportation contributes to increases in greenhouse gases that affect global warming. Although there is still uncertainty regarding the magnitude and distribution of both positive and negative effects, there is little doubt that human activity, including fossil fuel use for transportation, is contributing to increased atmospheric concentrations of greenhouse gases. In the USA, transportation accounts for about one-third of carbon dioxide emissions, or about 26% of total greenhouse gas emissions.

The Clean Air Act Amendments of 1990, passed after extensive debate in Congress, address the attainment and maintenance of National Ambient Air Quality Standards (NAAQS). These standards set the maximum allowable levels of various pollutants in the air. Non-attainment areas are classified for ozone, carbon monoxide, and particulate matter in accordance with the severity of the air pollution problem. Depending upon the degree to which an area exceeds the NAAQS, that area is required to implement various control programs and to achieve attainment of the NAAQS within a specified period of time. The areas that are furthest out of compliance were given the longest length of time to achieve the standards.

The Clean Air Act Amendments require states to develop State Implementation Plans (SIPs) detailing how they will meet the NAAQS within the legislated deadlines. As part of the SIPs, states are required to reduce volatile organic compounds emissions by 15% from 1990 baseline emissions over the 6 years following enactment. In addition to the 15% reduction, emissions arising from growth in vehicle-kilometers of travel have to be offset by other reductions in emissions. In more severe non-attainment areas, transportation control measures need to be implemented to reduce the use of vehicles by shifting travel to alternative modes such as transit and ridesharing, or shortening the trip length, or using other approaches including telecommuting.

The "conformity" provisions in the 1990 Act were expanded from the Clean Air Act Amendments of 1977. A conformity determination is made to assure that federally assisted projects or actions conform to a SIP. No project can cause or contribute to new NAAQS violations, nor increase the frequency or severity of any existing violations of any standard, nor delay the timely attainment of any required NAAQS. The process recognizes that transportation-related air quality issues have to be analyzed on a system-wide basis and be controlled through regional strategies to be effective. The Clean Air Act Amendments of 1990 expand the "sanctions" where states fail to carry out requirements of the Act including withholding of federal funding for highway projects.

Although transportation is a significant source of air pollution emissions, the net effect of aggressive transportation emission regulations has been a dramatic reduction in emissions and a measurable improvement in air quality in the nation. However, the transportation sector still faces a significant challenge to reduce emissions. Future improvements in air quality will require decisions about new regulations, the application of new technologies, and changes in travel choices. One strategy would combine increasing the use of alternative fuels, while improving vehicle efficiency, and fostering market-based emissions trading.

5.6. Land development

It is unlikely that much progress will be made in reducing congestion and air pollution without facing the land development element of the picture. It is the dispersed pattern of residences, businesses, and other activities in the USA that makes the automobile so essential for travel in that country. To make group modes of transportation such as bus and rail transit more workable, densities would need to be higher and developments designed more efficiently. This would require a degree of land use planning and control that has not been widely practiced to date, as well as changes in those policies that have supported dispersed, low density development.

Many forces have acted to bring about these development patterns including changes in the demographic diversity, including rising immigration; in the economic base of the nation and the shift to a service-based economy; in technology, especially communications technology; in life-styles, such as the increase of two-worker and single-parent households; and the trend to more flexible and flatter organizational arrangements that is reflected in the restructuring of many businesses in recent years. There needs to be a better understanding of these forces and those that will affect development patterns in the future. There also needs to be an understanding of the economic, social, and environmental implications of these forces.

5.7. Social equity

In the future, US policy-makers will be required to devote more attention to assuring that the environmental impacts of transportation do not fall disproportionately on disadvantaged communities and that all segments of society receive the benefits of transportation. Executive Order 12898, entitled "Federal Actions to Address Environmental Justice in Minority Populations and Low-Income Populations," focuses attention on the environmental and human health conditions in minority and low-income communities to ensure that all federal programs and activities do not use criteria, methods, or practices that discriminate on the basis of race, color, or national origin.

The Executive Order requires that the environmental impact process under the National Environmental Policy Act address environmental justice issues. Federal actions and projects must be analyzed to identify the human health, economic, and social effects on minority communities and low-income communities. Mitigation measures must be taken to address significant adverse environmental effects on minority and low-income communities. The affected communities must have opportunities to provide input in the identification of impacts and mitigation measures.

6. Conclusion

Transportation policy-makers in the USA will face a changing context driven by economic, demographic, and technological change. Transportation systems will need to be improved within the constraints of limited budgets and environmental impacts. A wide range of institutions have been created at the federal, state, and local level to address these issues, but these institutions will need to evolve as well to accommodate expanding issues and new participants in the policy-making process.

References

Hazard, J.L. (1988) *Managing national transportation policy*. Westport: Eno Foundation for Transportation.

Weiner, E. (1999a) *Urban transportation planning in the US – an historical overview*, revised and expanded edn. New York: Praeger.

Weiner, E. (1999b) "National studies of the US transportation system," *Transportation Quarterly*, 50:11–20.

US Department of Transportation (2000) *Transportation decision-making – policy architecture for the 21st century*. Washington, DC: US Department of Transportation.

US Department of Transportation (2001) *Government transportation financial statistics 2001*. Washington, DC: Bureau of Transportation Statistics, Department of Transportation.

This Page Intentionally Left Blank

Chapter 45

TRANSPORTATION POLICY IN CANADA

TREVOR D. HEAVER and WILLIAM G. WATERS II
The University of British Columbia, Vancouver

1. Introduction

Transport policy issues in Canada have many features in common with other countries but also reflect differences arising from its geographic size, patterns of economic development, and its federal political structure. Canadian national transport policy, the focus of this chapter, deals with inter-provincial and international transport of freight and passengers, and has traditionally been little involved with urban transport and the automobile. The latter is in contrast to transport policy concerns for developed countries with high population densities.

Historically, transport policy orientation was to unify the country and promote a national economy, whereas over time the importance of transport to facilitate economic activity and trade became more important than political unity objectives. The result has been a shift in the broad policy from one in which transport was used as an instrument to serve national policy, an approach that resulted in the promotion and protection of carriers, to an approach that largely separates transport policy from other national policy concerns. The interests of shippers and travelers in the efficiency of transport services are now given paramount attention. This shift has been associated not only with seeing the individual modes as a part of the transport system but, more recently, as a part of the wider national logistics structure essential for national competitiveness in global supply chains. The shift has resulted in a general, although not universal, recognition that transport cannot be used as an instrument of economic development separate from the logistics and other economic considerations of the industries transport is expected to serve.

The shift in policy has taken place as the country has grown and matured economically and politically. The financial relationship of the government to the industry was once as a major source of capital and subsidy. Now, there is growing concern that the government sees transport as a "cash cow," an industry to be taxed. However, within this relationship is the dichotomy between the rural and remote locations where transport services may not be viable and the areas of high density where they can prosper, be self-sustaining and even be taxed.

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

Canada's statement of national transport policy has changed with the new circumstances of the transport industry and the economy. In part, the changes have been in response to the evolving potential for reliance on competition in transport markets and, inevitably, in part they have been in response to specific concerns of interest groups. In what follows we attempt to identify both economic and social/political forces that affected the evolution of Canada's national transport policy.

As current transport policy is closely linked to the economic and political geography and to the history of a country, it is appropriate to provide a short review of the geographic and historic setting of transport policy in Canada.

2. The geographical, historical, and jurisdictional setting

Canada is the second largest country in the world (after Russia). It extends 5514 km east to west, and 4634 km south to north. Its population in 2004 was only 32.5 million (about one-tenth that of the USA) although approximately 85% of the population lives within a 300 km band adjacent to the border with the USA.

Joining the scattered concentrations of population by an all-Canadian route was a rationale for the support of railways during the formation of the country, and has remained a consideration, as in the federal support for the Trans-Canada Highway. The preservation of a level of transport capability separate from that of the USA remains a factor in Canadian policy. However, in view of the considerable interdependency between the two economies and transport systems, and the maturity of both, separating out the influence of a desire for national independence has become harder.

The Canadian federation consists of ten provinces and three territories. Under the Canadian constitution, responsibilities for transport are divided between the federal and provincial governments. With the exception of air transport and marine navigation, for which the federal government is solely responsible, the provinces are responsible for local works and undertakings. The latter includes roads, intra-provincial trucking and bus services, and intra-provincial railways, unless declared to be works to the general advantage of Canada, a provision that has been rarely used.

This division of responsibilities means that the largest part of Canadian society's investments in transport, which is in roads, is under provincial jurisdiction. It also means that issues of intergovernmental policy harmonization apply in many aspects of transport, internally as well as internationally. This is a significant complicating factor in Canadian transport policy.

The large size of the country, the small size of the economy, and its proximity to the USA account for the long history of public roles in the development and support of transport undertakings. They include support for the construction of the private Canadian Pacific Railway (CPR) in 1881, the formation of Canadian National Railways (CN) as a Crown corporation in 1922, the creation of Trans-Canada Airlines (now Air Canada) in 1937, and the assumption of federal responsibility for major ports (National Harbours Board, 1936) and airports (in the 1950s and 1960s). The federal (as well as provincial) government has also been actively involved in the regulation and subsidization of transport.

The dominance of long-distance transport by the railways led to active regulation of that mode to prevent abuse of its market power. The long distances and importance of resource products in Canada resulted in the dominance of rail issues in transport policy. Railways were to Canada what shipping was to Australia.

The first major shift in the policy of the government toward transport was caused by the successful growth of modes able to compete with the railways. During the 1950s, the loss of rail traffic and the effect of competition from truck and, where available, water transport gave rise to financial crises for the railways and pressures on the regulatory structure. The substantial development in transport policy arose from the recommendations of the MacPherson Royal Commission on Transportation (1961–1962). Notwithstanding the commission's name, it was appointed to look into the problems associated with rising railway rates, but it had the insight and vision to adopt a broader mandate concerning transport in Canada. Unlike commissions that had preceded it, this commission showed the independence to interpret the challenge before it widely and undertake a broad investigation of transport in Canada. The resulting report laid the foundation for modern Canadian transport policy.

3. The MacPherson Commission and the National Transportation Act 1967

The MacPherson Commission (1961–1962) saw the problems of the railways as symptomatic of deeper changes that were underway in the economy and transport industry. It saw that continuation of tightly regulated rail rates and subsidies from the government were inappropriate. The report was one of the first public documents in the world to grasp the significance of changing technology and rise of competition in transportation. The report distinguished "national policy," which focused on broad economic and social concerns, from "national transport policy," whose mandate they recommended should be concerned only with facilitating an efficient transportation system. Further, with the rise of competition, detailed regulation of transportation was unnecessary and interfered with the functioning of markets. However, in keeping with conditions of the time, the commission did not contemplate competition between railways, although it argued that intra-modal competition was already possible in the trucking industry. The commission called for greatly reduced regulation of railway markets, and for commercial forces rather than government regulations to guide transportation.

The commission also saw regulation as only one form of government intervention that affected the functioning of markets to the detriment of the industry. The commission argued that in its operational and promotional roles in transport, the government should avoid distorting markets by providing services at less than cost to certain modes, and/or by requiring carriers to provide unremunerative services in the public interest without compensation.

Many of the concepts of the MacPherson Commission were embodied in the statement of national transportation policy in section 3 of the 1967 National Transportation Act (NTA 1967). However, the effects of the commission were confined to the shift in the policy statement and to major revisions to the regulation of railway rates. Reducing regulation of other modes had to wait for two decades. The statement of national transportation policy reads in part:

> It is hereby declared that an economic, efficient and adequate transportation system, making the best use of all available modes of transportation at the lowest total cost is essential to protect the interests of the users of transportation and to maintain the economic well-being and growth of Canada, and that these objectives are most likely to be met when all modes of transport are able to compete ...
>
> (a) regulation of all modes of transport will not be of such a nature as to restrict the ability of any mode of transport to compete freely with any other modes of transport;
>
> (b) each mode of transport, so far as practical, bears a fair proportion of the real costs of the resources, facilities and services provided that mode at public expense;
>
> (c) each mode of transport, so far as practical, receives compensation for the resources, facilities and services that it is required to provide as an imposed public duty;

Subsequent paragraphs prescribed that rates and conditions for the carriage of traffic should not constitute an "unfair disadvantage" or "undue obstacle" to the movement of traffic, the development of industries, export trade, or movements of traffic through Canadian ports. Over the years, legislation has modified the wording and introduced qualifying statements to make the directive less focused, but the thrust of the statement has provided the general policy direction since 1967.

The NTA 1967 established an integrated regulatory agency to oversee all the transport modes. However, the hoped-for rationalization of regulatory attitudes across the modes was not achieved. The deregulatory push was primarily for railways. Railway rates were subject to regulated minimum and maximum levels, the former set at costs and the latter set deliberately at such a high level that applications would be very rare. Market forces were expected to set rates, subject to the opportunity for shippers to appeal rates and service conditions under a broad "public interest" provision, the working of which became controversial. Over the next decade or so, the rail industry changed from being dominated by tedious regulatory disputes to one guided by commercial principles (Heaver and Nelson, 1977). Rail regulatory hearings became a rare event. However, the process was not always smooth. There were – and continue to be – residual regulatory issues. It would be close to 20 years before deregulation was extended to other modes. The full implications of relying on market forces, competition, and an emphasis on economic efficiency evolved gradually.

4. Developments in transport policy to 2000

While the legislation of 1967 introduced a new policy with clear principles, it could not ensure that the principles would be implemented. Some glaring exceptions to the legislation remained. Over time, these exceptions have had to be faced as the levels of competition within the transport industry and faced by shippers have heightened and as the financial constraints on governments tightened. The result was a remarkable decline in federal involvement in transport by the end of the twentieth century.

Table 1 presents a concise chronology of major transport policy developments in Canada since 1960. It is a selected list of major policy reviews and legislative implications. Rather than elaborate chronologically, this chapter focus on a number of themes that characterize policy reform in Canada. These four themes are shared with many other countries, and are broad but related:

(1) regulatory reform (deregulation) and the evolution of residual regulatory roles of the government;
(2) reductions in subsidies for unremunerative services;
(3) a changing/declining federal role in the provision of infrastructure and an increased emphasis on cost recovery;
(4) the role of Crown corporations (public enterprises) and privatization.

After discussing these themes, this chapter subsequently identifies policy themes and directions likely to be important in the early twenty-first century.

4.1. The progress of regulatory reform (deregulation)

The dominant theme in transport policy reform is deregulation, a policy direction shared with most other developed countries. Canada was one of the first countries in the world to embark on deregulation, applied first in the rail industry.

The railway regulatory reform of 1967 quieted concerns about railway rates and regulation during a period of growth in the carriage by rail of bulk commodities, mainly for export. While there was dissatisfaction among some shippers with the limited grounds for regulatory appeal, the overall satisfaction of shippers was evident in the opposition of shippers and carriers to the prospect of a more interventionist position of government proposed in 1975 (the important exception

Table 1
Selected Canadian transport policy events since 1960

Policy reviews and/or legislation	Comment
MacPherson Royal Commission on Transportation, 1959–1961	Set up to review rail issues. Recommended a shift to relying on markets and competition and reduced government involvement to achieve an efficient transport system
National Transport Act (NTA) 1967	Proclaimed a transport policy separate from national policy, and adopted many of the recommendations of the MacPherson Commission
Federal policy review, *c*. 1975–1976	Motivated by political desires for intervention in transport to achieve national policy goals; some shipper interest in greater regulatory intervention. Bill C-33 not passed; the vast majority of shippers and carriers opposed more government intervention
Via Rail Canada Act 1978	Crown corporation Via Rail took over rail passenger services
Commissions on grain transport	Separate commissions reviewed the costs and financial losses on statutory grain rates (Snavely Commission, 1975), and reviewed problems, potential solutions, and subsidy mechanisms in grain transport (Hall Commission, 1975–1977; Gilson Commission, 1982)
Western Grain Transportation Act (WGTA) 1983	The WGTA capped the volume of grain to be moved under statutory rates, and offered a subsidy equal to current losses, but subsequent increases in costs were reflected in higher rates. The subsidy is paid directly to railways, with various provisions for review and assessment of performance
White Paper, *Freedom to Move: a Framework for Transportation Reform*, 1985	A government-led review sought to promote competition and deregulation, to rely more on markets and commercial forces to guide transport; heavily influenced by the deregulation impacts in the USA and elsewhere. It led to changes in all modes
National Transportation Act (NTA) 1987 Motor Vehicle Transportation Act (MVTA) 1987	Enacted the principles in the White Paper *Freedom to Move*. Emphasized intramodal competition in all modes. It strengthened negotiating options for shippers, including extended interswitching, competitive line rates, and final offer arbitration. The companion MVTA called for collaboration with the provinces to foster deregulation and harmonization of trucking

Table 1
Contd

Policy reviews and/or legislation	Comment
Shipping Conferences Exemption Act 1987. Originally 1970, also amended 1979 and 2001	Exempted certain practices of conferences from the Competition Act, allowing agreements on pricing and sharing capacity. The Act allowed individual lines to take independent rate action (1987) and to enter into confidential contracts (2001)
Privatization	Air Canada, 1988; Canadian National Railways 1995; Nav Canada, 1996
Royal Commission on National Passenger Transportation (RCNPT), 1989–1993	Prompted by issues with Via Rail, but had a broader mandate. Recommended greater reliance on market forces, with an emphasis on cost recovery and even commercialization of infrastructure provision. Made estimates of full social and environmental costs of modes, and recommended moving toward full-cost recovery, including environmental costs
National Transportation Act Review Commission (NTARC), 1993	A statutory review required 5 years after the NTA 1987. Various recommendations to modify the perceived balance of powers between shippers and carriers in negotiation. Recommended examining the feasibility of separating rail track from operations; transport mergers to be under the Competition Act
Airport, port and seaway devolution	1992, four airports to local authority status; 1996, others to Canadian Airport Authority status with modified arrangements. Major devolution for ports and the St Lawrence Seaway in the Canada Marine Act 1998
Canada Transportation Act (CTA) 1996	Limited legislative changes regarding rail shipper regulatory disputes; easier to abandon or transfer track. No longer a north/south distinction in air transport: all domestic air deregulated. No regulation of northern water carriage. Mergers and acquisitions to be under the Competition Act
Transportation Climate Change Table, 1999	Reviewed potential policy actions, and their costs and effectiveness in achieving the reductions in greenhouse gases to satisfy the Kyoto protocol
Canada Transportation Act Review (CTAR) Panel, 2000–2001	Recommended minor changes to regulatory frameworks, possibly open rail access (with compensation) in extreme cases. Reiterated the basic policy principles of the MacPherson Commission, and recommended greater devolution of infrastructure provision including roads, and recognition of environmental costs in transport pricing

Table 1
Contd

Policy reviews and/or legislation	Comment
White Paper, *Straight Ahead: a Vision for Transportation in Canada* 2003	Outlined government transport policy directions largely consistent with CTAR, including the importance of research on full costs, incorporating environmental costs
Proposed revisions to Canada Transportation Act in 2003	Draft legislation reflecting many CTAR recommendations

was continued regulation of grain transport, discussed separately below). During the decade 1975–1985, remarkable changes took place in attitudes and conditions.

Three forces accounted for the thrust for greater reliance on competition in transport markets. First, there was a shift in Canada as elsewhere to a belief in greater reliance on market forces, marked by the election of conservative governments during the 1980s. Second, deregulation of transport in the USA had direct implications for Canada as air travelers and some shippers were able to make greater use of more attractive deregulated services in the USA to the disadvantage of Canadian carriers. Third, the experience of Canadian shippers in negotiating rates with US railways gave them confidence in their ability to negotiate rates and services effectively even with the large railways. The NTA 1967 had provided greater freedom for Canadian shippers and railways to negotiate rates and services than enjoyed by their counterparts in the USA. However, the allowance of confidential rates in the USA by that country's Staggers Act of 1980 leapfrogged the role of commercial negotiations in the USA beyond that in Canada.

The changes were reflected in a major transport policy paper, *Freedom to Move: a Framework for Transportation Reform* (Transport Canada, 1985). It was the result of an internal government review, not the result of a commission or other major public review process. The proposals were to open up and promote competition in the transport system, reduce the burdens of regulation, and increase negotiating tools for shippers who had limited competitive alternatives. The proposals were the basis for the National Transportation Act 1987. This was the first legislation in Canada to address the regulation of rail, highway, and air transport in a single statute. The Act did not deal with ocean shipping or surface passenger transport. Liner shipping is subject to the Shipping Conferences Exemption Act 1970. Revisions to this Act in 1987 were in line with changes made in the USA in 1984 that required the shipping cartels to give their members the right of independent rate action, thereby enhancing competition.

Looking first at rail regulation, reliance on competition went beyond that in the USA by promoting competition among railways. This was done through two provisions creating access to alternative railways for a shipper located on only one line. Since 1918, railways had been required to move traffic to/from shippers located on the line of another railway when the shipper was located within 3 miles (4.8 km) of a railway interchange point ("interswitching"). The NTA 1987 extended the distance to 50 km. It also introduced "competitive line rates" (CLRs) which would apply to shippers on a single line beyond the reach of interswitching. Under the CLR provision a shipper could use a rate obtained from a railway at an interchange (a place of potential competition on a route) to lead to the creation of a rate on the line of the originating or terminating carrier where a captive shipper is located. For various reasons, there has been virtually no use of CLRs, particularly due to the reluctance of the dominant railways to poach each other's traffic, although arguably the existence of CLR provisions brings some disciplinary pressures on railways in rate negotiations.

The NTA 1987 also introduced final offer arbitration (FAO) as a way to encourage reliance on "commercial" processes rather than regulatory intervention.

A *quid pro quo* in the Act was greater freedom for the railways to abandon branch lines (discussed further below), and these provisions were further extended in the Canada Transportation Act 1996. The changes, especially in 1996, enabled the railways to rationalize their system more effectively and become more efficient. As was the experience in the USA, this also resulted in the formation of more "short line" railways.

The experience of US deregulation of airlines had an even greater impact on Canadian transport policy. Canadians could see and experience the growth of competitive services and lower fares available in the USA. There were deregulatory steps in Canadian air transport even before passage of the NTA 1987. The NTA 1987 freed entry of Canadian firms, although carriers were required to give advance notice of withdrawal from a route on which they were the sole operator. This was to allow for orderly replacement by another operator. Initially, air regulation persisted for "northern" air services (but with a "fit, willing, and able" test rather than "public convenience and necessity"), but full deregulation was extended to the north in 1996. Like most countries, Canada has reserved the domestic air market for domestic-owned carriers.

Under deregulation, the air industry underwent consolidation, as it had in the USA. With Canada's small domestic market, the sustainability of a domestic competitive air market is debatable. Consolidation left Air Canada in a market-dominant position, although there has been both bankruptcy and expansion of low-cost air carriers. At the time of writing, Air Canada is suffering a loss of domestic market share and is under bankruptcy protection. The Canada Transportation Act Review (CTAR) Panel (2001) recommended that Canada pursue an "open skies" market with the USA, beyond the liberal cross-border flights already between the two nations. However, the prospects for this being implemented are uncertain.

The NTA 1987 had companion legislation, the Motor Vehicle Transportation Act (MVTA), to encourage deregulation of trucking. The MVTA did not impose the federal authority on inter-provincial trucking but called for provincial collaboration to encourage competition and deregulation. Progress was slow, but deregulation gradually came about across the country, concurrent with efforts for greater coordination and cooperation on matters of provincial jurisdiction such as hours of service and road safety issues. Conflicts across the government jurisdictions have made progress slow.

The practices of liner shipping companies that are members of conferences (cartels) are subject to the Shipping Conferences Exemption Act. First passed in 1970, the Act exempts conference lines from the Competition Act for certain practices, such as the exchange of pricing information subject to some conditions. Prior to the 1980s, the majority of Canadian shippers supported the conference system. Since then, the majority no longer support the exemption. Canada has continued to allow conferences, mainly to be consistent with its major trading partners. Revisions in 1987 and in 2001 have enabled more competition among members by requiring in 1987 that conferences give members the right of independent rate action and, in 2001, that individual confidential contracts are allowed. Two significant differences exist between Canadian and US law that result in more competitive environment in Canada: there is no policing of rates below those published, and non-conference lines are unregulated.

The most recent public policy review was the CTAR Panel (2001). It reviewed the functioning and issues of the CTA 1996 and various other legislation and issues. The question of balance between shippers and carriers in a regulatory process was reviewed. Some parties were advocating some type of open access regime to enhance rail competition. The CTAR Panel did not agree with it as a general policy, although its recommendations suggested the possibility of granting access as an ultimate threat to a carrier deemed to be unresponsive to shippers' complaints. Draft legislation in 2003 did not include this component but included other recommended measures to deal with residual rail regulatory issues. Regarding air transport, despite some concerns about possible market dominance, the CTAR Panel concluded that it is impossible to efficiently regulate airfares. Reliance on market forces is necessary even if the markets are not perfect.

4.2. The subsidization of unremunerative services

A common problem in transport policy is the financing of infrastructure or services where the market size is too small to support viable services but they are deemed to be of social/political importance. The NTA 1967 called for subsidies for unremunerative services, but the implementation of the principles took many years. The most significant cases are reviewed briefly. The first situations described deal with rail services, as these were the first focus of government attention.

Grain carried at statutory rates

After 1897, grain moving from the prairies to the west coast for export and moving to the Great Lakes was required (except for years affected by the First World War) to move under rate levels by order of Parliament at one-half cent less per ton-mile than those existing in 1897. It is not surprising that, in spite of technological changes, by 1960 the railways lost money on this traffic. This was the largest single traffic type handled by the railways, but the railways received no direct subsidy for their losses. It took a series of commissions of inquiry before prairie residents would concede that the railways lost money on grain. Even then, there was considerable opposition by farm interests to the provision of explicit subsidies to the railways. Grain and other bulk commodities boomed in the 1970s, and cash-strapped railways could not afford to invest in needed capacity. Pressures from other bulk shippers led to reform. The Western Grain Transportation Act of 1983 changed the rate levels and called for subsidies to continue the historical benefit to prairie producers. These subsidies persisted at increasing levels until swept away in 1995 under the severe financial constraints of the federal government. It had taken a generation to bring about needed reforms. To this day, there are various regulatory restrictions that apply to grain transport that do not apply in other rail markets. While these restrictions still affect the rate of innovation in the logistics of grain handling and transportation, significant changes have occurred. The excessive focus of prairie farms and the related economy on grain production and export has diminished in favor of more diversified production and more processing of various types, especially in the livestock industry, on the prairies.

Intercity rail passenger services

In 1967, it was accepted that the CPR and CN operated passenger services at a loss. The NTA 1967 provided for a subsidy limited to 80% of the losses. Ostensibly, the limit was to provide an incentive to the railways to be more efficient, but there was limited scope to do this. The CN was more positive than the CPR in efforts to advance passenger services, but the most obvious way to reduce remaining losses was to let the services decline. In 1978, the federal government established a Crown corporation, VIA Rail, to take over intercity rail passenger services, paying a fee to CPR or CN for use of their tracks. The services included those in the "Corridor," from Windsor to Quebec City (the highest-volume intercity route in Canada, especially Toronto–Ottawa–Montreal), transcontinental service, and some rural/remote services. Rail passenger service is a minuscule portion of intercity passenger travel in Canada. No routes are profitable, so the survival of VIA requires ongoing financial support and periodic injections of capital. There is no economic case for the service, but it survives due to political support to retain some vestige of long distance rail passenger service. It is an anomaly in the general direction of Canada's transport policy.

Uneconomic railway branch lines

In 1967, the railways also continued to face severe restrictions in the abandonment of uneconomic branch lines. There were provisions for subsidy when abandonment was denied, but it was a lengthy process to qualify for subsidies and an excessive number of low-density branches had broader implications for the efficiency of the rail system. Further, the efforts to preserve the old grain logistics system were doomed by the closure of distributed small grain elevators in favour of new large elevators. Over time, the growth of subsidies and the closure of lines were mitigated to a limited extent by the transfer of some branch lines to short-line operators. However, it was not until the NTA 1987 that substantially greater freedom was introduced for the railways to abandon lines (after due notice allowed other operators or local communities the opportunity to assume responsibility for the lines). The policy was part of a broad shift by the federal government to make lower levels of government responsible for then unremunerative services that bring local rather than national benefits. This practice is relevant, also to the policy of the federal government in relation to its responsibilities for ports and airports.

4.3. Changes in the provision and finance of transport infrastructure and related services

The federal government was involved directly in transport investment and operation in many ways. These included the provision of ports and airports and related navigation services and the operation of the Canadian sections of the St Lawrence Seaway. (Crown ownership of CN and Air Canada are considered separately later.) The services have been devolved in quite different ways. What is common is a shift to largely commercial mandates. The services are not considered to be "instruments" of national policy.

Three forces led to profound changes in federal involvement in transportation in the 1980s and 1990s, consistent with the economic principles embodied in the NTA 1967. The first was the increased competitiveness of transport and other markets. The second was concern for the efficiency of services provided. The third and most important factor was growing concern over the levels of government spending and the resulting deficits. Provision of infrastructure and other government operations were a financial drain on the government of the day, so it looked for ways to rid itself of these obligations. A few details of the changes are relevant.

Since 1936, the National Harbours Board (NHB), a Crown corporation headquartered in Ottawa, had been responsible for eight deep-sea ports in Canada. The level of detailed management from Ottawa was considerable. The administration of airports was even more directly under political control. The later development of airports and the sole responsibility of the federal government for aviation led to the Department of Transport assuming responsibility for most airports. Subsequently, but with different pressures in the two types of facility, the local port and airport communities argued for greater independence to achieve more efficient and responsive managements. By the 1980s, the federal government recognized the need for change. Users demanded more responsive local management and federal finances were under increasing pressure. Devolution enabled the government to introduce changes that would result in more efficient services and would shift financial responsibility off taxpayers and onto users. Change was evident first in ports but, so far, it has gone further with airports.

The first major devolution of authority for ports occurred with the Canada Ports Corporation Act of 1983. Ports of national importance and with the prospect financial viability would become corporations, but still as agents of the Crown. Later, in the Canada Marine Act 1998 (CMA 1998), the individual port corporations (and local commission ports) were renamed authorities and given greater local autonomy. At the time of writing, there are 19 port authorities. However, several restrictions were placed on the authorities, for example on financing structures and property disposition. These restrictions, and payments to the federal government (as well as the level of taxes levied by some municipalities) are controversial. In 2003, the Canada Marine Act Review Panel recommended changes, but no action has been taken to date. Since 1995, port devolution policy has resulted in the transfer of all but 67 of 549 regional and local ports to locally based public or private bodies, while 34 remote ports are retained by Transport Canada.

Airports had not been administered under specific airport legislation, so that their devolution was easier in some ways than that of ports. The first four local airport authorities were created in 1992, when community-based groups in four cities negotiated the transfer of the airports to not-for-profit local airport authorities (LAAs). Subsequently, in 1994, the National Airport Policy confirmed the intention of the government to commercialize the national airport system and to divest outright regional, local, and small airports. Transport Canada continues to own, operate, and financially support 13 remote airports. By June 2003, 124 of 136 airports had been transferred primarily to not-for-profit community-based entities. The airports are to be managed and funded by the local authorities with the larger airports paying rent or lease fees to the federal government. A common feature of the boards of the airport and port authorities is that directors cannot be users of the facilities.

Part 3 of the CMA 1998 transferred management of the Canadian portion of the St Lawrence Seaway to a not-for-profit corporation, the St Lawrence Seaway Management Corporation, to pursue a commercial approach to the St Lawrence Seaway operation while protecting the interests of the users and long-term viability of the seaway. Five of the nine directors of the corporation represent users of the seaway. The corporation has been successful in bettering its business plans.

The approaches taken to the provision of air and marine navigation services are in sharp contrast. The model for the former was adopted from New Zealand. Under the Civil Air Navigation Services Commercialization Act, Nav Canada, a non-share capital private corporation, assumed responsibility for the Air Navigation Services' (ANS) network and facilities on November 1, 1996, for Can. $1.5 billion. The company is responsible with setting charges for air navigation services levied on airlines and aircraft operators to recover the costs incurred by Nav Canada in providing air traffic control, flight information, and other ANS services. A key to the success of Nav Canada, which enjoys a natural monopoly, is its governance by a stakeholder board.

Marine navigation and ice-breaking services, on the other hand, are still provided the by federal government, now by the Ministry of Fisheries and Oceans, not Transport. There is debate over the efficiency of the services and the extent and manner by which cost recovery should be achieved.

In summary, the 1990s saw a significant reduction in federal involvement in the provision of transport infrastructure, accompanied by a stronger emphasis on cost recovery. Perhaps as important as the change in the magnitude of user charges is the change in attitudes to them. There is now a wide recognition that user charges are important to bringing efficiency to the behavior of both users and providers of infrastructure. Users are generally willing to pay as long as the services are provided efficiently and the structure of charges is deemed rational. However, many contentious issues remain. Are the governance structures of new organizations and the competitive and regulatory frameworks within which they operate sufficient to assure efficiency in the long run? There is also concern that the government has moved too far from a policy of using transport infrastructure investment to promote economic growth, to now seeing transport as an industry to be taxed.

4.4. Privatization of Crown corporations providing transport services

The historical importance of transport in Canada had led to two major service providers, CN and Air Canada, being Crown corporations. Two developments made it impractical and undesirable to continue Crown ownership. First, the government lacked the resources to be able to contribute capital. Reliance on debt by the corporations created inappropriate financial structures. Second, in a highly competitive environment, Crown corporations could not be used in ways to achieve special non-commercial objectives. There were no monopoly profits in business segments to spend on non-economic goals. Further, non-economic goals can be achieved best if subsidies are provided to the most efficient operators.

Air Canada was privatized in 1989. It was the largest air carrier in Canada, and took over the failing Canadian Airlines International in 1999. Air Canada has faced the challenges of government restrictions due to its market dominance, the rise of aggressive low-cost air carriers, and a highly cyclical industry, especially post-September 11, 2001. As noted above, the future structure of the Canadian air industry is uncertain as Air Canada is under bankruptcy protection at the time of writing.

CN was privatized in 1995 in spite of considerable uncertainty about the ability of any management to achieve profitability in light of the history and circumstances of the corporation. The company's transformation into a profitable and highly efficient railway attests to the potential for change within organizations faced with competition and driven by performance goals and able to innovate.

5. Challenges of Canadian transport policy entering the twenty-first century

The twentieth century saw Canada evolve from a nation still under formation using transport as an instrument of public policy to achieve national integration, to a modern economy heavily reliant on trade where transport is seen as a facilitating infrastructure-based service supplied largely by commercial entities. There are policy challenges remaining; some are residual elements of traditional policy concerns (regulatory disputes still arise), while others are in new policy areas (e.g. the importance of environmental concerns). This section reviews major policy themes that face Canada – most of which are shared with other nations. These are:

- the residual role for regulation;
- the growing emphasis on cost recovery of infrastructure and governance of infrastructure provision;
- the increased importance of harmonizing policies with close trading partners;
- the importance of urban transportation challenges;
- the growing influence of environmental concerns;
- security and border issues;
- the ongoing challenge of low-density markets.

5.1. The decline of transport regulation as a major policy issue

For the first half of the twentieth century the regulation of transport was a major issue. However, the general success of deregulation and a reliance on full competition, including intramodal and intermodal competition, has left transport more like any other industry. Regulatory issues are notable by their fewness and distinct nature.

The effectiveness of competition in transport markets has increased as barriers to trade have diminished, and as the incidence of transport costs to the value of goods has declined. This means that not only is competition of carriers within a mode and from alternate modes more likely than in previous decades but that the logistics alternatives of shippers in sourcing or marketing their goods are more likely relevant. Shippers have a greater choice of sources and markets, so they can often negotiate on the basis of bypassing a carrier through buying or selling in alternative locations with different requirements for transport services. The combination of competition from carriers and the logistical alternatives of shippers means that most markets are contestable and competitive. One of the lessons from Canada is that competitive forces often exist even for shippers who appear to be "captive" to a single carrier. Nonetheless, there are some situations in which the alternatives of shippers are limited and the market power of carriers could be a problem. There are residual regulatory needs, particularly for rail shippers, and the CTAR Panel (2001) proposed modifications to the existing regulatory framework for dealing with abuse of a dominant position. But it is important to keep this in perspective: regulatory concerns are now of very limited national importance in contrast to their significance a few decades ago.

A regulatory issue that may emerge in Canada is the possibility of cabotage right to foreign airlines, a practice that has been generally denied in the regulation of aviation. Cabotage arises as an issue in Canada because of the structural instability that has existed in the airline industry since 1987. While deregulation has worked well in facilitating competition and generally lower fares, it has been accompanied by the financial failure of several companies. While the troubles of Air Canada are attributable in part, at least, to its high cost structure, there is concern that the size of the Canadian market might not be large enough to support a competitive domestic industry. Allowing foreign competition is one option to increase competitive supply. But this is complicated by international regimes, which traditionally focus on "national" carriers. The CTAR Panel (2001) recommended pursuing an "open skies" arrangement with the USA, but the prospects of this are uncertain. It seems likely that it is merely a matter of time before the internationalization of business in goods and services catches up with the airlines.

5.2. *The changing emphasis in cost recovery in provision of infrastructure*

Fiscal necessity and appreciation that cost recovery is associated with more efficient behavior by the providers and users of services have resulted in considerable progress in cost recovery. Deficits still exist in marine services, but they are being tackled. Associated with the emphasis on cost recovery has been the devolution of responsibility for infrastructure provision and finance to local authorities. There are controversies over levels of user charges as well as issues of governance of the corporatized infrastructure providers.

Highway finance and taxation

Roads are predominantly a provincial responsibility. The federal government is responsible for very few roads, but periodically becomes involved in shared projects with the provinces, such as the Trans-Canada Highway completed in 1965.

The federal government does not have a "user charge" to recover its expenditure on roads as it does for other modes of transport. Fuel tax levies are treated as general revenue. There is a clamor by users and provincial governments for the federal government to stop extracting money from the transport system, either by increasing spending or by returning some or all of fuel tax revenues they collect. The prospect of this happening is uncertain, but there are compelling arguments. First, unlike the Goods and Services Tax (GST), the fuel tax paid by businesses is an input tax that cannot be recaptured by firms. It is, therefore, inconsistent with the government's main intention, when introducing the GST, to remove distortions in the industrial structure of the country through the burden of taxes on businesses. Second, as the tax applies to fuel used by all modes, the railways argue that it is a tax that they pay for which they get no benefit and that puts them at a disadvantage to the US railways that do not face such a tax. Third, there is a recognized need for greater expenditures to maintain the country's intercity, rural, and urban networks in the face of constraints on levels of government to raise taxes. Canada, like other countries, faces a deficit in infrastructure investment. Finally, there are arguments that dedication of the fuel tax to a highway expenditure program (at all levels of government) would result in more efficient (less politically motivated) expenditure. The CTAR Panel (2001) suggested introduction of road management funds, institutions analogous to airport or port authorities to manage the provision and finance of regional road systems. It may be too radical a change for the near term. However, given the pressures on both federal and provincial governments to place priority on health care and education, infrastructure user charges will have to play a major role in financing that infrastructure.

Issues in the governance of infrastructure providers

The federal government has reduced its direct involvement in infrastructure provision, devolving airports and ports to local authorities, and air navigation to Nav Canada. Responsibility for marine navigation remains in government, but has been moved from the Ministry of Transport, to Fisheries and Oceans.

While it is widely accepted that the devolution of responsibility for the provision and financing of facilities and services has been beneficial, there are concerns about the governance structures. For example, the structure of the airport and port authorities means the boards of directors are not responsible to any person or organization. Nor are facility users directly represented on these boards, unlike Nav Canada and the St Lawrence Seaway Management Corporation. Consequently, there are questions about accountability and the danger that airports, in particular, may over-build and over-charge in the absence of competition. The concern highlights the absence of pricing oversight in the Canadian model. It implicitly assumes sufficient internal constraints and external competition on the authorities that there is no need for regulatory oversight. Another issue is the degree of financial independence. The main concerns for the port authorities relate to limits on their borrowing. There is concern that non-profit authorities could aggravate the business cycle. Non-profit is thought to protect users and consumers from monopoly price increases, but, on the other hand, non-profit entities are not expected to incur deficits over any substantial period of time. There is a fear that infrastructure suppliers could have sufficient market power to offset market downturns with price increases, at the very time that customers facing depressed and competitive conditions would be wanting to lower prices to stimulate business.

In spite of these qualms, there is a broad consensus that the devolution of infrastructure provision has led to more timely investments, improved efficiency, and greater emphasis on user-finance for infrastructure. Issues of governance are a matter of "fine tuning" the model rather than fundamental structural flaws in the concept.

The return governments expect on assets employed in transportation

Although not achieving full cost recovery of marine expenditures, the federal government is characterized as having moved from its historic position of treating transport as a means to promote economic growth to now seeing transport as an industry to be taxed additionally and differently than other industries. Cited are not only concerns about the fuel tax, but also the government's requirements for levels of payments made by port and airport authorities that are seen by many as excessive and, perhaps, inappropriate altogether. The issue of these payments to the government is aggravated by the common use of taxes to support similar infrastructure in the USA. This situation is returned to separately later.

The possibility of viewing transport as a "cash cow" is not limited to the federal government. Each province and municipality with transport services passing through it may try to take advantage of being able to levy a charge on a small part of a much larger movement. This strategy may have severe consequences for the services in a competitive environment. Currently, it is an issue for the application of municipal property taxes in ports. In British Columbia, the level of property tax levied on port property in some municipalities has led the province to introduce the Port Property Tax Act 2004 to roll back and cap excessive taxes.

5.3. Urban transportation challenges

Urban transport has not played a significant part in national transport policy in Canada because it falls under provincial jurisdiction. The lack of federal involvement and support in urban transportation is unlike most other OECD countries. However, the scale of Canadian urban and highway transport problems is growing, and the capacity of the provincial and municipal governments to deal with the problems is increasingly questioned. There are pressures for change in the level of federal involvement.

Despite Canada's large land area relative to population, 77% of the population lived in urban areas in 2001. The urban regions are the growth poles of modern economic activity. Unfortunately, the growing urban communities, like cities everywhere, face growing traffic congestion and air pollution. Arguably, the major transportation problems facing the majority of Canadians are urban transportation challenges.

To date, the federal government's involvement in urban transportation has been confined to the occasional demonstration project (e.g., contributing toward a light rail line in Vancouver prior to the World's Fair in 1986) and participating in projects associated with federal jurisdiction, such as ports and airports. In spite of arguments for user pay, which for urban transport implies collecting revenue from the local users, there is a growing argument for federal involvement and financial support. In part, this is because of the drain of fuel tax revenue from the auto sector into the general revenue of the federal government.

There are strong indirect reasons for a federal interest in urban regions. Although resource-based industries are still important in many regions of the Canadian economy, the cities are the engines of growth for a modern service and information economy. Even much of the economic activity linked to resources takes place in an urban setting. Further, the cities are the location of the major environmental challenges. In spite of the jurisdictional issues, it is likely that the federal government will become more involved in urban transportation, although the form and institutional arrangements are yet to be determined. *Ad hoc* financing of specific infrastructure projects will likely be the first step, but all parties know that urban regions need a consistent and recurrent source of funding to carry out long-term plans and investments. Further, user charge systems bring greater accountability to funding decisions.

A frequent problem in dealing with urban transport is the fragmentation of communities among local governments. The CTAR Panel recognized that, often, cities are institutionally ill equipped to make efficient decisions on a regional basis. The CTAR Panel recommended new institutions that would combine road and transit funding and be answerable to the region. A regional transportation authority might receive funding from federal and provincial sources and be responsible for direct financing from tolls or similar sources of funds. However, the prospects for such agencies are uncertain to say the least.

5.4. Environmental concerns in Canadian transport policy

As elsewhere in the world, the environmental concerns associated with transport have become increasingly recognized and important. Canada has signed the Kyoto Protocol, although there is no specific implementation plan to achieve the Kyoto standards. There is no mention of environmental concerns in the Canada Transportation Act (1996). Transport Canada has sustainable development strategies, and extensive consultation has taken place on alternative policies to achieve Kyoto standards (Transport Canada, 1999). But there is no specific set of measures, no timetable targets, and no official directive in the transport legislation. The omission from the Canada Transportation Act may be viewed as an oversight that revisions to the Act will rectify. The issues will be how extensive and meaningful these considerations will be. The Royal Commission on National Passenger Transportation (1992) and the CTAR Panel (2001) both recommended moving toward including full environmental and social costs in transport decision-making, including developing taxation/pricing principles that would force users to internalize the external costs. The government White Paper *Straight Ahead: a Vision for Transportation in Canada* (Transport Canada, 2003) indicated that this is the direction of policy. It calls for more research on implementation strategies and procedures. Nonetheless, these are all recommendations and not yet legislative directives. There is no doubt that environmental concerns are going to be an important influence on national (and regional) transport policy decisions in years to come, but the policy mechanisms and degree of rigor remain to be determined. At present, environmental concerns do not play a major role in Canadian (and US) transport policy, as compared with Europe, but this will change.

5.5. The conflict of national transport policies

Disputes among governments within Canada about the appropriate levels of taxation and revenue collected from transport are one aspect of a wider phenomenon: the conflict among transport policies of different governments. Efforts over the years to harmonize provincial transport policies in the interest of reducing barriers to trade have been largely successful. However, the growth of trade and the increased flexibility in logistics systems now means that differences among national policies are more likely to have trade-distorting effects than previously. Similarity of Canadian policies with those of other countries may be an ideal, but it is not always practical.

For Canada, the policies of the USA, which has an economy about ten times greater than its own, are very important. Also, an adjacent country feels the main brunt of conflicting policies. The North American Free Trade Agreement has left many more contrasts in policy than the economic integration associated with the formation of the EU, and, unfortunately, the transport industry is replete with them. In the face of these practices, what are the appropriate policies for Canada?

Current concern exists about how Canada should respond to the common use in the USA of public investment in transport as an instrument to shape economic development and national security in contrast to the user pay and even revenue generating position adopted in Canada. Debate also exists about whether Canada should open its transport markets to cabotage to the advantage of shippers and travelers but to the disadvantage of current Canadian firms if cabotage is not provided to Canadian firms to benefit from improved access to the US market.

The actions of US states and cities are also of concern. For example, British Columbia ports are at a disadvantage compared with the ports of Seattle and Tacoma because of the net tax gains of these US ports. How does this affect the national interest? Is the national interest the ability of Canadian ports to move a certain part or all of national trade? Is the national interest in being able to move that share of US trade that might result from "fair" competition? Or should Canada accept and take advantage of the willingness of others to supply services at less than full costs even if it adversely affects employment in Canada? The last view is unlikely!

Global transport policy issues are important in shipping and aviation. Opening entry into international aviation markets is inhibited by the regime of formal multilateral and bilateral agreements. The concept of freedom of the seas has left these markets open, although comity of nations has inhibited making pricing agreements illegal in liner shipping. The ready access of ship owners to low-tax regimes in open-registry countries has led many countries to respond. In Canada, the opportunity since 1990 to manage international shipping without exposing the net income annually to tax has enabled growth in international shipping management, including major international firms such as Teekay Shipping.

5.6. Security and border issues

After 2001, recognition must be given to the emergence of security as a major border issue. Notwithstanding the general and effective reduction of visible barriers to trade and travel across borders prior to 2001, the changes in international security issues at the start of the twenty-first century have raised issues of security to new heights. They exist widely and absorb people's time and societies' resources. The major issues are:

- What costs in terms of human freedoms of many types to accept in the name of security?
- What economic costs to accept in pursuing security?
- Who should pay for security?
- What structures of charges to achieve security are appropriate?

In Canada, the border with the USA heightens security issues. Also, in the USA, considerable federal money is being poured into security. It may not be as much as US carriers, ports, and airports argue is necessary, but it is much greater than the funding provided in Canada. Canada needs to provide a level of security at least equal to that desired by the USA because keeping the border open is more important to Canada than to the USA, where security trumps trade. The long-term effects of security issues remain to be seen.

5.7. The problem of low-density markets

This is a concern raised in the CTAR Panel (2001) report. The emphasis on competitive markets for providing transport services and the increased reliance on user finance of infrastructure heightens concerns in low-density markets that they will suffer unduly. All countries face this problem, but it is more acute for a large country with low population.

Some markets cannot support many suppliers, where, perhaps, only one carrier would be viable. The extent to which government policy can protect shippers and travelers in such markets is limited. It is impractical to expect government regulation to intervene effectively in every market for every resident who finds competition is limited. The situation of small markets is even more of a problem when it comes to infrastructure. All infrastructure exhibits at least some economies of scale; sizeable initial investments are required, especially relative to potential traffic volumes. As a result, full-cost recovery from local users might not be realistic. The economics of larger volumes make cost recovery a practical policy for larger markets, but not necessarily for small ones.

Political considerations require providing policing and access to even the most remote regions of the country. The CTAR report recognises there is no immediate answer to the question of what the minimum or desirable level of support for low-density infrastructure is or how it should be financed. They are challenges that have existed throughout Canada's history, and they exist at all levels: national, provincial/territorial, and municipal. As Canada relies more on commercial provision of services and infrastructure, it will be necessary to address explicitly the conditions and levels of subsidy appropriate for low-density regions. Such decisions are a challenge for a democracy.

6. The future of transport policy

The transport industry in Canada, like elsewhere, has changed to be much more like any other industry. The change has come about with the greater maturity of the modes and the greater effect of competition of all forms on the transportation companies. The distinctness of the industry has also been reduced by its ever-closer integration into the logistics and supply chain decisions of companies. This enhances its importance but diminishes its distinctiveness. However, the fundamental importance of mobility to society and the difficulties of ensuring efficient and adequate provision of infrastructure and services in a safe and sustainable manner throughout a country the size of Canada mean that transport policy will continue to have distinctive prominence in Canada

References

Canada Transportation Act Review Panel (2001) *Vision and balance*: Ottawa: CTAR.

Heaver, T.D. and J.C. Nelson (1977) *Railway pricing under commercial freedom: the Canadian experience*. Vancouver: Centre for Transportation Studies.

MacPherson Royal Commission (1961–1962) *Report of the Royal Commission on Transportation*. Ottawa: The Queen's Printer, Volumes I–III.

National Transportation Act Review Commission (1993) *Competition in transportation: policy and legislation in review*, vols I and II. Ottawa: Minister of Supply and Services Canada.

Royal Commission on National Passenger Transportation (1992) *Directions*, vols I – IV. Ottawa: Minister of Supply and Services Canada.

Transport Canada (1985) *Freedom to move: a framework for transportation reform*. Ottawa: Transport Canada.

Transport Canada (1999) *Transportation and climate change: options for action. Options paper of the Transportation Climate Change Table*. Ottawa: Transport Canada.

Transport Canada (2003) *Straight ahead: a vision for transportation in Canada*. Ottawa: Transport Canada.

This Page Intentionally Left Blank

Chapter 46

TRANSPORTION POLICY IN NEW ZEALAND AND AUSTRALIA

DEREK SCRAFTON
University of South Australia, Adelaide

1. Introduction

Australia and New Zealand are neighboring island nations in the Southwest Pacific, though not close neighbors – it is almost 2200 km from Sydney, on Australia's east coast, across the Tasman Sea to Auckland. The transport policies of the two countries have many common features, including deregulation and privatization, during the second half of the twentieth century, but also some significant differences, particularly those engendered by Australia's much larger size and population, and the complications resulting from the division of powers in Australia's federal constitution. Partly as a result of these differences, reform of transport has been faster in New Zealand than in Australia.

Changes in transport policy in both countries have taken place against a background of intense economic and social change in the past 25 years. Tightly regulated and protected economies have given way to more open trading economies, and multicultural societies have replaced a colonial past.

This chapter commences by describing the recent developments in transport policy in New Zealand, where successive governments have structured a domestic transport system to meet the changes in demand and to support the country's competitiveness in the uncertain world trade system. The section on Australian transport policy summarizes the roles of the federal and state governments, the impact their jurisdictional responsibilities have on the pace of change in transport, and the potential to develop national transport policy in a federal framework. It is in this latter area that the contrast between transport policy formulation in New Zealand and Australia is most marked.

2. Transport policy in New Zealand: the 1980s and early 1990s

In the early 1980s there was widespread recognition that the New Zealand transport sector was neither responsive to its customers nor operating at a

Handbook of Transport Strategy, Policy and Institutions, Edited by K.J. Button and D.A. Hensher

desirable level of efficiency or investment. While some changes had already been made before 1984, it was the election of a reformist government in that year that accelerated gradual reform into a whirlwind of change that embraced not merely transport but the whole of New Zealand society.

Transport policy after 1984 and until well into the 1990s had very clear goals. There was an emphasis on minimizing input and operating costs, improving safety performance, and substantially improving the overall level of service to passengers and users of freight services, both domestic and international. To implement these goals, successive governments actively pursued four main policies in the transport system: encouraging fair competition; reducing public ownership; improving accountability; and improving safety.

2.1. *Encouraging fair competition in the transport sector*

Following many decades in which the New Zealand transport system had become one of the world's most highly regulated, the 1980s and 1990s were marked by the rapid removal of most of the limits on the quantity of transport that operators could offer in any sector. The domestic aviation, airports, ports, and coastal shipping markets were opened to effectively unrestricted competition by local or international operators; the statutory dominance of the rail sector in the land transport freight market was eliminated by 1986; and access to the road freight transport sector was opened to all who met safety-focused entry standards. In the land transport passenger market after 1989, rail, bus, and ferry operators could directly compete for commercial services and bid for fixed-term contracts for services subsidized by regional councils. At the same time, the complex taxi licensing system was replaced by a system that concentrated on the quality of operators and drivers, rather than the number of vehicles in the market. Issues related to fair competition in transport became the responsibility of the economy-wide Commerce Commission. Internationally, New Zealand actively promoted "open skies" air services agreements, and sought wherever possible to reduce or remove other restrictions on international transport.

2.2. *Minimizing public ownership of transport and related systems*

With transport policies that sought to impose greater market discipline on transport operators, it was a logical corollary that both central and local government would question their ongoing role as direct operators of transport businesses. New Zealand's financial difficulties through much of this period provided a further impetus for a withdrawal from many previously publicly owned transport operations, both in terms of realization of asset value and transfer of public capital investment risk. High-profile transport asset sales to the private sector included Air New

Zealand, the New Zealand Shipping Corporation, New Zealand Rail (now known as Tranz Rail), and the key airports at Auckland and Wellington. Equally significant, but perhaps less in the public eye, was the privatization of most of the public road maintenance and construction sector, and the sale to the private sector of all but two municipal public passenger transport operations. Safety inspection and related businesses in the air, land, and maritime sectors were also sold.

2.3. *Improving accountability within the transport system*

Where asset sales were difficult or impractical, substantial structural and operational changes were put in place to improve public agency efficiency. In the land transport sector, a series of changes between 1989 and 1996 established the National Roads Fund. Other than a general excise tax on petrol and charges for accident compensation, all revenue from road users (excise tax, road user charges, and vehicle-licensing fees) is paid into this fund. Funding is then allocated by a Crown entity – Transfund New Zealand – to local authorities for roads, public passenger transport, and other land transport purposes, to Transit New Zealand, a Crown entity responsible for the state highway network, and to pay for all road safety activities by the police and other agencies. Other than the police, the responsible agencies are required to put all such expenditure out to competitive tender, and cannot generally compete for such work themselves.

By 1986, with the completion of a major reform begun in 1977, all diesel-powered vehicles had ceased to pay fuel tax (other than the economy-wide goods and services tax) but instead directly pay road user charges based on the weight and characteristics of each vehicle, the distance traveled, and a range of costs imposed by each vehicle on the infrastructure.

New Zealand's ports were reformed into competitive commercial businesses owned by local authorities. A number have minority private sector shareholding, and are quoted on the New Zealand Stock Exchange. The former central employer of waterfront labor, the Waterfront Industry Commission, was abolished in favor of local employment by port companies, while stevedoring became more contestable. Regional airports have largely followed the same corporatization path, although, apart from Auckland and Wellington, none has private shareholding. Former branches of the New Zealand Ministry of Transport, such as air traffic control and weather forecasting services, have been commercially restructured into the publicly owned Airways Corporation of New Zealand and MetService New Zealand. In 1990 the New Zealand Ministry of Transport itself was restructured into a small multi-modal policy agency, with very limited operational functions. Six agencies working in cooperation with the ministry (Transfund New Zealand, Transit New Zealand, and the four safety agencies noted below) were established to manage operational and regulatory functions.

2.4. Improving safety within the transport system

In the early 1990s the operational management of transport safety licensing and regulation, and the promotion of transport safety, became the responsibility of the Land Transport Safety Authority (road and rail), the Civil Aviation Authority (which also manages aviation security), and the Maritime Safety Authority (which also handles marine pollution risks). The Transport Accident Investigation Commission became the primary investigator of transport accidents. With the exception of the Transport Accident Investigation Commission, and a small number of defined public good activities, these agencies are required to support themselves from user charges. The directors of each of the three modal authorities are the licensing authorities for the transport sector, and have significant statutory independence to carry out their roles of controlling entry, monitoring safety performance, and withdrawing licenses and certificates as appropriate. To maximize the effectiveness of enforcement activities, the former New Zealand Ministry of Transport Traffic Safety Service was merged into the police in 1992. The National Roads Fund (and thus road users) meets the full costs of police road safety activity, including the highway patrol and a specialist heavy vehicle unit.

2.5. Outcomes

In any assessment of the success of the changes outlined above, it is essential to remember that this pace of reform was common to most sectors of New Zealand society in the 1980s and until the mid-1990s. While the transport sector was changing, for example, the whole local government system and the public service were restructured, agricultural subsidies were largely eliminated, the tax system was fundamentally redesigned, and most other sectors of the economy were exposed to commercial competition.

The process that took place in the transport sector is often described as deregulation, though, given the accompanying volume of legislation, it might better be characterized as re-regulation, as economic regulation and statutory limitations were replaced by quality control through, for example, safety and occupational standards. While the basic processes were undoubtedly influenced by contemporary concepts of the limited role of government, transport changes were primarily to improve the efficiency and safety of transport systems.

The complexity of change swept through the whole New Zealand economy, and selecting one sector for examination in isolation can sometimes give a misleading impression of the contemporary environment. The process of change continues in an open economy, and it may be too early to give an unequivocal judgement of the changes. Nevertheless, by most measures the combination of enhanced competition, greater private sector involvement and public sector reform made the transport reforms of the 1980s and early 1990s an overall success. Consumer service levels have risen and prices have fallen in the airline industry, in the ports sector, on long-distance road and rail services on the main routes, and in the taxi sector. Freight service has dramatically improved in reduced costs, productivity, and levels of service as road, rail and coastal shipping have reshaped, improved, and innovated to meet the changing transport task.

Input costs for many elements of the transport sector have fallen in a competitive environment. Road construction and maintenance costs have fallen by 10–15% in the face of intense technological innovation by the construction industry – and this trend continues. Investment in ports, airports, and the road network has increasingly focused on areas of traffic demand rather than merely supporting local prestige projects. Contracting and competition have reduced the costs of delivering public transport. In the public sector, more commercial structures have meant that long-delayed investment in air traffic control services, for example, has moved the Airways Corporation to become a world leader in service and technical innovation.

The restructuring of safety management systems has maintained relatively low levels of risk in the aviation, rail, and maritime sectors, reduced significant public sector financial risk, and reduced the road toll by 40% since the late 1980s.

However, privatization of key transport industries does not necessarily remove all financial risk for the public sector. Whatever their causes, business failures enveloped Air New Zealand in 2002 and Tranz Rail in 2003, forcing the New Zealand government to intervene to protect shareholders and to maintain a level of service that the wider community deemed necessary. In Air New Zealand's case, the government had to recapitalize the business in a way that made it the majority shareholder; while Tranz Rail's infrastructure is being taken back into public ownership to protect the long-term viability of rail freight operations. Neither Air New Zealand nor Tranz Rail is likely to revert to the protected status of the early 1980s, but the lesson that public risk is not eliminated by privatization of major transport systems has been a salutary one for New Zealand. (Similar failures and the need for remedial action arose in transport in Australia, too.)

In the public sector, reductions in the costs of road improvements and public passenger transport services were significantly offset in the early part of the 1990s by substantial cuts in investment levels, as New Zealand faced a major public expenditure crisis. Whereas more commercially structured businesses such as ports and airports had largely eliminated investment backlogs by the late 1990s, the road and public transport networks – particularly in Auckland – have yet to overcome past under-investment.

Changes on the scale outlined here have significant social costs. While the human costs of reform in the ports and rail industry were handled in a relatively sensitive manner, most observers accept that the productivity enhancement on the scale that characterized much of the transport sector in the 1980s and early 1990s had a significant impact on local – especially rural – communities and society as a whole, the consequences of which are still in progress.

3. Transport policy in New Zealand: into the twenty-first century

By the late 1990s, attitudes to structural reform in New Zealand were changing. While there was a general acceptance that many of the transport reforms had been necessary and successful, there was a growing public concern at the pace and impact of reform on society as a whole.

In a key change of direction, proposals put forward in 1998 for merging the management of the state highways owned by Transit New Zealand and the local road network owned by 75 local authorities into a small number of companies jointly owned by central and local government, each able to set its own prices for road use, were never introduced into parliament in the face of substantial opposition at the consultation stage.

Public dissatisfaction with the process of reform was increasingly accompanied by a growing realization that there needed to be greater attention paid to the social and environmental costs imposed by the transport sector on society as a whole. A seminal series of research documents on the land transport system released by the New Zealand Ministry of Transport in 1995–1997, collectively known as the Land Transport Pricing Study (New Zealand Ministry of Transport, 1995a,b, 1996a–c, 1997), identified substantial costs that were being imposed on society as a whole by a narrow focus on road input and operating costs. Increasing and visible air pollution in Auckland and Christchurch led to a study – the Vehicle Fleet Emissions Control Strategy (New Zealand Ministry of Transport, 1998) – that questioned New Zealand's history of inaction in this area. Increasing numbers of local community groups, such as pedestrian and cyclist advocates, began to question the long-term sustainability of elements of contemporary urban transport policy.

The policies of the coalition government elected in late 1999 reflected these wider concerns about the impacts of transport and the processes of reform. While the new government did not seek to fundamentally reverse the changes of the previous two decades, it wanted to place these reforms in a broader context of sustainable development, as it sought to return New Zealand's per capita income to the top half of those countries that belong to the Organisation of Economic Cooperation and Development.

In late 2002, the New Zealand Transport Strategy was released, which committed the government to an explicit set of goals and objectives, stating that "By 2010, New Zealand will have an affordable, integrated, safe, responsive, and sustainable transport system" (New Zealand Government, 2002). To achieve this goal, the strategy has five main objectives: assisting economic development; assisting safety and personal security; improving access and mobility; protecting and promoting public health; and ensuring environmental sustainability

3.1. Assisting economic development

While economic efficiency promoting trade remains a key aim of transport policy, "efficiency" has been given a wider meaning to focus on the way all modes can contribute to the total transport system in the most sustainable and cost-effective manner. The process of project analysis in the renamed National Land Transport Fund (formerly the National Roads Fund) is being revamped to include a wider input for government strategic policy, as well as incorporating the wider social and environmental impacts of individual projects. The purpose statements of Transit New Zealand and Transfund New Zealand are being modified accordingly.

Significant additional resources have been made available for public transport, walking, cycling, alternatives to roads, and the development of regional transport networks in growth areas, as well as making a start on the substantial main road investment backlog. Legislation has begun to develop public–private partnerships in the road infrastructure sector, and tolls will be allowed on new roads. Development of the existing road user charges system into an electronic charging system has been approved in principle, and detailed work on a business case is now under way.

The problems of Auckland are a major concern. While additional funding has been made available for the incomplete motorway network, the government has also repurchased the urban rail passenger system from the private sector. The long-term planning of the urban transport network to support the Auckland Growth Strategy is well under way, including work on congestion pricing.

Most fundamentally of all, a major study (the Surface Transport Costs and Charges Study) is in progress, to evaluate the total costs imposed by road and rail transport on society, and the way in which these charges are met. The result of this study is likely to be a key factor in the government's stated desire to promote sustainability by maximizing rail's share of land freight transport.

3.2. Assisting safety and personal security

Transport safety continues to be a priority. Long-term safety strategies are being developed for each mode, and the road transport sector has already been set the goal of a 35% reduction in fatalities by 2010. New safety legislation is planned to raise railway safety performance.

The events of September 11, 2001 have meant a substantial increase in aviation security activity, including domestic screening. Work on improving port security is already advanced. With a stated desire to see public transport play a larger role in the cities, issues related to personal security are also a matter of concern.

3.3. Improving access and mobility

Support for public transport systems has already risen substantially, by paying regional councils (who manage passenger transport contracts) for the number of passenger-kilometers actually traveled, and providing seed (or "kick-start") funding for new services. As a result, Auckland public passenger transport has risen by 7% a year since 2001, while Christchurch increased passengers by 22% in 2001–2002. Specialist total mobility passenger services are under review, with the goal of providing more consistent levels of service across New Zealand.

3.4. Protecting and promoting public health

Following a major report on the health impacts of road vehicle emissions (New Zealand Ministry of Transport, 2002), which identified the "invisible" road toll from vehicle emissions as being similar to the "visible" road toll, major initiatives are underway to reduce vehicle emissions. New specifications for vehicle fuel will be in force by the start of 2006, while plans for vehicle emission standards and testing are under advanced development.

Other initiatives are being progressed to manage water run-off from roads and the impacts of transport noise on communities. Active promotion of the health and transport benefits of walking and cycling is underway, with a special focus on the 30% of current car trips that are under 2 km.

3.5. Ensuring environmental sustainability

The long-term goal of transport policy is focusing to an ever greater degree on the implications of sustainability for managing the total transport task. It is increasingly recognized that growth in transport demand does not automatically have to follow economic growth, and that there are substantial opportunities for reducing the impact of transport systems.

New Zealand acceded to the Kyoto Protocol on Climate Change in 2002, and a range of related measures are under development, including vehicle fuel efficiency labeling and the future role of carbon taxes. Regional and local transport planning systems are under review, as is a fundamental study of the way in which transport can contribute to sustainable settlements.

4. Australian transport policy: the constitutional framework

Compared with New Zealand, where the development of transport policy is a function of a national government taking into consideration the needs of local governments, the transport industry, users and shippers, the situation is complicated in Australia by the existence of a federal government that has to share the policy formulation task with six sovereign states (New South Wales, Victoria, Queensland, Western Australia, South Australia, and Tasmania), plus two territories, which have, for most transport activities, responsibilities similar to the states. In some federal systems, the states have specific powers and the federal government assumes the residual powers, whereas the Australian constitution vests specific powers in the Commonwealth (federal) government and leaves the residual powers with the states. These residual powers are considerable, giving the state governments wide jurisdiction over transport within their state borders (Scrafton and Starkie, 1985).

Despite this more complex jurisdictional context, it is still possible to compare the development of Australian transport policy with that of New Zealand, as the underlying drivers of policy change are similar: a strong initial emphasis on economic efficiency, governments withdrawing from direct service provision, deregulation, and privatization, later softened to give greater consideration to social and environmental goals. What is clear is that the outcomes in Australia are patchy: there are wide geographic variations and different policies being used to achieve similar goals, with reforms being implemented at different times across the jurisdictions.

The pace of reform varies across the transport modes, and within modes in different states, partly due to varying political considerations (in Australia at any time there is likely to be a mix of right- and left-wing governments in power), but more often driven by legal imperatives or socio-economic opportunities. In most cases the reforms to services have been faster than those involving infrastructure, partly due to the relative ease of separating operations from infrastructure, and the ability to cost services and set prices for services more accurately than charging for use of infrastructure.

The private sector has always played a major role in Australian transport, mainly through the ownership and operation of personal and commercial road vehicles, aircraft, ships, and some railways. Government involvement has been dominant in the construction and maintenance of roads (mainly state and local governments), airports, and airways (mainly the Commonwealth government), operation and subsidization of railways and public transport (states), and regulation of all modes of transport. Deregulation of long-distance freight and passenger transport by road began in the 1950s, followed by reduced protection for freight railways in the following decade, but it was not until the 1980s that reduced levels of regulation were affecting all modes of transport to varying degrees. Then in the 1990s, governments sought to divest themselves of direct operation of transport (e.g. airlines, airports, urban bus services) through privatization, franchising, or contracting.

The public sector still owns the major part of Australia's infrastructure, with all three tiers of government involved in supplying transport infrastructure. Although state and local governments dominate public investment in transport infrastructure, up to 40% of their revenues come from the Commonwealth government in the form of general and specific purpose payments. While the Commonwealth is constitutionally responsible for the communications industry, it gradually expanded its role to provide national highways, railways, airports, etc., then in recent years changed its policy and reduced its holdings in, for example, railways and airports. The states are responsible for a broad range of transport infrastructure, including roads, railways, and ports, while local governments provide roads and some airports. Private involvement in infrastructure is not new, but in recent years greater reliance by governments on contracting out, build, own, operate, and transfer type arrangements and privatization has led to increased private sector participation in railways, roads, ports, and airports. This trend toward greater private sector involvement is likely to continue (Economic Planning Advisory Commission, 1995).

5. Australian transport policy: recent developments

Over 2000 million tonnes of freight is carried around Australia annually, and over 500 million tonnes exported. Road transport accounts for 75% of domestic freight by weight, with rail most of the remainder (much of which is bulk product), though there is a significant coastal shipping trade. Ocean shipping dominates the international freight trade, though air transport accounts for about 25% when measured by value. Almost all international travel is by air, and private road vehicles make up the majority of domestic travel movements (Bureau of Transport and Regional Economics, 2003a).

The following section summarizes the major developments in rail, road, air, and marine transport in Australia, selecting examples from different jurisdictions to illustrate the different responses to policy changes.

5.1. Railways

The historical development of Australia's railways resulted in separate state-owned systems operating on different track gauges that proved to be ill-equipped to deal with competition from other modes in the twentieth century. In addition, state ownership meant heavy political interference in management, rate-setting, and service decisions (Fitch, 2001).

There is still one major government-owned and operated railway in Australia – the mainly narrow-gauge Queensland Rail system, which garners most of its revenues from the haulage of coal from inland mines to coastal ports. The rest of the former state freight railways have been sold or leased to private companies. The largest operator is Pacific National, a company formed by the Toll and Patrick transport groups to acquire and merge the operations of the former government-owned National Rail Corporation (mainly Commonwealth, New South Wales, and Victoria) and FreightCorp (New South Wales), creating a company that operates over most of the nation's standard gauge rail network. Pacific National is an above-track operator, and most of the tracks over which it operates continue to be owned and maintained by government-owned track authorities.

In addition, there are a number of major regional-based private companies operating freight trains, e.g. Rail Australia (based in Victoria), Australian Railway Group (mainly South Australia, Western Australia, and the Northern Territory), and Tasrail. All of these companies, and others (including Queensland Rail), are able to take advantage of the access provisions of competition law to operate beyond their home territory and compete with one another for business over main and branch lines.

Government organizations continue to own most passenger train services in Australia, including the suburban trains in the larger mainland cities, though in some cases (e.g. Melbourne) they may be contracted out to a private operator.

Whereas in New Zealand the government resumed ownership of the rail infrastructure in 2003, in Australia most of the track, signals, and other fixed infrastructure has remained in government ownership, though it may be leased to a private operator or to another government body, e.g. the standard gauge main lines in Victoria are leased to the Australian Rail Track Corporation, the regional freight lines to Rail Australia, and the suburban network to one or more franchisees. Legal and contractual argument continues over access to rail infrastructure, particularly access to track and terminals (Bureau of Transport and Regional Economics, 2003b).

A longer-term problem is how to pay for new rail infrastructure, particularly extensions to the network. The Alice Springs to Darwin rail line, opened in January 2004, was a public–private partnership involving the Commonwealth, South Australian, and Northern Territory governments and the private sector. Similar joint ventures constructed the new rail lines to Sydney and Brisbane airports.

While there are some residual problems resulting from these major changes to Australia's rail system, particularly the existence of an overly complex regulatory framework that is a modern version of the problem of differing rail gauges, the overall level of rail service is a great improvement to that prevailing before the changes. On the freight side, the major challenge facing the railways is to deliver service of a standard that allows them to participate in integrated logistics, increase the use of the network, and generate adequate revenue to justify upgrading the track, particularly that on the east coast corridors.

5.2. Roads and road transport

Lessening of controls over road transport began in 1954, when the Australian High Court determined that laws restricting traffics to rail were unconstitutional in that free interstate trade was inhibited. As that included most inter-city freight and passenger movement, the decision effectively ended government controls on long-distance road transport in Australia. It took the state governments many years to apply similar deregulatory policies to intrastate transport, and once Tasmania removes route licensing, deregulation of road freight traffic will be nationwide. The only controls remaining in place are technical standards (e.g. size and weight of trucks), safety related (including occupational safety, such as limits on hours of driving), or environmental, such as excluding heavy vehicles from certain roads and restrictions on night time deliveries in residential areas.

However, competition is still controlled on country bus services in all states to varying extents, and reform is painfully slow; in effect, protection of single-operator coach routes has replaced protection of rail passenger services. On the other hand, the demand for rural and regional bus service continues to decline, so regulation is less relevant: survival of these services is now the policy issue, and the need for subsidy is more significant than regulation *per se.*

The establishment of the National Road Transport Commission (NRTC) in 1991, as part of a program to encourage greater competition in all economic sectors of the Australian economy, expedited reform in road transport, particularly in achieving greater uniformity in technical and safety standards, improving efficiency in administrative procedures affecting road transport, and setting nationally uniform charges for the use of heavy vehicles. While in the past it might have been argued that such reforms could or would have occurred without such a national body, nobody now would deny that the focus the NRTC brought to the implementation of reforms over a decade did succeed in overcoming what had been deeply entrenched differences between states.

The NRTC's success can be measured by the agreement of all governments, in 2003, to expand the scope of its remit to cover similar areas of technical and safety regulation over rail transport, as the re-branded National Transport Commission.

The ownership and management of public roads in Australia is a function of state and local government departments and agencies, with the Commonwealth responsible for financial matters covering the designated national highway system (Dowcra, 1993). Although the private sector has always been involved in road construction and maintenance, the main reforms to roads in the late twentieth century included greater direct private involvement in the financing and management of roads, particularly in New South Wales and Victoria. This increased involvement has caused major rethinking of the potential for reform in roads, and is a stimulus for consideration of road pricing – previously a political minefield. Privately owned roads such as the CityLink network in Melbourne are real-life laboratories for electronic tolling, differential charging schemes, and asset management programs. When this experience is added to the NRTC's work on heavy vehicle charges and knowledge of overseas application of congestion charges in cities, a framework is in place for improved decision-making for financing of Australian roads.

The Commonwealth government has made it clear, in the AusLink Green Paper released in November 2002 (Australian Department of Transport and Regional Services, 2002), that it intends to move away from the formula-based funding arrangements embodied in the existing federal/state roads agreement to a needs-based assessment process, with an emphasis on national corridors, as a step toward strengthening the link between pricing and investment in roads, as well as investing in transport more generally, rather than just in roads. Although there is some opposition in the states toward implementation of the AusLink proposals, the concern is related more to the quantum of funds that might be available under the new program than to any objection to the policy changes foreshadowed by the Green Paper (Scrafton, 2003). This concern is understandable, when the warranted upgrading of the national highways alone between Brisbane, Sydney, and Melbourne will require over Aus. $10 billion. One can visualize a future when Australia will have a single road-planning and funding authority for main roads, with the existing state agencies acting as regional project managers.

5.3. Urban transport

Reforms to urban transport embrace suburban railways in all major mainland Australian cities, trams in Melbourne, Sydney, and Adelaide, public and/or private bus services in all large cities and many smaller towns, taxis and hire cars, community transport, private cars, cycles, and motorcycles, and a wide range of urban freight vehicles. Compared with the advances seen in national and regional land transport, those in urban areas are less obvious. For example, little change has taken place in the high-cost urban rail sector or in the tightly regulated taxi industry. There has been expansion of the suburban rail systems in some cities, though the benefits have to be seen more in the context of improving the urban environment than in economic terms. Investment is required to lessen congestion problems on strategically important road links in the metropolitan cities, irrespective of investment levels in public transport.

Private and public bus operations have coexisted in Brisbane, Sydney, and Melbourne metropolitan areas for many years, albeit in fairly rigidly defined operating areas, and with government control of fare levels. In recent years, franchising or contracting out of former government bus services in Melbourne, Adelaide, and Perth (and some smaller cities) has contributed to an expansion of the market for the private bus sector, and in turn led to improved productivity and to creation of a critical mass of private bus operators in all cities (Hensher and Daniels, 1993).

5.4. Aviation

After some 30 years (from the 1950s to 1980s) of tight control, mainly by the Commonwealth government, Australian aviation "has been subject to reform pressures almost continuously" (Bureau of Transport and Communications Economics, 1993). The changes commenced with the abolition of the domestic two-airline policy in 1990, which resulted in greater competition, though it proved difficult for a third carrier to break into the trunk route market. However, the collapse of Ansett Airlines in 2001 provided an opening for Virgin Blue to gradually assume the role of a major domestic trunk carrier. The majors are complemented by a number of smaller airlines serving the regional towns and remote Australia.

In general, the deregulated Australian domestic aviation scene in the twenty-first century is one that is likely to continue to evolve. The introduction of Virgin Blue, with a cost structure similar to that of the low-cost carriers found in USA and Europe, creates a competitive environment to which the longer-established operators will need to respond if they are to survive. Creating subsidiary airlines with lower-cost structures is one approach, but experience overseas shows it may not be successful.

While there is still a clear distinction between Australian domestic and international air services, the direction is toward seamless operations, led by Qantas, once a Commonwealth government owned international carrier that was merged with the former government domestic operator, then privatized. While international routes are still subject to bilateral agreements, the Trans-Tasman routes are closely linked to domestic services in New Zealand and Australia, while the influence of Asian-based airlines using fifth-freedom rights create competitive services and fares for passengers in and out of Australia. Competition across the Pacific Ocean is still limited, but, overall, aviation reform is providing a greater range of services and prices for airline users.

Ownership and management of all major airports and many smaller aerodromes and airstrips were the responsibility of the federal government department liable for aviation, but as early as 1958 the Commonwealth began divesting itself of local airports, through a scheme providing financial and technical support to new local owners, including local governments. However, the process was very slow, until given a boost in the late 1980s with a revision to the arrangements, after which local airports were encouraged to levy user fees. Now there are over 200 "small" airports in Australia owned and/or operated by their local community or by private companies. Some of those serving regional towns and cities and tourist destinations have substantial terminals supporting regional air services. As tourism grows to regional Australia, so some of these airports can expect to attract direct links to the major cities and even to some overseas airports.

The big change to airport policy came toward the end of the twentieth century, with the corporatization of the major city airports in 1988 under the Federal Airports Corporation, followed by the sale of the airports to private sector organizations in the 1990s. The first tranche to be sold or leased were Melbourne, Brisbane, and Perth, followed by the airports at Adelaide, Hobart, and other smaller cities, leaving Sydney, the dominant international aviation gateway to Australia, until last. Although the process was driven by the Commonwealth government, the owner of the airports, state governments took an active interest in the sales and in planned infrastructure improvements because the airports are seen as important elements in attracting tourists to the regions. However, the next generation of large aircraft might engender a change to commercial aviation that sees fewer international gateway airports and greater use of feeder services.

5.5. Maritime

Merchant shipping is a vital element of Australia's transport network, both domestic (coastal shipping) and international. State government entities control most of the ports, although several ports have been sold or leased to the private sector in recent years, e.g. in Victoria and South Australia, where Geelong, Portland, and Adelaide are run by companies that have strong links to the road transport industry. In addition, contracting out of non-core services such as cargo handling, terminal construction, towage, security services, and recreational boating facilities. In general there is underutilized berth capacity at most Australian seaports (National Transport Planning Taskforce, 1994)

It is interesting to note that much of the momentum to reduce port costs comes from the pace of change in New Zealand's ports, which in turn was partly stimulated by the threat of port reform in Australia! One of the factors that slowed port reform in Australia was the existence of private sector monopolies, such a pilots and stevedores, which together with labor were able to thwart change. Because port authorities have no direct control over the actions of key players in the performance of the ports, disputes may require resolution in industrial and/or competition tribunals (Bureau of Transport and Communications Economics, 1995).

6. The way forward

By the early 1980s the transport systems of New Zealand and Australia were increasingly failing to manage the increasingly complex transport tasks that the countries' changing domestic and international trade situation required.

The policy changes of the 1980s and early 1990s in both nations launched a major restructuring of the transport sector to promote efficiency, responsiveness to demand, and to improve safety – and generated a period of change whose consequences are still being worked through. While it cannot be claimed that every one of the reforms of this period was successful, on balance they did deliver a transport sector better able to meet the pressures that a changing domestic and world environment was creating.

But change is never easy, and as the 1990s advanced, public interest and transport policy-makers began to be more concerned with the social and environmental consequences of change that were often ignored in the swift restructuring of the 1980s.

By the start of the twenty-first century, transport policy in both New Zealand and Australia is building on past reforms on the way to the long-term goal of a sustainable transport system – one that combines the operational and structural efficiencies gained in the 1980s and 1990s with the realization that any transport system must be closely focused on the way in which it interacts with society and the physical environment. In general, New Zealand's unitary system of government, compared with Australia's more complex federal system, seems to be a factor influencing the pace of transport reform: New Zealand has been a leader in implementing policy changes, with Australia slower to take initiatives, e.g. in signing up to the Kyoto Protocol.

Governments will continue to have a role in transport, primarily as strategic planners and the providers of some infrastructure. If governments do not perform well in these tasks, people and goods will still be moved, but transport will be more expensive and less effective that it could be. While expenditure on transport by governments is substantial, it is small in comparison with the private sector's expenditure on owning and operating vehicles. So public sector supply driven solutions are no longer relevant, if indeed they ever were: the demand for transport evolves primarily from the private sector and the transport function is only part of the broader logistics and distribution system and the need for accessibility, and that demand will set the agenda for policy reform.

Overall, transport has responded fairly well in its contribution to micro-economic reform in New Zealand and Australia. However, the outcomes are variable geographically and across the modes, and are susceptible to changes that are not always for the better.

To continue to meet the complex and rapidly changing transport task in relatively small trading nations distant from all markets, but to do it in a way that maintains and enhances the environmental and social fabric, thereby maximizing the benefits to all New Zealanders and Australians, is a demanding challenge. The momentum of reform needs to be maintained, together with continued operational efficiencies and renewed efforts to ensure investment in productive transport infrastructure. As an example of the size of the challenge to be faced, in 2020 general (non-bulk) freight flows between Australia's major cities will be double what they were in 2000 (Bureau of Transport and Regional Economics, 2003c).

References

Australian Department of Transport and Regional Services (2002) AusLink – towards the national transport plan, Green Paper. Canberra: Australian Department of Transport and Regional Services.

Bureau of Transport and Communications Economics (1993) The progress of aviation reform, Report 81. Canberra: AGPS.

Bureau of Transport and Communications Economics (1995) *Review of the Waterfront Industry Reform Program*, Report 91. Canberra: AGPS.

Bureau of Transport and Regional Economics (2003a) Australian transport statistics. Canberra: AGPS.

Bureau of Transport and Regional Economics (2003b) *Rail infrastructure pricing: principles and practice*, Report 109. Canberra: AGPS.

Bureau of Transport and Regional Economics (2003c) *Freight between Australian cities*, Information sheet 22. Canberra: AGPS.

Dowcra, G. (1993) *Roads policy and australian federalism. BTCE occasional paper 106*. Canberra: AGPS.

Economic Planning Advisory Commission (1995) *Private Infrastructure Task Force report*. Canberra: AGPS.

Fitch, R.J. (2001) "A critical study of the operational and financial performance of the south Australian railways," unpublished Ph.D. thesis. University of New South Wales, Sydney.

Hensher, D. and R. Daniels (1993) *Productivity measurement in the urban bus sector 1991–92*. Sydney: Institute of Transport Studies, University of Sydney.

National Transport Planning Taskforce (1994) *Building for the job: a strategy for Australia's transport network*. Canberra: AGPS.

New Zealand Government (2002) *New Zealand transport strategy*. Wellington: New Zealand Government.

New Zealand Ministry of Transport (1995a) *Land transport pricing study: the cost of roading infrastructure*. Wellington: New Zealand Ministry of Transport.

New Zealand Ministry of Transport (1995b) *Land transport pricing study: roading as an economic good*. Wellington: New Zealand Ministry of Transport.

New Zealand Ministry of Transport (1996a) *Land transport pricing study: environmental externalities*. Wellington: New Zealand Ministry of Transport.

New Zealand Ministry of Transport (1996b) *Land transport pricing study: safety externalities*. Wellington: New Zealand Ministry of Transport.

New Zealand Ministry of Transport (1996c) *Land transport pricing study: national traffic database*. Wellington: New Zealand Ministry of Transport.

New Zealand Ministry of Transport (1997) *Land transport pricing study: options for the future*. Wellington: New Zealand Ministry of Transport.

New Zealand Ministry of Transport (1998) *Vehicle fleet emissions control strategy: final report*. Wellington: New Zealand Ministry of Transport.

New Zealand Ministry of Transport (2002) *Health impacts of road vehicle emissions*. Wellington: New Zealand Ministry of Transport.

Scrafton, D. (2003) "The Auslink Green Paper – proposals, questions and consultation," *Transport Engineering in Australia*, 8:73–79.

Scrafton, D. and D. Starkie (1985) "Transport policy and administration in Australia: issues and frameworks," *Transport Reviews*, 5:79–98.

This Page Intentionally Left Blank

AUTHOR INDEX

This Page Intentionally Left Blank

SUBJECT INDEX