Policy Research

WORKING PAPERS

Transport

Infrastructure and Urban Development
Department
The World Bank
December 1992
WPS 1070

Economic Fundamentals of Road Pricing

A Diagrammatic Analysis

Timothy D. Hau

Most economists agree that road pricing benefits society by curtailing congestion. Efficiency analysis demonstrates why the public rejects congestion pricing. A dedicated road or transport fund is more viable when the road users are charged not only for the damage caused by heavy vehicles but also for congestion.

Policy Research

WORKING PAPERS

Transport

WPS 1070

This paper — a product of the Transport Division, Infrastructure and Urban Development Department — is part of a larger effort in the department to evaluate options for charging for road use. Copies of the paper are available free from the World Bank, 1818 H Street NW, Washington, DC 20433. Please contact Jennifer Francis, room S10-063, extension 31005 or 35205 (December 1992, 96 pages).

Hau presents a conceptual framework for road pricing based on a rigorous diagrammatic — but nonmathematical — framework derived from first (economic) principles. His analysis of traditional arguments about road pricing shows why implementing congestion pricing as practiced in the past has encountered obstacles. Partly, it is because both types of road users — the tolled and the tolled off (those who avoid the road to shun a toll) — are shown to be worse off under a constant value of time, except for the government. And when differences in time valuation are taken into account, primarily those with very high time values are better off. Unless congestion toll revenues are earmarked and travelers perceive that the money is channeled back in reduced taxes, lower user charges, or improved transport services, neither the tolled nor the tolled off will support road pricing. Only where there is hypercongestion is everyone better off with congestion pricing.

In the absence of scale economies or diseconomies, the level of economic profits — toll revenue collections less a road's fixed and non-use-related costs — serves as a surrogate market mechanism indicating that a road should be expanded or downsized. The decision to let roads deteriorate over time is itself an act of disinvestment.

Hau shows that if a road authority levies economically efficient charges for congestion, it is possible to make money on a road. Roads can be profitable in urban areas in the long run because land rents are high; congestion tolls reflect the associated high opportunity costs.

On urban roads with indivisibilities and diseconomies of scale, efficient pricing may curtail the extent of profitable undertakings. On rural roads with indivisibilities and economies of scale, marginal cost pricing can produce short-run profits. Economic efficiency is enhanced by pursuing optimal pricing in the short run and optimal investment in capacity in the long run. The rule is to implement short-run marginal cost pricing while varying road capacity over the long run.

Insights by Newbery, Small, and Winston — about the economic implications of the extensive damage that heavy vehicles cause to roads — enrich the basic Mohring model. Charging for both the external and variable cost of road damage, by assigning a fee based on vehicle weight per axle, can help cover deficits arising from road congestion.

Even if a road network is broadly characterized by increasing returns to scale in building and strengthening roads, the deficit could be closed by diseconomies of scope. A road network that accommodates both cars and trucks costs more than the sum of an autos-only and a (smaller) trucks-only road system. So the surplus associated with diseconomies of scope offsets the potential loss associated with scale-specific economies. A dedicated road or transport fund is all the more viable because road users are charged not only for the damage caused by trucks and heavy vehicles but also for congestion.

The Policy Research Working Paper Series disseminates the findings of work under way in the Bank. An objective of the series is to get these findings out quickly, even if presentations are less than fully polished. The findings, interpretations, and conclusions in these papers do not necessarily represent official Bank policy.

Produced by the Policy Research Dissemination Center

Economic Fundamentals of Road Pricing: A Diagrammatic Analysis

by

Timothy D. Hau

Transport Division
Infrastructure and Urban Development Department
The World Bank

Economic Fundamentals of Road Pricing: A Diagrammatic Analysis

by

Timothy D. Hau

CONTENTS

ACKNOWLEDGEMENTS iv

ABSTRACT v

I. CONCEPTUAL GUIDELINES 3

II. CAUSE OF CONGESTION: PUBLIC/PRIVATE OWNERSHIP OF ROADS 4

III. FOUNDATIONS OF ROAD CONGESTION -- THE CLASSICAL CASE (IN THE SHORT RUN) 8

IV. DEMAND AND SUPPLY 10

V. VERY CONGESTED ROADS AND BOTTLENECKS 12

VI. THE WELFARE IMPACT OF ROAD PRICING 13

VII. THE EFFECTS ON THE TOLLED, THE TOLLED OFF AND THE UNTOLLED 13
Policy Implications 17

VIII. SHORT RUN EQUILIBRIUM 18

IX. CONVERGENCE TOWARDS LONG-RUN EQUILIBRIUM UNDER CONSTANT RETURNS 20

X. OPTIMAL INVESTMENT 22

XI. LONG-RUN VS. SHORT-RUN MARGINAL COST PRICING 24

XII. TRADE-OFF BETWEEN INDIVIDUALS' TIME AND TREASURY ACCOUNTS 25

XIII. FIRST-BEST OPTIMAL PRICING AND INVESTMENT RULES 28
Empirical Considerations 28

The First Rule -- The Optimal Pricing Rule 28
The Second Rule -- The Optimal Capacity Rule 29
Perspective on the Result 29

XIV. RELAXATION OF ASSUMPTIONS 31
1) Differences in Time Valuation 31
2) Demand Variability and Peak-Load Pricing 32
3) Indivisibilities 34
Optimal Pricing and Investment with Indivisibilities: An Example 35
4) Returns to Scale 37
A. Economies of Scale and Rural Roads 38
B. Diseconomies of Scale and Urban Roads 39
C. Diseconomies of Scale and Indivisibilities 41
D. Economies of Scale and Indivisibilities 42
E. The Extent of Indivisibilities vs. Divisibility and Their Effects on Scale (Dis)economies 44
F. Empirical Evidence on the Scale Economy Issue 45
G. Recent Results on Cost Recovery 47
5) Variability of Road Thickness 51
The first rule -- the optimal pricing rule 52
The second rule -- the optimal capacity rule 52
The third rule -- the optimal durability rule 52
Economies of Scope vs. Diseconomies of Scope 53

XV. SUMMARY AND CONCLUSIONS 56
A. The Role of a Road Fund 61
B. The Role of a Transport Fund 62

APPENDIX: 65
Measurement of the Welfare Impact of Road Pricing 65

FIGURES 71

REFERENCES 89

LIST OF FIGURES

Figure 1 Derivation of a Travel Time-Flow Curve of an Urban Highway 73

Figure 2 Derivation of the Marginal Cost Curve and Congestion Toll 74

Figure 2(a) ‘Dynamic’ Phenomenon of Traffic Growth: The Relaxation Effect 75

Figure 3 Welfare Impact due to the Introduction of Road Pricing in the Peak Period: Short-Run Marginal Cost Pricing 76

Figure 3(a) Welfare Impact due to the Introduction of Road Pricing in the Peak Period: Short-Run Marginal Cost Pricing 'Hypercongestion' Case 77

Figure 4 Effect of the Introduction of Road Pricing in the Peak Period on the Off-Peak Period 78

Figure 5 Introducing the (Short-Run Average) Fixed Cost, SRAFC, of a Road, Short-Run Optimal Toll with Economic Profit 79

Figure 6 Long-Run Equilibrium of an Optimally Designed Road with Both Optimal Pricing and Optimal Investment 80

Figure 7(a) Constant Returns to Scale with Road Divisibility: Doubling Optimal Road Capacity, K, and Traffic, Q, Result in Doubling Fixed Cost, Variable Cost and Total Cost (FC, VC, TC) and Toll Revenues ($t^* \cdot Q^*$) 81

Figure 7(b) The Relationship between Short-Run Average Total Cost and Long-Run Average Total Cost and Marginal Cost with Perfect Road Divisibility and Constant Returns to Scale 81

Figure 8(a) Road Indivisibilities under Constant Returns 82

Figure 8(b) Optimal Pricing and Investment with Indivisibilities: Expansion from a 2-Lane Road to 4-Lane Road 82

Figure 9 Economies and Diseconomies of Scale in the Provision of Road Capacity with the Growth of Travel Demand 83

Figure 10 Doubling the Number of Streets -- Road Capacity -- Results in Quadrupling the Number of Intersections and Traffic Lights and Doubling Waiting Time 84

Figure 10(a) Original Scenario -- Existing Street Configuration 84

Figure 10(b) Final Scenario -- Number of Streets are Doubled 84

Figure 11 Diseconomies of Scale: Urban Road Network with Perfect Divisibility 85

Figure 12 Economies of Scale: Rural Roads with Perfect Divisibility 86

Figure 13 Decreasing Returns to Scale and Extent of Indivisibilities 87

Figure 14 Increasing Returns to Scale and Extent of Indivisibilities 87

ACKNOWLEDGEMENTS

A rudimentary version of this research paper was presented at the 2nd International Conference on Privatization and Deregulation, Tampere, Finland, June 16-20, 1991. Part of the findings of this research was presented at the 6th World Conference on Transport Research, Lyon, France, June 29 - July 3, 1992; Seminario de Tarificación Vial 1992 (Road Pricing Seminar), sponsored by the Ministry of Transport, Santiago, Chile, July 20-21, 1992; the Symposium on Road Pricing, sponsored by the Swedish National Road Administration and the Department of Regional Planning, Royal Institute of Technology, Sigtuna, Sweden, November 9-10, 1992; and the Southwest Congestion Pricing Conference, sponsored by the University of Houston's Center for Public Policy and Citizens Advocating Responsible Transportation, Houston, Texas, January 4-5, 1993. Comments by participants at these places are gratefully acknowledged. I have benefitted from the discussions and comments of many people on the controversial subject of road pricing and I take this opportunity to thank William Vickrey, Michael Beesley, Kiran Bhatt, Richard Bird, Andrew Evans, Stephen Glaister, Phil Goodwin, José Gómez-Ibáñez, Chris Hendrickson, Peter J. Hills, Jake Jacoby, Jan Jansson, Peter Jones, Theodore Keeler, Damien Kulash, Odd Larsen, Henning Lauridsen, Douglass B. Lee, Peter Mackie, Anthony May, Robert McGillivray, Herbert Mohring, Richard Musgrave, Yew Kwang Ng, Esko Niskanen, Tae Oum, Farideh Ramjerdi, Gabriel Roth, Kenneth Small, Antti Talvitie, Tom Tietenberg, Burkhard von Rabenau and Sir Alan Walters. I am also grateful to World Bank staff, including Esra Bennathan, Philip Blackshaw, José Carbajo, Anthony Churchill, Shanta Devarajan, Asif Faiz, Jeffrey Gutman, Clell Harral, Ian Heggie, Chris Hoban, Gregory Ingram, Frida Johansen, Kyu Sik Lee, William McCleary, William Paterson, Richard Scurfield, Zmarak Shalizi and Larry Summers.

Support services from Margarat Callan, Pam Cook, Barbara Gregory, Beatrice Sito and Gabriella Steinmetz are gratefully acknowledged.

These acknowledgements do not mean that they share in the views expressed in this paper. I alone retain full responsibility for its contents.

Timothy D. Hau

ABSTRACT

Economic Fundamentals of Road Pricing: A Diagrammatic Analysis

by

Timothy D. Hau

This paper presents a conceptual framework for road pricing based on a rigorous diagrammatic -- but non-mathematical -- framework derived from first (economic) principles. It throws light on congestion pricing systems and issues surrounding short-run and long-run marginal cost pricing, scale economies and diseconomies, indivisibilities, road durability, the peak-load problem in urban transport and the financial viability of the public provision of road services. The paper integrates the ideas of Mohring, Strotz, Vickrey, Walters, Keeler, Small, Winston and Newbery into a single analytical framework.

Analysis of the traditional road pricing arguments demonstrates why congestion pricing as practiced in the past has understandably encountered obstacles to implementation. This is partly because both types of road users, the tolled and the tolled off (those who avoid a road in order to shun the toll), are shown to be worse off -- with the exception of the government -- under a constant value of time. Even if differences in time valuation are taken into account, it is still essentially the case that primarily those with very high time values are made better off. Unless congestion toll revenues are earmarked and travellers perceive that the money is channeled back in the form of reduced taxes, lower user charges or improved transport services, neither the priced nor the priced off would support road pricing. It is only in the case of hypercongestion can congestion pricing be shown to make everyone better off.

In the absence of scale economies or diseconomies, the level of economic profits, i.e., toll revenue collections less the fixed and non-use related costs of a road, serves as a surrogate market mechanism indicating that a road ought to be expanded or downsized. The decision to let roads deteriorate over time in and of itself is an act of disinvestment. Further, it is shown graphically that if a road authority were to efficiently charge for congestion, it is possible to make money on the road. Profitable roads arise in urban areas in the long run because land rents are high and congestion tolls reflect the high opportunity costs. Yet, efficient pricing in the presence of both indivisibilities and diseconomies of scale in urban roads may curtail the extent of profitable undertakings, whereas pursuing marginal cost pricing under the restrictive conditions

of both indivisibilities and scale economies in rural roads could result in profits in the short run. Economic efficiency would be enhanced if optimal pricing were pursued in the short run and optimal investment in capacity were pursued in the long run. The rule is therefore to implement short-run marginal cost pricing while varying road capacity over the long run.

Recent extensions by Newbery, Small and Winston have enriched the basic Mohring model that this paper develops diagrammatically by incorporating the fact that heavy vehicles are responsible for extensive road damages. Charging for both the external and variable cost of road damages on a vehicle weight per axle basis can help cover the deficit that may arise from congestion tolling. Even if a road network is broadly characterized by increasing returns to scale in road construction's use and strengthening, the deficit could be closed by diseconomies of scope. The existence of scope diseconomies in highways means that a road network that accommodates both loading and traffic volume, as found universally, costs more than the sum of an autos-only and a (smaller) trucks-only road system. Hence, the surplus associated with diseconomies of scope offsets the potential loss associated with scale-specific economies. The viability of a dedicated road or transport fund is enhanced by the fact that the road pavement is charged in two dimensions: once when traffic flow creates congestion and another when traffic loadings cause road damages.

Economic Fundamentals of Road Pricing: A Diagrammatic Analysis

1. In major conurbations of both developed and developing countries, congestion is incessant during the peak periods and oftentimes the interpeak.[1/] Yet the traditional methods of effectively curtailing congestion are few, and their usefulness limited. On the supply side, the expansion and improvement of roads is restricted by increasingly tight fiscal and physical constraints. On the demand side, however, the problem can be addressed by pricing or regulation. This paper uses the pricing (or market-based) approach to grapple with congestion because of its inherent flexibility and power of discrimination.[2/] The aim of road pricing is to internalize the externalities generated by road use. My focus here is on removing the external effects caused by motorists by charging for congestion and road damage externalities. Congestion is recognized as an important type of externality from vehicle usage in both developed and developing countries in that it represents a large share of total estimated road use costs (Newbery (1988bc, 1990) and Newbery, Hughes, Paterson and Bennathan (1988).[3/]

2. This research begins with a series of two papers on road pricing in theory and practice. The first paper presents a conceptual framework for road pricing. It gives an interpretative abridgment of the literature on the theory of optimal pricing, investment and durability of roads

[1/] By the turn of the century, more than four-fifths of the world's most heavily-populated agglomerations are projected to be in the developing world (World Bank (1991), p. 3 and p. 22).

[2/] The regulatory approach on quantity -- the so-called command and control measures -- suffers from its inability to provide correct market signals to induce the most efficient trips to be undertaken. In contrast to pricing incentives, it generates virtually no revenues for the public sector (see Pozdena, Schmidt, and Martin (1990)).

[3/] The total cost of undertaking a trip -- both 'private' and 'social' -- involves: 1) congestion (which is borne by road users); 2) pavement wear (which is covered by the road authority); 3) air and noise pollution; and 4) costs of accidents (both of which are borne by society at large); in addition to 5) vehicle operating costs. 'Social costs' (i.e., both internal and external costs) here are distinguished from 'private costs' (i.e., internal costs), the latter of which are self-financed. Those traveller-borne internal costs include: 1) vehicle operating costs as well as 2) time costs and delay in congested traffic (Newbery's definition of road use costs refers only to the external costs.) This paper deals mainly with the marginal cost pricing of congestion and pavement wear charges as opposed to marginal *social* cost pricing, which is defined to include -- in addition to private costs -- all the *external* costs of congestion, air pollution, noise pollution, accidents, road damages and externalities. Hence, strictly speaking, 'congestion pricing' refers solely to the pricing of congestion externality whereas 'road pricing' refers more generally to the pricing of all externalities from mobile sources.

based on the works of Herbert Mohring, Robert Strotz, William Vickrey, Alan Walters, Theodore Keeler, Kenneth Small, Clifford Winston and David Newbery.[4/] It aims to integrate their ideas and principles into a single analytical framework. I am convinced that the presentation of a rigorous and unified diagrammatic -- but non-mathematical -- framework derived from first economic principles casts important light on congestion pricing systems and on issues surrounding short-run and long-run marginal cost pricing, scale economies and diseconomies, indivisibilities, the peak-load problem in urban transport, optimal road durability, financial viability and cost recovery in the public provision of road services. While there are several automatic road user charging and electronic toll collection systems in use in parts of Norway, Italy, France, and the United States, as well as bills in parliament to implement various forms of road pricing in Santiago and Stockholm, one *congestion* pricing system is currently operating (Hau (1992)). That is Singapore's Area Licensing Scheme, which is now in the process of being converted into the Electronic Road Pricing System, to be operational in 1995 (see Watson and Holland (1976, 1978), Behbehani, Pendakur and Armstrong-Wright (1984)). Even so, the charging of vehicles by daylight hours in Trondheim, Norway (with further differential pricing for electronic tag users) since 1991 could be regarded as a rudimentary form of congestion pricing.

3. Recent technological breakthroughs in automatic road user charging have brought electronic road pricing much closer to reality. Thus, the second paper (Hau (1992)) in this research presents a taxonomy of alternative technologies of congestion pricing. It drops the crucial assumption that the implementation of short-run marginal cost pricing is costless (Hau (1990)). In particular, it compares and contrasts the methods of manual versus electronic charging schemes. Electronic approaches are analyzed progressively by discussing first, the increasingly popular electronic toll collection mechanisms which incorporate recent technological advances in automatic vehicle identification -- commonly known as AVI -- as well as smart cards. The appropriateness and applicability of electronic toll collection mechanisms to electronic

[4/] I owe much intellectual debt to these authors and I am very grateful for the opportunity of discussing my work with several of them.

road pricing are then dealt with. The companion paper also analyzes the relative cost-effectiveness of each technology and performs benefit-cost analysis where data permits. The implications of using each of these technologies for relieving congestion are discussed and policy recommendations are drawn.

I. Conceptual Guidelines

4. Rising real incomes result in increased aspirations for the ownership of private automobiles. Barring major restraint measures, an increasing number of motor vehicles means that travel demand swells concomitantly. Because municipalities are finding it increasingly difficult to finance new road construction and improvements, the rate of growth of travel demand outstrips the growth of road capacity (Hau (1988)). Transport planners, with limited options available, find it very difficult to combat effectively mounting traffic problems in the face of increasing urbanization. The resulting traffic explosion is an illustration of Parkinson's Law or Downs' law of peak-hour expressway congestion, in which commuter traffic ascends rapidly to the level of new capacity in urban areas (Downs (1962), Mohring (1965)). Traffic engineers have long been familiar with this "fundamental law of highway congestion" in which latent demand expands to fill the gap created whenever highway capacity is improved (Small, Winston and Evans (1989), p. 85). In this section, I set up the conceptual guidelines that allow authorities to curtail traffic congestion in an efficacious manner at the same time as satisfying the World Bank's general guidelines for public sector projects and urban transport policy (World Bank's Operational Manual Statement No. 2.25 (1977), World Bank's Urban Transport Policy Paper (1986), and Julius and Alicbusan (1989)).

5. In a nutshell, the essential principles include:

1) implementing short-run marginal cost pricing (short-run efficiency) to generate maximum net benefits for society: *efficiency pricing*;

2) undertaking investment in infrastructure whenever the additional benefits exceed the costs (long-run efficiency) to society of doing so: *economic viability*;

3) investing in transport services when benefits exceed costs; promoting public transport services especially when revenues exceed costs: *financial viability*;

4) maintaining 'fairness' among beneficiaries, for example, via benefit taxation -- *equity* -- where possible; and

5) using pricing and cost recovery policies to improve the efficiency of managing the public sector -- *cost-effectiveness* and *managerial efficiency* -- if possible.

II. Cause of Congestion: Public/Private Ownership of Roads

6. Roads which are infrequently utilized possess the characteristic of nonrival consumption among users and are traditional examples of public goods. Joint consumption means that roads yield services that are simultaneously enjoyed by more than one user, without substantial detriment to the satisfaction of others. If roads are totally nonrivalrous, then neoclassical economic principles dictate that roads ought to be provided for by the public sector and financed from general revenue taxation (and perhaps land value taxation), fully taking into account the social opportunity cost of public funds. On the other hand, roads which are heavily utilized have the nature of rival consumption among users and are called congested public goods. These variable-use congested public facilities then approximate to varying degrees the rival nature of private goods. Private goods are of course provided contingent upon payment, excluding those who are not willing to pay for them. Alas, with free access to roads, people are not barred from the use of scarce services, resulting in overuse. Hence market failure due to nonexcludability calls for governmental intervention in the form of better designed road user charges and motivating charges which correct for externalities. (The failure of the voluntary pricing mechanism due to nonexcludability is referred to as the 'free-rider problem' in the public finance literature (Boadway and Wildasin (1984), Chapter 3)). Note that it is the traditional inability to exclude motorists from the use of crowded streets that is *the* cause of market failure. However, when refinements in automatic vehicle identification technology become even more cost-effective, automatic road user charging means that market failure is somewhat overcome. This is because prices are then able to reflect both the intensity of demand and the true economic cost of road use. The problems of nonexcludability and preference revelation would then be virtually

eliminated and the advantages of price incentives reaped. Thus the standard public finance text (see, for example, Musgrave and Musgrave (1989, Chapter 4)) argues that both the nonrival consumption characteristic of public goods and the nonexcludability -- or rather, costly excludability -- of congested public goods are causes of market failure, calling for government intervention. Hence, roads which possess the attribute of congested public goods , and thus have a partially rival consumption characteristic, ought to be treated by the relevant governmental authorities as mixed or impure public goods, if not private or club goods (Buchanan (1965)).

7. The foregoing discussion leads naturally to the transport economist's definition of congestion. That is, as more and more vehicles join a traffic stream, the travel times of all motorists making trips are raised, resulting in delay to all. In essence, the congestion phenomenon is one of excess demand, given a fixed road network.

8. Yet the fundamental reason why congestion occurs so ubiquitously is because property rights are not clearly delineated, yielding market failure. If roads are both privately owned and competition prevails in the provision of roads, usage would be (Pareto) optimal in the absence of scale economies (Knight's conjecture (1924)).[5/] If roads are not privately owned, access to and usage of roads are effectively free to the traveller, resulting in excessive use of those roads where traffic is heavy enough to produce significant adverse interaction among vehicles. (The

[5/] We know that competition results in economic efficiency. In the context of roads, owners maximize the return to their land by setting the price equal to the short-run marginal cost of production, where the difference between the short-run marginal cost and the average variable cost represents the quasi-rent to the fixed factor of production. (The notion of quasi-rent is discussed in footnote 29.) This is the analog to society mimicking the decentralized competitive landowner's rent maximizing behavior by charging the optimal (Pigouvian) toll -- the divergence between marginal cost and average variable cost (see discussion in Section VIII and the proof of this conjecture in Vickrey (1968)). Knight further argues that the case of increasing returns is dealt with by the market allowing only one agent to remain behind to exhaust scale economies (see Walters' (1954) interpretation of Knight's conjecture (1924), which does not accord with my interpretation here). I argue that the government is left with the role of being the agent that charges the congestion toll, in order to prevent monopoly pricing. After all, it is unclear *a priori* that the informational requirements and transaction costs of public ownership would exceed the social costs (such as rent-seeking) of regulating private ownership. In other words, it is debatable whether government failure is necessarily greater than market failure, especially in developing countries. For instance, the government's power of eminent domain over public projects may reduce its land acquisition and road construction costs whereas the common property nature of roads would encourage rent dissipation.

overcrowding occurs *despite* the fact that motorists already have to pay fuel taxes and registration/license fees and purchase taxes for owning a vehicle, which can be regarded as entry fees only.)[6/] On the other hand, it is unlikely that private ownership of roads would yield a perfectly competitive market structure because alternative routes are in general not perfect substitutes. Hence, those road owners might exploit their locational attributes and monopolistic prerogatives by raising charges above marginal cost. In addition, incentives would arise for collusion. Thus competition would at best be imperfect in the case of road provision where lumpiness is found. The relevant comparison, based on efficiency consideration, is the welfare cost associated with government intervention vis-a-vis the dead-weight loss associated with the monopolistic or oligopolistic pricing practices of private ownership. In addition, it seems reasonable to argue that there are other grounds -- such as equity -- for regarding landowners' super-normal profits, or collusive tendencies, as undesirable.

9. In practice, virtually all roads belong to the government and not to private owners. Because of free-access roads, one should not be surprised to be confronted with the pervasiveness of congestion. This is the common property resource problem, which yields an externality.[7/] One might ask whether or not some congestion-prone roads ought to be privatized to internalize the externality. The key factor is whether (sufficient) competition could be made to prevail to ensure, in the absence of regulatory restraints, that these roads are not overpriced. A testable

[6/] The fuel taxes which motorists pay are generally used to contribute to: 1) the maintenance and operating (and in part the investment) costs of roads, 2) debt service on highway bonds, and/or 3) general tax revenues. (Independent of how fuel tax revenues are disposed of, the main problem of the uniform fuel tax is its inability to vary by time and place of use (see Hau (1992)).) The combination of a fuel tax, first registration taxes and annual license fees can be regarded as, to a rough order of approximation, a two-part or multi-part tariff.

[7/] As long as property rights are well-defined, exchangeable and enforceable, private bargaining would internalize all 'externalities' and yield an efficient allocation of resources independent of who is held liable for creating external effects in the first place (Coase (1960, 1988)). Since Coase's so-called theorem holds only if the costs of transactions are nil and income effects are negligible, it applies mainly to the case of small numbers. With congestion and mobile-source externalities, I argue that the large numbers of people affected clearly result in large transactions costs, and hence market failure, beckoning government intervention. Further, Cooter (1982) argues that Coase does not distinguish between zero transactions costs and zero bargaining costs, and that the generality of the "Coase Theorem" is overstated.

hypothesis in contestable markets theory is to check whether the condition of potential competition and not actual competition is satisfied (Baumol, Panzar and Willig (1982, Chapter 1)). One surmises that it may be difficult to assume that it would be effective in practice, especially in the case of roads where irreversibility and lumpiness exists (Baumol and Lee (1991)). Also, noncomparability may occur because a competitor would find it difficult to duplicate a road on the same alignment and would likely settle for inferior alignment. Economic efficiency could be enhanced in some cases if there were a *mixture* of both private and public roads, since the existence of publicly-provided *free*ways might limit the degree to which those who use private expressways would be charged monopoly prices. Similarly, private toll roads could serve to over-zealous governments from overcharging on public toll roads.[8/]

10. In short, the congested public good nature of roads suggests that these facilities should be treated more like club goods or private goods, yet the status quo appears to suggest that public ownership of roads is institutionally preferable. If so, municipalities ought to simulate the workings of a competitive, private firm and industry by setting congestion tolls on public roads to internalize the congestion externality. This thereby deals with the problems of nonexcludability and preference revelation explored earlier. If a private ownership arrangement is deemed beneficial at times, the government could exercise its power of eminent domain in reverse by altering the structure of property rights of the relevant facilities from public to private ones. This could be done, for example, by auctioning off government land to the highest bidder (and simultaneously reducing general taxes and/or providing more public goods), or, alternatively, by designating the competitive provision of highway services to corporations or private contractors via "franchise bidding" (Mills (1981)). I have argued, on both efficiency and equity grounds -- following our conceptual guidelines -- that the facility's ownership arrangement ought

[8/] Note that a relatively uncongested toll road would yield inefficient traffic allocation if its effect is to price traffic off to alternative public roads which are already congested. See Johansen (1989) and Catling and Roth (1987) for a discussion of toll roads and road privatization.

to reside with the government, with the proviso of setting cost-effective congestion tolls.[9/] I next show how a congestion toll can be derived from a transportation engineering speed-flow curve. In the process, the equivalence of efficient congestion tolling and the short-run marginal cost pricing of vehicle flow is established.

III. Foundations of Road Congestion -- The Classical Case (in the Short Run)

11. Road congestion is well founded in the economics and transportation engineering literature (Dupuit (1844, 1849), Pigou (1920) and Walters (1961a, 1961b)).[10/] It begins by considering a representative driver cruising under low traffic conditions along a given stretch of urban road with fixed beginning and end points. The representative driver would be able to achieve a mean free speed that balances the benefits to him of a faster trip against the costs to him of higher energy requirements and a greater risk of an accident. *Ceteris paribus*, as other vehicles enter the road thereafter, density increases, speed drops and travel time (or delay) lengthens (and accident probability rises). The causality is as follows: traffic density determines speed and not vice versa. Parallelling the theory of fluid dynamics, traffic flow is the product of density, in vehicles per kilometer, and speed, in kilometers per hour. Note that the rectangular area in Fig. 1(a) -- the speed-concentration relationship -- is equivalent to traffic flow, expressed in vehicles per hour (see Gerlough and Huber (1975, Fig. 4.12) and May (1990, Chapters 7, 10), for example). Hence, traffic flow is the product of traffic density and speed, with traffic flow attaining a maximum at F^{max} with speed at S^{m} in Fig. 1(b) -- the speed-flow relationship (see Haight's (1963) "fundamental diagram of road traffic" -- a flow-density curve -- and similar

[9/] While I am convinced of the advantages of market forces, I have some reservations about their extent. For example, some have argued that many social services should be provided for by the market, and also that transport infrastructure should also be privatized (Roth (1987, Chapter 6) and Catling and Roth (1987)). It is difficult to see, in the absence of road divisibility and competitive forces, how marginal cost pricing would be pursued and maintained by private road owners. From a public choice perspective, the same argument could be reversed and applied with respect to governmental authorities. This explains why I argue that perhaps a mixed system of public/private ownership/regulation would be able to reap the strength of both.

[10/] While Walters (1968, pp. 31-34, Chapter 3 Annex) discusses optimal investment, his works emphasize the short run character of road pricing. The approach taken in this paper is to highlight both the short-and-long-run nature of the congestion problem in order to explore the issue of cost recovery.

figures in Morrison (1986)). (Touted maximum flow figures of the 'capacity' for a typical expressway are about 1800-2000 vehicles per lane-hour at 50-55 kilometers per hour (30-35 mph) (see Gerlough and Huber (1975, Chapter 4) and any Highway Capacity Manual, for example, Transportation Research Board (1985, Chapters 2-3)).[11/]

12. Given a fixed distance of say a kilometer of road, the traffic engineer's speed-flow curve can be straightforwardly converted to a travel time-flow curve as travel time is the reciprocal of speed, with vehicles-kilometer per lane-kilometer-hour on the horizontal axis (see Fig. 1(c)). Using a constant value of time as a shadow price for the representative driver, travel time is then converted to a money basis which yields time cost, called the average variable cost, AVC (see Fig. 2). Low traffic volume corresponds with relatively high speed, so fuel cost would be high. With high traffic flow and low speed, however, fuel cost would also be kept high because of fuel inefficiencies caused by the alternate acceleration and deceleration associated with dense traffic. These two factors roughly cancel one another out, leading to the plausible assumption that the costs of operating an automobile (which include fuel, oil, maintenance and depreciation costs) are approximately independent of the level of traffic flow (Mohring (1976, Chapter 3), AASHO's (1960) 'Red Book'). A fixed money cost for the vehicle operating cost can therefore be added to the time cost portion to form the generalized cost - an accepted construct of transport economists (Nash (1976) and Button (1982)). Similarly, the road's variable maintenance cost, which is assumed to be proportional to the traffic level, following Walters (1968, p. 24), can be added up also (see Fig. 2).[12/] So it is the time cost element that is mainly responsible for the upward-sloping portion of the AVC curve. The AVC curve climbs upwards because significant negative interactions occur before traffic reaches maximum basic capacity, Q^{max}; it is variable in the sense that as traffic flow, Q, is increased, congestion delay actually sets in rapidly

[11/] Economists' notion of 'capacity' is whenever congestion delay begins, which is considerably less than traffic engineers' concept. Further, vehicular flow is sometimes normalized by the capacity of a certain road to yield a 'volume-capacity' ratio.

[12/] Hence the marginal cost curve, MC, when summed up vertically with the vehicle operating cost and variable road maintenance cost, yields a translated marginal cost curve, *MC*. The same notations apply to the average variable cost curve.

at substantially below the traffic level Q^{max} (contrary to the engineering notion of a constant average variable cost curve extending up to the point $Q^{max} \equiv F^{max}$). After the engineering or basic capacity, $Q^{max,}$ is reached, AVC becomes an 'inverse supply' curve. Note that the standard supply curve is nonexistent in the context of roads.[13/]

IV. DEMAND AND SUPPLY

13. The 'supply' side can be made to be congruent with the demand side when a conventional demand curve is specified to depend on the travel cost (price) facing a traveller for a single trip. When an initial demand function, Q^d, intersects the AVC curve at point U (Fig. 2), a (stable) equilibrium is said to exist at Q^o.[14/] This is an equilibrium point because travellers' willingness-to-pay curve, i.e., the inverse demand function, equals the average variable cost curve -- the function upon which travellers base their travel decisions. After a small excursion of the demand in the neighborhood of the equilibrium point, unfettered use results in a return to the observed traffic level, Q^o, hence the equilibrium is considered stable.

14. Basic price theory says that whenever the average variable cost rises, it means the marginal cost curve lies above it.[15/] The vertical difference between the two cost curves is

[13/] The solid portion of the speed-flow curve and travel time-flow curves in Figs. 1 and 2 denote the 'normal' part of the curves. The non-solid backward-bending portion of the cost curve (Fig. 2) means that time cost increases because traffic flow is reduced after the engineering capacity is reached. The backward-bending curve, far from being fictitious, has been substantiated in the literature (Gerlough and Huber (1975, Chapter 4), Keeler, Small and Associates (1975, Fig. 1)). Schiff (1991) explores this backward-bending case.

[14/] The demand function, Q^d, is an observed, constant-money-income Marshallian demand curve, with the usual regularity conditions. It approximates the exact Hicksian demand curve, which yields the true reflection of willingness-to-pay and marginal benefit. Formally, at equilibrium $GC(Q) = Q^d(P)$, and $Q^d(GC(Q^o)) = Q^o$, where GC is the generalized cost. Since GC is simply a translation of AVC, the interpretation of one is synonymous with the other.

[15/] Marginal cost is obtained as follows: $MC \equiv \Delta C(Q)/\Delta Q = AVC(Q) + Q \cdot \Delta AVC(Q)/\Delta Q = AVC(Q) \cdot (1 + \varepsilon)$ where $C(Q)$ is the cost function, ε is the elasticity of the AVC curve, i.e., the rate at which time cost rises with respect to a one percent rise in traffic flow (Walters (1961)). The first term composes of only time cost and the second term is the marginal (external) congestion cost, set equal to the congestion toll. Marginal cost pricing of a trip, P, is achieved by setting P = MC. This is known as the first-best optimal pricing rule: our first optimality rule. (Note that AVC depends parametrically on the capacity level K, and can be expressed as AVC (Q, K).

the marginal (external) congestion cost -- the additional delay that one driver imposes on the rest -- which is not taken into account by the last driver who joins the traffic stream. In fact, since each driver chooses whether or not to travel according to the AVC curve -- being the decision curve -- he or she totally ignores the resulting external congestion cost imposed on fellow motorists. We thus have the optimal point Y at which the marginal cost curve intersects the (peak) demand curve in Fig. 2. In other words, Q' is the associated optimal output in the sense that the generalized cost which includes external congestion cost and other variable costs (i.e., constant (unit) operating cost of a vehicle and variable maintenance cost of a road), is equated to the price. The congestion cost is the additional time cost that a motorist imposes on others, calculated by taking the increment in average time cost caused by the added trips and multiplied by the number of vehicles in the traffic stream. The Pigouvian tax applied to roads is that optimal toll which closes the wedge between the marginal cost and average variable cost curves by emitting the correct signal and creating appropriate (dis)incentives.

15. This Pigouvian toll-tax is equal to the marginal external congestion cost. It is also known as the net-benefit maximizing, economically efficient, (Pareto) optimal and marginal cost toll. (If one prefers, it can be regarded in graphic terms as a sin tax, even though it has not yet included the cost of environmental and other externalities.) Hence the marginal cost pricing of trips in the short run (given that a road is fixed) yields a first-best Pareto optimal allocation of resources. The optimal road user charge is then comprised of a congestion toll and another component which covers the variable road maintenance cost (see Fig. 2's legend).

Without loss of generality, the inclusion of the average vehicle operating cost and variable road maintenance cost -- both being constant with respect to traffic -- simply alters both the left-hand side and the right-hand side by the same amount). Implementing marginal cost pricing in this case means pricing the *difference* between marginal cost and the average variable cost of a trip, plus a component which covers the cost which the motorist imposes on the community (see Fig. 2's legend). We shall return to an intuitive discussion of this subtle but important distinction. Note further that marginal cost rises asymptotically to the engineering capacity level of Q^{max} and is undefined for the AVC curve at points beyond V.

V. VERY CONGESTED ROADS AND BOTTLENECKS

16. Observe that at the equilibrium point U in Fig. 2, the resultant throughput is significantly less than the road's maximal flow capacity of point V. The backward-bending 'supply curve' exists because a one percent increase in density results in more than a one percent decrease in speed when very dense traffic is reached. (That point X is a stable equilibrium point can be seen intuitively by perturbing the price level while point w is unstable in that any disturbance will result in a movement to U or X.) The point here is that quite a few cities, for example, Bangkok, Budapest, Buenos Aires, Hong Kong, Jakarta, Mexico City, Santiago, Sao Paulo, Seoul, Singapore and Taipei are faced with extremely congested situations such as point X, certainly during peak of the peak.

17. Consider the dynamic phenomenon of traffic growth as shown in Fig. 2 (a). Suppose the initial demand curve intersects the average variable cost curve at point 1. As traffic grows, the observed number of vehicles per lane-hour increases from point 1 to 2 to 3 (which is identical to point U of Fig. 2). A further increase in demand beyond point 4 would result in a discontinuous jump to the backward-bending part of the unit variable cost curve at point 5. Intuitively, traffic congestion worsens rapidly and 'jams up' all of a sudden at times. After a while at this point, travel demand would start to slack off to points 6 and 7 (which is identical to point X of Fig. 2). Note that traffic would be moving at a snail's pace as queuing develops. This corresponds diagrammatically to the positive sloping part -- as opposed to the relatively smooth traffic of the 'normal' downward sloping portion -- of the speed-flow curve, Fig. 1(b). As travel demand continues to diminish towards the end of a rush hour, say, the traffic level would touch point 2 briefly but would then end up at point 8. Thereafter, further slackening of demand results in traffic returning to the upward sloping portion of the unit cost curve at point 1, say. This is known as a relaxation phenomenon in an engineering and physics context. Therefore, the points between 4 and 8, such as the equilibrium point W, are never reached in reality![16/] We conclude that it is precisely the daily recurrent peaking characteristic of travel

[16/] In other contexts such as Walters (1987), point W is described as an unstable equilibrium.

demand that calls for innovative solutions. That the traffic level at those rush hour periods seems to be near gridlock -- an illustration of 'hypercongestion' -- is too important a case to be ignored (Walters (1987)).

VI. The Welfare Impact of Road Pricing

18. The purpose of this section is to highlight one of the main points of this paper, the proof of which is relegated to the appendix. Economists *know* that road pricing results in improvement in welfare to society, yet politicians and the public almost unanimously regard it with skepticism. Why? To economists, the increase in welfare comes about because of the imposition of an externality-corrective (toll-)tax. Yet, for those motorists remaining on the road, the congestion toll is similar to a tax increase. Under conditions of 'normal' traffic (or non-hypercongested situations), note that the toll-tax paid by the motorist exceeds the valuation of time savings resulting from road pricing on average, so that the 'tolled' is worse off.[17/] Those who are priced off the road to an inferior mode or time of travel in order to avoid paying the toll are also worse off, while those who remain on other indirectly impacted roads are either worse off, if congestion arises there, or just as well off, if there is no resulting congestion. It turns out that the government, in collecting toll revenues, becomes the main party that is better off.

VII. The Effects on the Tolled, the Tolled Off and the Untolled

19. In the appendix, I establish the proposition that marginal cost pricing of trips maximizes the net benefit to society in the sense that a Pareto-efficient situation is attained, that is, no one can be made better off without making someone else worse off. The quantity approach (the 'American' approach) yields the welfare gain due to road pricing of area *l* (shown by the triangle UYZ in Fig. 3). The equivalent area, *e+f-d*, is the welfare gain using the change in total benefits and costs approach (the 'British' approach). The latter area is the cost saving of area *e+f+g+k*

[17/] The 'tolled, the tolled off, and the un-tolled' are terms coined by Zettel and Carll (1964). My approach differentiates from both Zettel and Carll and Wohl and Hendrickson's (1984, pp. 114-116) in that I apply the standard neoclassical method of welfare analysis in the appendix to derive my results and policy implications.

less the loss of the use value of area $g+k+d$. We note that there are two groups of travellers that are clearly adversely affected by road pricing: the tolled and the tolled off. Briefly, the British approach to the calculation of welfare gain says that those who remain on the road after road pricing is introduced incur a cost by making a toll payment of the rectangular area $b+c+e+f$. On the other hand, the traveller benefits in the form of reduced travel time -- 'forcibly' induced -- and the resultant time savings is the smaller rectangular area $e+f$. Hence the consumer-traveller would regard this exchange as not getting good value for money because the traveller is still faced with a net payment of area $b+c$ relative to the no toll situation: this area is his loss of consumer's surplus. Despite such a trade, the traveller would undertake the trip because his willingness-to-pay still exceeds the price.[18/]

20. The other identifiable group includes all those marginal users whose willingness to pay is not high enough and hence are tolled off the road. As a group, their loss in valuation, the vertical trapezoidal area $d+g+k$, exceeds their saving in time cost of the area $g+k$ by the (welfare) loss in consumer's surplus of area d. So both groups are necessarily worse off vis-a-vis the original situation. This is shown to be true despite an argument that those who remain benefit from reduced time cost! They *do* benefit from reduced travel time, but they have to exchange money for time. In addition, the "tolled on" is either just as well off, or worse off, depending on whether or not congestion arises. So it appears that, a move from a Pareto inferior position to a Pareto optimal state leaves everyone worse off! How could this happen? It comes about because we have not yet accounted for one agent: the government. The toll revenue collected by the government is considered an actual gain to society in that it is counted once only. Indeed, it may have a greater value than the dollar amount itself if it replaces other (income) revenue sources with an excess burden.

21. Suppose the toll revenue is collected and then put aside. An 'efficient' amount of traffic and congestion would still exist on the road in question, yet society -- an aggregation of both gainers and losers -- would be definitely worse off. Unless the public in general, or road users

[18/] It is as if a blackmailer -- the government -- carves out part of the motorist's consumer's surplus.

in particular, can partake in the tax proceeds either in the provision of public goods and/or reduction in tax revenues, they will definitely be worse off from this imposition of an 'optimal' congestion toll. If funds are not channeled back to road users, the government, or the rest of society, gains, but only at the expense of those faced with road pricing.[19/] However, because of the nature of the Pigouvian tax, which possesses an asymmetric price signal, generators of externalities are taxed but those affected by the externality are *not* supposed to be compensated. This is a requirement of optimality in the case of *both* public and private good-type externality-corrective taxes (Baumol and Oates (1988, Chapter 4)).[20/]

22. Two choices to those who are tolled off the road are to use the road during the off-peak period or switch over to public transport during the peak. The analysis is similar in spirit for both classes, so I will illustrate it for the off-peak period, for simplicity, following the same notation used thus far. Substitution of travel demand from peak to off-peak has the effect of shifting the off-peak demand curve to the right, from Q^{d}_{op} to $Q^{d'}_{op}$ (Fig. 4). It is easy, as is sometimes done, to regard the area *m* as an additional 'benefit'. This procedure is incorrect because the area bounded by the demand curves in the substituted off-peak period is in fact a pseudo-benefit and is already accounted for by the welfare gain to road pricing in the peak period, i.e., the triangular area *l*, in Fig. 3 (Mishan (1988, Chapter 8)). Intuitively, since there are no changes in consumer's or producer's surpluses, there are *no* net changes in benefits or costs: the additional trips made during the off-peak period are entirely self-financing in the short run. If congestion were to set in during an off-peak period with relatively medium levels of congestion, say, during the inter-peak, efficiency suggests that another congestion toll level be set to internalize the congestion externality. In this way, traffic will settle to another equilibrium

[19/] The Smeed Report of 1964 and others effectively *assume* away the problem by stating that the congestion toll revenues will be returned to the population in a lump sum nondistortionary manner.

[20/] Note that Baumol & Oates (1988, 2nd edition, Chapter 4) corrects the mistake made in the earlier edition (1975, 1st ed., Chapter 3), which says that only the victims of private good-type externalities ought to be compensated for after the imposition of a Pigouvian tax. This result is contrary to accepted notions of justice. However, Baumol and Oates (1988, pp. 236-240) point out that if the tax revenues were funnelled back *indirectly,* then there would perhaps be only insignificant divergences from Pareto optimality. The idea is that wealth effects on consumption ought to be minimized.

level with a smaller congestion toll in the inter-peak period. This is the idea behind peak-load or differential pricing. A dynamic process takes place among the peak and inter-peak periods until an equilibrium is settled upon. Proper cost-benefit analysis requires that the changes in net benefits be calculated only in the periods which encounter changes in travel time due to congestion or decongestion. With a two market model, the welfare effects of road pricing in the peak period are simply repeated in the inter-peak period whenever congestion is encountered.

23. It turns out that there is a case often overlooked in which *everyone*, including the government, can be shown to be better off. This is the case of 'hypercongestion', where density is beyond the point of maximum flow. Here the traffic density is so high that both traffic flow and speed diminish, with the generalized cost P^o and traffic Q^o occurring at point A (Fig. 3(a)). (Even though it violates economic rationality to end up at such a point A, this type of traffic jam does in fact transpire fairly regularly, though limited to peak periods such as peak-of-the-peak.) The implementation of a marginal cost toll would result in travellers reverting to the normal non-hypercongested portion of the speed-flow curve (such as may be observed downstream of a bottleneck), which corresponds to the lower branch of the AVC curve at point C in Fig. 3 (a). In addition, the travellers have to pay a unit toll payment of distance BC, resulting in a generalized cost to the motorist of P′, which is still *lower* than P^o. Because of the price decrease from P^o to P′, traffic therefore increases from Q^o to Q′ correspondingly. In this case, literally *everyone* is made better off: the tolled, what I call the "tolled on," and the government.[21/] If the speed-flow curve is as depicted in Fig. 1 (b), then hypercongestion appears more often than is commonly realized: it occurs whenever speed drops to half (approximately 60%) of the maximum speed limit.

[21/] It can also be shown, using the technique of welfare analysis elaborated in the appendix, that the net benefit to society from such a rational move is equal to area $o+p+q+r+s+t+u+v$ (= area $o+p+q+r+s+t+w+x$).

Policy Implications:

24. Ever since the French engineer, Jules Dupuit (1844), introduced the powerful concept of consumer's surplus to analyze issues of the efficient pricing of (toll) roads, economists have embraced that notion, extending and deepening knowledge in that area. Efficiency analysis carefully shown here indicates that society would unequivocally gain from road pricing. The question that arises naturally is why road pricing, with one single exception worldwide, has failed in the sixties, seventies and eighties to get off the ground. I show, using the analytical framework developed above, that road pricing as sold in the past is most likely doomed to political failure. This is because almost all motorists, including both the ones who are tolled and tolled off, find that they are invariably worse off as a result, except in the case of hypercongestion (see footnote for qualification).[22/] The sole unmistakable gainer is the government. If road users do not perceive or are not persuaded that they benefit from the government's newly collected revenues in the form of provision of transport services and worthwhile public expenditures or receive transfers in the form of reduced general tax payments, albeit indirectly, it is highly unlikely that these groups would acquiesce to the pricing of existing roads. Cooperation would be more likely if they are guaranteed a reduction in motor vehicle-related taxes such as import duties, first registration taxes, annual license fees and/or fuel taxes. In particular, the replacement of -- rather than an addition to -- existing vehicle-related taxes by cost-effective congestion tolls would especially be welcome by road users.

25. A natural question then is as follows: is there a theoretical argument for dedicated funds or earmarking so that society as a whole would benefit from the implementation of road pricing? Restated another way, is there a way in which road users can act as beneficiaries and be

[22/] The policy implications discussed here are based on the assumption of constant value of time -- the assumption used by Walters and others in deriving the average variable cost curve (see section XIV, subsection 1, below). My result that road pricing makes everyone worse off except the government is meant in an average sense. People's time valuations differ in reality, so that this result is modified to say that only those with fairly high valuation of time would be better off, and everyone else still worse off. The condition is that the weighted valuation associated with the time savings rectangle of area $e+f$ must exceed their total money payment of area $b+c+e+f$ based on a weighted congestion toll. Otherwise, even those who remain on the tolled road would continue to pay but would actually be worse off relative to their status quo.

indirectly ‘compensated for’ the payment of tolls by satisfying a commonly accepted notion of fairness, while carefully skirting the first-best pricing rule? I think that the answer is ‘yes’, although not entirely without qualifications.

VIII. Short Run Equilibrium

26. The proposition that optimal pricing of and investment in a highway system parallels the short and long run equilibrium conditions of a competitive industry of a textbook commodity was first shown by Herbert Mohring (Mohring and Harwitz (1962), (Mohring (1965, 1976)).[23/] Economic analysis of transport problems is simplified considerably by explicitly recognizing the traveller as both consumer and producer, and as producer-traveller he purchases factor inputs such as travel time from himself. This short-run/long-run approach is advanced both cogently and lucidly by Mohring and used by established transportation economists such as Keeler, Small, Kraus, Glaister, Morrison, Winston and Oum.

27. In the short run, some inputs for producing a textbook commodity are regarded as fixed.[24/] Under competition, an economically efficient output level is achieved when the market-determined price equals the short-run marginal cost of producing that good. A competitive producer purchases variable inputs by hiring labor and procuring raw materials, in addition to investing in a fixed input, capital. Thereafter, the firm combines the inputs via the production process and creates commodities to be sold to a consumer. In other words, the producer uses the revenue which he obtains from selling the good at the given price to pay for the variable inputs of labor and raw materials, plus the fixed input in the form of quasi-rent on the capital equipment, normally regarded as the accounting profit. The graphs for the textbook commodity are similar to (but not exactly the same as) the case of roads considered in Figures

[23/] See survey by Winston (1985).

[24/] The standard definition of short run is a situation in which some productive inputs are regarded as fixed, and hence certain costs would be fixed. However, the definition of long run refers to a situation in which all inputs vary. In the road context, the duration of the long run depends on the rate at which, say, the size of a road and hence the basic or engineering capacity, can be varied.

2 and 3.[25/] In transport, the short-run marginal cost of a trip, which we derived rigorously from an engineering speed-flow curve, is to be set equal to the 'price' of a trip. Recall that transport is unusual in that the traveller is both a producer and a consumer. Analogous to the parallel case of a textbook commodity, the road user, when undertaking a trip, supplies some of his own variable inputs, which include vehicle operating cost and time cost. We have seen that the competitive level of trips exceeds the efficient quantity in the presence of congestion. Hence, because the quasi-rent of a highway facility would be dissipated due to free competition, an optimal toll should be imposed to capture this quasi-rent. Clearly, even though the dictum of short-run marginal cost pricing prevails in both cases, the optimal toll does *not* equal the short-run marginal cost of producing an output but is equivalent to the *difference* between marginal cost and average variable cost. This is a subtle but important distinction between transport and widgets.

28. The optimal toll is the efficient charge referred to in Mohring and Harwitz's (1962) mathematical statement of this problem and Newbery's (1989) Proposition 1. The optimal user charge is then the optimal toll plus another component required to cover the variable maintenance cost of a road discussed in Walters' (1968, p.24) and Newbery's (1989) Proposition 2 (see Fig. 2's legend).[26/] As discussed before, to focus our efforts here, the optimal user charge should ultimately include air, noise pollution, accident cost and road damage externalities.

29. Walters (1968, Chapter 2) defines the term 'user charge' as the money charge that governmental authorities levy on travellers for the congestion cost they impose on others and for the variable maintenance cost of the road incurred due to their use of that road. In the absence of congestion, the user charge covers the unit road maintenance cost component only and would be independent of the traffic level. Walters (1968, Chapter 4) then coins the term "economic

[25/] With standard goods, both the short run marginal and average variable cost curves can decline and swing upwards, whereas I have shown that both the short run marginal and average variable cost curves in transport *never* decline but only rise upwards.

[26/] More precisely, Newbery's (1989) Propositions 1 and 2 include the *invariate* maintenance cost due to weather.

user charge (EUC)" to be equivalent to the generalized cost concept or user price employed here. This latter usage might lead to a possible misunderstanding about the demand and supply side or even double- counting and is therefore avoided here.

30. To recapitulate, short run equilibrium in transport occurs when the government, in the form of a highway agency, behaves in an optimizing fashion just as a private competitive firm would were it possible to organize the industry in a competitive fashion. The optimal user charge should not be set equal to the price but to the difference between the marginal cost and the average variable cost of a trip.

IX. Convergence Towards Long-Run Equilibrium Under Constant Returns

31. So far we have confined ourselves to short-run equilibrium. The fixed cost component has been deliberately left out of the analysis of the marginal cost of a trip.[27/] The motorist is oblivious to the capital cost of a road, and his behavior is independent of it. However, from the highway agency's planning point of view, the capital cost of a road is very much taken into account. Once a highway is built, however, it is regarded as sunk. The sunk cost of a road, once incurred, is irrelevant to a planner: only current and future costs, not historical cost, serve as a correct guide to planning future investment. Since the variable road maintenance cost is assumed to be constant, the marginal cost of a trip thus remains the same.

32. In the long run, however, a highway agency can vary the fixed capital input by expressway expansion, if the investment is deemed justifiable. On the other hand, if a rural road has been built as a result of past planning errors, it can be allowed to deteriorate or be downgraded (or be even auctioned off). Expanding a road until the additional benefit equals the additional cost of building it would yield maximal net benefit to the community. How might this be done without resorting to a full-scale cost-benefit study?

[27/] The fixed cost of a firm is defined to be the minimum amount of outlay necessary to start production.

33. To see how this might be done, we introduce the fixed cost, i.e., the cost of construction, together with the ‘invariate’ maintenance, depreciation and operating costs of a road that are incurred by a governmental authority in Figure 5.[28/] We then convert the entire fixed cost into the cost per time period of a unit of capital for utilizing the flow of highway services. This is done in order to make it commensurate with the average variable cost of a trip discussed thus far. The summation of the short-run average fixed cost and the average variable cost curve yields the average total cost curve. Charging the optimal toll of the distance t′ in Figure 5 seems to be more than sufficient to cover the short-run average fixed cost of the facility. In this case, the optimal toll, t′, exceeds the short-run average fixed cost of the facility, SRAFC′, by the unit profit difference of π′. In general, there is no *a priori* reason why toll revenue collections based on short-run marginal cost pricing cannot cover the non-use-related costs of a given highway facility.

34. In the case of a textbook commodity, whenever the quasi-rent being earned by a firm’s existing capital equipment exceeds its cost, there is an incentive to expand production. Ultimately, the quasi-rent earned by the existing capital equipment would then be equal to its (fixed) cost.[29/] Putting it another way, upon seeing the existence of economic profits, other firms enter the industry also, shifting the industry supply outward, increasing output and lowering price as a result. The unrestricted mobility of resources and the entry and exit of firms serve as the instrument by which profits would be competed away in due course. When capital is freely varying, long-run equilibrium is reached when zero economic profit occurs. Equivalently stated, the quasi-rent earned on the firm’s capital equipment equals its cost, i.e., the market return of the

[28/] The term *invariate,* i.e., non-traffic related maintenance cost is found in Walters (1968, p. 23).

[29/] The quasi-rent to a firm is the accounting profits *plus* interest expense on borrowed funds, if any. The accounting profit is the firm’s total revenue less its contractual costs, including interest expense on borrowed funds, wages to labor, cost of raw materials, and rental cost of leased buildings. Accounting profit less the market return of the owner-supplied assets is the economic profit. Alternatively, the quasi-rent on invested capital is the total revenue less the variable costs, i.e., including wages, cost of raw materials and rental cost, but excluding interest expense on borrowed funds. Total revenue less total variable and fixed costs yields economic profits. The fixed cost of invested capital includes the entire opportunity cost of capital, regardless of whether funds are borrowed or not. (See the third footnote of Appendix (footnote 71) and Mohring (1976, pp. 8-11)). I am grateful to Herbert Mohring for clarifying these points.

cost of reproducing the invested capital.[30/] This condition holds under constant returns to scale, where a proportionate increase in all inputs is compatible with the same proportionate increase in outputs. Given fixed factor prices, total cost also doubles, so marginal cost remains constant in the long run. With a slight but crucial modification, this analysis carries over to the case of roads. When the quasi-rent of the existing capital stock exceeds the normal market return on the cost of reproducing the invested capital plus the highway facility's invariate maintenance and depreciation costs, new investment is expected to find its way in that road segment of the highway industry if the appropriate price signals are given. Equivalently, in the long run, if toll revenues -- which recover quasi-rents throughout time periods -- exceed the entire fixed cost of the existing facility, the highway authority would have the appropriate incentive to expand a stretch of that road until all economic profits are eroded away. As we have seen in the case of roads, the variable cost is composed of the user-supplied time and operating costs and are fully self-financing. The non-use related costs are then financed separately by the road agency via toll revenue collections. In this way, full costs are covered and there is no need to raise charges when there are constant returns to scale.

X. Optimal Investment

35. In the long run, toll revenues would then exactly cover the amortized cost of construction, invariate maintenance and depreciation costs of roads -- a powerful result first shown by Mohring and Harwitz (1962, Chapter 2) and Mohring (1965) -- under the technical conditions of constant returns to scale in road construction, maintenance and road use. Constant returns to scale intuitively means that the cost of building and maintaining an expressway is proportional to the capacity. Constant returns to road use yields an intuitive interpretation: travel time depends solely on the volume-capacity ratio. If the engineering capacity and the traffic flow

[30/] The (opportunity) cost of a resource used *here* is the highest return of it *elsewhere*, hence the cost is the market return on reproduction cost and not historical cost.

were doubled, unit travel times would remain the same.[31/] The final long-run equilibrium is shown in Fig. 6. By faithfully pursuing the policy of marginal cost pricing of a trip by charging a congestion toll -- the difference between the marginal cost and the unit variable (time) cost -- and by expanding or appropriately reducing the capacity of the road until there is zero economic profit, the output (of vehicle-kilometers per lane-km per hour) is considered optimal. At a moment in time for an existing road, output is optimal in the sense that, given the marginal-cost price, the efficient level of trips is achieved. Undertaking either more or less trips would involve lowering the net benefit to the community. In the long run, output would be 'doubly' optimal if it is the efficient level of trips for that link of road which has been optimally built. Diagrammatically, not only does the implementation of a congestion toll internalize the external congestion cost, it can be seen that the toll covers the short-run average fixed cost of the road in a stationary state. Recall that for homogeneous traffic the average fixed cost of a road is simply defined to be the difference between the average total cost and the average variable cost. Clearly, collecting a unit toll would cover the entire average fixed cost of the road and yield zero profit only because the existence of economic profit or loss serves as a quasi-market mechanism in the investment decision of whether to expand or contract the highway capacity. With zero profit, the minimum point of a short-run average total cost curve is obtained. For any given level of output, the minimum total cost of yielding this output would be obtained only if the optimal investment level in capacity had been chosen. Looking at it the other way, the optimum size of a road is obtained by drawing a locus of all the minima of the various short-run average total cost curves of different sizes (and capital costs) of highways and choosing that particular size of

[31/] *If* 1) the capital and invariate maintenance cost of highway capacity, KC, is directly proportional to the engineering capacity, K, i.e., KC (K) = *a*K, where *a* is a constant, *then* there exists constant returns to scale in highway construction (and invariate road maintenance). (In mathematical jargon, KC is homogeneous of degree one in capacity.) The engineering capacity is measured by lane-width and is treated as a continuous variable. Further, *if* 2)(a) traffic can be expressed in terms of a homogeneous unit, Q, in vehicles per lane-hour, and the time cost function AVC(Q,K) depends directly on the traffic flow but is inversely related to the capacity and (b) *if* doubling both highway capacity and traffic flow result in the travel time of a trip remaining the same, *then* there exists constant returns to road use. (Mathematically, the AVC function is homogeneous of degree zero in traffic volume and capacity.) With constant returns to road use, AVC(Q,K) can be formally rewritten as AVC(Q/K), where Q/K is the volume-capacity ratio. Since unit vehicle operating and variable road maintenance costs are both independent of the level of output, and capital cost, KC, is proportional to lane expansion, ATC(Q,K) = ATC(Q/K) holds also. These two technical conditions are crucial to Mohring and Harwitz's (1962, pp. 85-90) so-called theorem and to Keeler and Small's (1977) extension of Mohring's theorem.

the road associated with the point where the demand curve intersects its marginal cost curve. We do this because the demand reflects motorists' maximum willingness-to-pay and hence the incremental benefit of the last trip. Since the minimum points of the short-run average total costs under constant returns are of the same height, the long-run average total cost is a horizontal line tangent to all the minima. With the long-run average total cost being constant, so also is the long-run marginal cost. Hence, it is only at the minimum SRATC point that *long-run marginal cost pricing* holds.

XI. LONG-RUN VS. SHORT-RUN MARGINAL COST PRICING[32/]

36. Intuitively, the long-run marginal cost of producing a trip yields the total cost of undertaking a trip to the community when all fixed and variable inputs can be varied continuously in the long run. Proponents of long-run marginal cost pricing argue that the market return to capital investment would presumably be fully covered. Yet the equivalence of short-run and long-run marginal cost pricing holds only in certain cases, including the static demand and single period case considered here. As shown in Fig. 6, long-run marginal cost pricing would cover *all* the variable costs, including time cost, vehicle operating cost and variable road maintenance cost, *plus* the fixed construction, invariate maintenance, depreciation and operating costs of the road. In fact, short-run marginal cost pricing covers the entire capital cost of the facility just as much as long-run marginal cost pricing does, as can be seen diagrammatically in Fig. 6. After all, both the short-run and long-run marginal and average costs are equal in the long run, with both sets of cost curves intersecting the demand curve at the same point. However, if a road is not optimally constructed but underbuilt, then long-run marginal cost pricing would send out too low a price signal, thereby exacerbating congestion. Short-run marginal cost pricing, on the other hand, would give the correct signal of higher willingness-to-

[32/] Prest (1969, p.8), Walters (1968, p.33), Bennathan and Walters (1979, p.33) and Bird (1976, pp.33-39) argue for short-run marginal cost pricing whereas others argue for long-run marginal cost pricing. Since the issue of long-run vs. short-run marginal cost pricing has been with us for some time, a clarification is in order (see the recent debate between Jordan (1983a, 1983b, 1985) and Vickrey (1985)). Vickrey points out that the concept of long-run marginal cost becomes obfuscated when several demand periods, e.g., peak, interpeak and off-peak, occur diurnally, given the case of a transportation infrastructure that has already been constructed.

pay and also yield positive toll revenues and economic profits as a by-product. Short-run marginal cost pricing is the rule to use whenever long-run equilibrium is not reached (and of course when it is). Looking at it another way, if short-run marginal cost is below long-run marginal cost at the current output, it means that the road has been overbuilt. But, of course, this does not mean that the size of the expressway should be or indeed can be varied instantaneously whenever demand fluctuates daily. Rather, it means that the price ought to be varied according to demand patterns using short-run marginal cost pricing.

37. Indeed, Professor William Vickrey has emphatically argued that there can be *no* solution to the urban transportation problem without peak-load pricing. Proper time-of-day pricing can be implemented only using short-run marginal cost. (We shall explore this point further in the section on demand variability.) Pursuing short-run marginal cost pricing period by period by varying road capacity incrementally over time would not only guarantee the best use of society's resources but would also enable road agencies to recover all costs -- as an incidental by-product -- in the long run. It is therefore recommended that short-run marginal cost pricing be used since the concept of long-run marginal cost cannot be unambiguously defined whenever cyclical variations in demand are involved.

XII. Trade-off between Individuals' Time and Treasury Accounts

38. Another way of obtaining the optimal investment level for roads is to answer the following question: what is the minimum cost to the community of road building, taking into account *both* the highway agency's desire to minimize the fixed cost of capital facilities and the travelling public's desire to save time?[33/] By minimizing the total cost -- the sum of these two costs: the variable (time) cost of trip makers and the fixed cost of the governmental authority -- a trade-off is found between individuals' time and the treasury's accounts. Given a non-optimal capital stock (K′) associated with a particular highway, as in the previous graph,

[33/] Solving the problem of cost minimization is equivalent to solving the problem of maximization of net benefit to the community.

Fig. 5, it can be seen that the least cost for the community involves having a road that is too small, for the level of demand depicted. Expanding the capacity of the road may reduce the user's trip cost evaluated at a given traffic level. Long-run equilibrium is reached when the minimum point of the short-run average total cost curve (equals the short-run marginal cost curve) intersects the demand curve. For the governmental authority, road capacity is a choice variable. By increasing its size, the volume-capacity ratio drops in the short run, and so does time cost. However, the cost of road capacity increases. Intuitively, the highway agency continues to expand the road until the marginal benefit from saving users' time costs is just offset by the marginal cost of one unit of capacity.[34/] It is at the output, Q* with an optimally built road K*, in Fig. 6 that the valuation of the last trip taken just equals its marginal cost, that is, the incremental cost of the trip to others, the motorist's own time cost in congested traffic, plus the vehicle operating cost and the road maintenance cost.[35/] The highway agency, by setting an optimal road user charge which is equal to the congestion toll and the variable road maintenance cost component, would be able to induce the motorist to travel up to the point where the price of a trip equals its short-run marginal cost. By pursuing this pricing policy for each stretch of road, the use of a non-optimal, existing highway network would be optimized. Further, by expanding highway capacity up to the point where the quasi-rent of each capital facility just covers the cost of reproducing it, with zero (economic) profit remaining, the net benefit to the community would be maximized. *By symmetry*, abandoning or downgrading roads is necessary when economic losses occur. The decision of not maintaining roads is tantamount to the act of disinvesting roads.

[34/] Formally, given a particular level of output, the cost-minimizing authority would expand the road up to the point where the marginal valuation in time savings due to a unit increase in capacity, $- Q \cdot \Delta AVC(Q,K)/\Delta K$, equals to the marginal cost of a unit of capacity, R. R is the rental cost per time period of capacity, which includes the invariate maintenance and other operating costs of a road, depreciation and imputed interest on invested capital. The negative sign would offset the inverse relationship of AVC and K, yielding a positive magnitude for the entire term. Alternatively, the road is to be expanded up to the point where the marginal external congestion cost just offsets the marginal cost of investment in capacity. This is the second optimality rule: the optimal investment in capacity rule.

[35/] The superscript * symbol indicates that that variable is optimized.

39. Henceforth, to simplify both our discussion and the diagrams, we ignore the individual's vehicle operating cost and the variable road maintenance cost since they are self-financing.[36] Note that our conclusions thus far hold under the assumption of constant returns to scale and perfect divisibility of roads. Consider a three-lane road with capacity K_3 in Fig. 7(a), where output is now measured in vehicles per hour.[37] Since the highway authority has efficiently priced the road by setting the congestion toll t_3* and optimally built the road by investing at K_3*, it can be seen that the toll revenue covers the fixed cost of the road. So far we have simply translated Fig. 6 into Fig. 7(a) with the costs borne by motorists conveniently left out but not forgotten. Assume that both traffic volume and road width, i.e., the number of lanes, are doubled and that the inputs to each of the component costs under constant returns are doubled, then i) the fixed costs of construction, maintenance and depreciation, ii) the variable time costs, and iii) the total costs are all doubled.[38] Intuitively, the geometric doubling of the rectangular areas of road construction and maintenance costs is synonymous with the condition of zero economies of scale (given fixed input prices) in road construction and invariate maintenance. Similarly, the horizontal doubling of the rectangular areas of the time costs supplied by individual users is akin to the technical condition of zero economies of scale in road use.

[36] I have therefore grouped the non-traffic related maintenance and operating costs as part of the short-run average fixed cost curve. Bear in mind that the variable road maintenance cost (not drawn in Fig. 7(a)) is being recovered by a separate road user charge component as shown in Fig. 2. With the condition of zero economies of scale in total road maintenance, the total maintenance cost is thus exactly doubled. Also, economic profits from now on will be referred to as 'profits'.

[37] Three lane roads are found in Australia where they are referred to as, 'two-and-a-half lane roads' by Hoban (1987).

[38] The stringent assumption of a road being finely divisible will be relaxed and indivisibilities introduced later on.

XIII. First-Best Optimal Pricing and Investment Rules

Empirical Considerations:

40. By estimating behavioral travel demand functions using state-of-the-art logit mode choice models, one could obtain empirical estimates of marginal valuation of time (as a function of income levels) (Hau (1986)). When combined with a fine-grained, parametric transportation corridor supply model of the San Francisco Bay Area (Talvitie and Associates (1978)), multi-market demand and supply could be equilibrated and cost-benefit analysis of alternative policies performed (Hau (1987)). Heuristically, the adjustment process towards the final equilibrium parallels that of the cobweb equilibrium model of adjustment. Optimal tolls and welfare gains and losses could thus be simulated with appropriate specification of the marginal travel time function. The results obtained are for a short-run equilibrium model of demand and supply.

41. Even with poor data, one could make some progress in empirical work. Given an estimate of a speed-flow curve and the corresponding travel-time flow curve, we know how these engineering curves can be converted to a short-run average variable cost curve of a trip, using an estimate of the value of time. A ‘supply’ elasticity estimate together with a value for unit variable cost would yield a one-to-one correspondence between the short-run average variable cost and the marginal cost (see footnote 15 for formula). A rough estimate of the demand elasticity and the traffic level of a particular road would yield a first order approximation of the proper congestion toll. Now, in order to maximize aggregate net benefit, two operating rules should be followed by the road authority.

The First Rule -- The Optimal Pricing Rule: For each stretch of road, short-run marginal cost pricing is fulfilled by setting a toll at the excess of short-run marginal cost over short-run average variable cost. The intuition is that this congestion toll would serve to internalize the congestion cost that a driver imposes on others. In addition, the motorist is charged another component which covers the variable maintenance and operating costs of a road which he imposes on the road authority. Thus the public authority’s imposition of an optimal road user charge would cover both the external cost of congestion as well as the variable road maintenance and operating costs.

The Second Rule -- The Optimal Capacity Rule: Under constant returns to scale and optimal pricing, whenever economic profit is found in the operation of one road link, the procedure would be to expand the capacity of that stretch of road. The existence of a loss under short-run marginal cost pricing suggests that the road has been overbuilt. By altering the capacity of each road in the long run according to the quasi-market signal of profits and losses, the entire highway network's investment level in capacity would be optimized, with the fixed cost of each road covered. Alternatively, the road authority, by trading its direct resource costs against individuals' travel time, follows the rule of setting the marginal travel time savings equal to the marginal cost of investment for an additional unit of capacity. The capacity of a road is expanded until the marginal capital cost equals the marginal (external) congestion cost.

Perspective on the Result:

42. What we have described is the long-run equilibrium of an optimally designed capacity of a road network under constant returns. If the road authority were: 1) to pursue the efficiency-enhancing policy of pricing according to the marginal cost of a trip, and 2) to minimize the sum of the direct resource costs of providing a road and the value of user-supplied travel time inputs, then the road would be both efficiently utilized and optimally expanded. Notice that the optimally designed road has a positive amount of external congestion cost. This results from the road agency's desire to minimize both the sum of the direct cost of the investment in capacity and individual drivers' travel time cost. In our simple framework, congestion delay would never be entirely absent, contrary to what environmentalists and road users would prefer, because achieving zero congestion is very costly to the community. In other words, an optimal amount of congestion externality is a valid concept, just as an optimal amount of pollution has long been recognized in the environmental economics literature. What if there is no congestion on a particular road? Zero congestion means that that stretch of road has been overbuilt (or priced non-optimally) and should perhaps be downgraded or even abandoned. If excess capacity occurs

all the time, the road possesses the non-rival consumption characteristic of a pure public good. Then we are faced squarely with the standard task of provision of public goods. If resources are plentiful, financing the shortfall via general revenue taxation has been the conventional dictum. Similarly, if public resources are scarce, the opportunity cost of public funds must be accounted for.[39/] *If* lump sum taxation is infeasible and resources are severely inadequate because of political constraints, then it is possible for one to consider the feasibility of financing via Ramsey pricing.[40/] If it is uncertain whether the (marginal) deadweight losses from general revenue financing exceed those obtained from Ramsey taxation, financial and other considerations such as equity may then have to be appealed to in order to justify the potential use of Ramsey pricing in the road sector.

43. By contrast, a road would possess the rival consumption characteristic of a private good when excess demand occurs. Hence a congested road is also regarded as a congested variable-use public facility. Because of this mixed good nature, and based on the theory derived here from first principles, the provision of road services ought to reside with the public sector. Under the condition of constant returns, the optimal toll revenue, which captures the quasi-rent earned from the invested capital, would cover the entire fixed cost of the road in the long run. No residual or overhead cost need be allocated. If profit exists, then it is because there is insufficient road capacity (or pricing at a level above marginal cost). The road is therefore not in long-run

[39/] In the companion paper (Hau (1991)), I present calculations of the opportunity cost of financing alternative charging mechanisms.

[40/] Frank Ramsey's (1927) inverse elasticity formula under the case of independent demands was discovered in response to Pigou's question of how to set tax rates in order to minimize the welfare losses associated with meeting a tax revenue requirement. The problem of the choice of optimal tax rates which are subject to a revenue constraint is formally equivalent to that of the setting of optimal prices which are budget-constrained. Ramsey pricing in the presence of externalities requires that it be computed on the basis of marginal cost and a *fraction* of the marginal external cost (Oum and Trethaway (1988)). Clearly, in the presence of leakages, non-excludability and partial rivalry in consumption, the requirements for the implementation of Ramsey pricing are considerably more stringent than those of pursuing marginal cost pricing. On conceptual, empirical and implementation grounds, marginal cost pricing is superior to even the simplest form of Ramsey pricing: the inverse elasticity rule. The computation of Ramsey prices, for instance, requires the estimation of marginal cost as a prerequisite and the determination of revenue targets. Despite all the problems with which Ramsey pricing is fraught, research into this interesting issue is potentially useful. After all, increasingly tight fiscal constraints, countries will demand alternative funding mechanisms.

equilibrium. The existence of profit serves as a surrogate market signal to expand capacity. The motto is as follows: "If a road makes money (i.e., economic profit), expand it, else not." Similarly, if a road loses money, it suggests that planners made the wrong decision or were given over-optimistic forecasts of travel demand. In that case, marginal cost pricing is still to be adhered to, with the congestion toll set close to nil. A user charge component is still needed to cover the variable road maintenance cost. Thus it may even be worthwhile to abandon a money-losing road and save on any annual invariate maintenance costs that might arise. Efficient pricing, financial viability and cost recovery are therefore entirely consistent with one another under constant returns to scale in long-run equilibrium.

XIV. RELAXATION OF ASSUMPTIONS

44. The above discussion assumes that the government aims to maximize welfare of the community by simulating the workings of a competitive industry and pricing highway services at marginal cost. There are a few major assumptions that need to be relaxed: 1) constant value of time, 2) static demand, 3) perfect divisibility, 4) constant returns to scale and 5) variability of road thickness. We consider the relaxation of each assumption in turn.

1) Differences in Time Valuation

45. The traditional presentation of road pricing and my ensuing critique assume a constant value of time (Walters (1961a)).[41/] The diagrammatic analysis in Figs. 2 and 3 implicitly assumes that every driver is identical and maintains the same time valuation. What happens when there are heterogeneous motorists, with different time valuation and tastes? A mathematical proof that generalizes the above result for homogeneous drivers to heterogeneous ones with different values of time is shown by Mohring ((1975), (1976, Chapter 4 Appendix)) and Strotz (1964a, 1964b), but the intuition behind it is not difficult. Instead of the optimal toll being based

[41/] In his pioneering work on road pricing, Walters (1961) (and authors thereafter) assumes that traffic is homogeneous, with all vehicles and drivers being the same, with the resultant identical valuation of time for all.

on a representative driver's value of time, the time value is now a weighted average of the different motorists' valuation of time, weighted by the number of trips taken by those motorists who actually use the facility. If a traveller's time value and the number of trips are close to the average, he will pay the average toll payment. If another motorist's time value is higher [lower] than average, he would be willing to pay more [less] than the average toll payment for taking a trip. He thus would be willing to, though begrudgingly, pay the difference. The congestion toll, $P' - P''$, in Fig. 3 then can be labelled the *weighted* congestion toll, and the constant value of time is re-interpreted as the *weighted* average valuation of time. For a trip with a sufficiently higher than average time value, the time saving of a trip, ($P^o - P''$), can be even higher than the *weighted* congestion toll, $P' - P''$, thus making the motorist better off. On the other hand, for a (shopping) trip having a lower time valuation, the user still has to make the average payment and therefore would be made worse off. Nevertheless, they both remain on the tolled road, as opposed to being tolled off, because their individual trips' marginal valuation or maximum willingness-to-pay still exceeds the generalized cost of their respective journeys. The use of alternative values of time would relax the point I made earlier that road pricing would make *all* groups except the government worse off. By relaxing the assumption of a constant value of time for everyone, those people with high values of time would be made better off at the expense of those with low values of time. This intuitive analysis assumes that everyone is faced with the same toll, as in the workings of a competitive economy, and that a perfectly discriminating monopolistic authority is non-existent.

46. We conclude that a single transportation facility with differences in values of time would not alter fundamentally our derived result, using the standard assumption of constant returns. Again, with efficient pricing, financial viability and full cost recovery are achievable.

2) Demand Variability and Peak-Load Pricing

47. We have in fact considered the case of variable demands, *inter alia*, when we discussed the welfare impact of road pricing on the peak and the off-peak periods. There it was shown that a congestion toll is needed during the peak when there is excess demand but not during the off-

peak when there is excess capacity. Notice that because there is free-flowing traffic in the off-peak, no tolling is required because no external cost of congestion is generated. Therefore no quasi-rent is being earned on the invested capital and only the variable costs are paid for by the traveller. On the other hand, during the peak period, a positive quasi-rent is earned. (Suppose further that there is an inter-peak period, some quasi-rents are also generated.) With highway capital stock remaining unchanged, the systematic, diurnal nature of travel demand (as opposed to the static, invariant demand case) means that the *sum* of quasi-rents (rather than just the quasi-rent from the singular peak period itself) of the invested capital should be compared with the cost of the highway facility. In other words, when *all* the quasi-rents over the entire demand cycle are summed up and compared with the capital cost, expansion of the highway is either warranted or not, under constant returns. The same conclusions obtained thus far again hold.[42/]

48. An interesting implication is that the entire capital cost of the highway is ‘allocated to’ and borne by peak travellers, mainly rush-hour commuters. This surprising result may seem ‘inequitable’, yet it is perfectly consistent with efficiency analysis. After all, it is peak users themselves that create congestion and they that demand the use of heavily congested expressways which require massive infrastructure developments. Without them, the optimal size of the road would be considerably smaller. The result of allocating all capital costs to users of the peak period has long been recognized in the literature on the pricing of public utilities a la Boiteux (1960). Following our earlier example, the extent to which there is another period -- the interpeak or shoulder period -- with even a modest amount of congestion, would allow for differential pricing and thus the allocation of some capital costs to these interpeak travellers. The optimal investment rule is then to expand a road until the *sum* of the quasi-rents over the demand cycle equals the entire capital cost of the facility under constant returns. By implementing peak-load pricing and altering the investment level of the highway facility, depending on whether profits are positive or negative, the highway network is again optimized. Hence, the

[42/] The output variable needs to be redefined as vehicles per lane per cycle, with the cycle being the duration of a particular charging period.

consideration of demand variability and peak-load pricing would not change the status of our conclusions, in the presence of differences in valuation of time. The fact that the fluctuating demands over the various peak, off-peak and inter-peak periods of a demand cycle are linked by a fixed capital facility and the observation that the consumption of trips must be satisfied by the production of trips during that particular time period combine to yield a simple modification of our result. Pricing, financial viability and cost recovery are again consistent with one another.

49. Keeler and Small (1977) show rigorously how the Mohring-Harwitz framework developed here is extended to the case of variable demands using peak-load pricing in the presence of independent demands and no indivisibilities.[43/] By assuming the demand in each period in fact depends on other periods, i.e., the case of dependent demands, the derived results still go through (Mohring (1970)).[44/]

3) Indivisibilities

50. While still retaining the assumption of constant returns, but accounting for differences in values of time and demand variability, we proceed to drop the assumption of a road being finely divisible. Road construction, in fact, involves significant indivisibilities that cannot be ignored. For example, a road must possess the minimum width for accommodating a standard-sized automobile and should also, ideally, be bi-directional. In the perfectly divisible case, the long-run average total cost curve which envelopes a *continuum* of closely-packed short-run average total cost curves at their minimum points is made horizontal. A flat LRMC curve also coincides with the corresponding LRATC curve (see Fig. 7(b)). Due to the presence of

[43/] It is due to the assumption of independent demands that long-run marginal cost pricing (equals to short-run marginal cost pricing) still holds at each time period. The concept of long-run marginal cost pricing is blurred in the case of jointness of demand.

[44/] Using the distribution of current demand distribution as given, which is synonymous with assuming independent demands, would result in upward bias in peak periods and downward bias in off-peak periods because of the possibility of substitution (Keeler and Small (1977)).

indivisibilities, however, the formerly neat and continuous pattern of the LRMC curve is broken (Neutze (1966), Kraus (1981b)). The new long-run average total cost curve is now composed of a series of short-run average total cost curves, where $SRATC_2$, $SRATC_4$ and $SRATC_6$ denote a two-lane, four-lane and six-lane road respectively. The long-run average total cost curve is a series of short-run average total cost curves connected together in a scalloped-like pattern (see the **solid** LRATC curve labelled ABCDEFG in Fig. 8).[45/] The long-run marginal cost curve takes on the various short-run marginal cost curves in the respective regions, resulting in a discontinuous shark's tooth-shaped LRMC curve (see the **thick** LRMC curve in Fig. 8(a)). Hence one is always working with the short-run curves themselves since one could only operate with a capital facility which is given. By now, it should be clear that whenever a short-run marginal cost curve rises above a short-run average total cost curve, profits can be obtained under short-run marginal cost pricing. Thus, if demand happens to intersect the short-run marginal cost curves in the region of Q_2Q_{2-4}, Q_4Q_{4-6} and Q_6Q_{6-8}, and multiples thereof, then the road makes money in the long run under constant returns. *Per contra*, to the left of the outputs, Q_2, Q_4 and Q_6, namely in the region of OQ_2, $Q_{2-4}Q_4$ and $Q_{4-6}Q_6$, the road loses money. With a 2-lane road, as traffic increases, the road's large fixed cost is spread over the additional traffic, and as congestion sets in, the road begins to make money. In other words, as travel demand continues to grow along the trend, adherence to short-run marginal cost pricing suggests that the road would go through an unavoidable cyclical pattern of deficit, surplus, deficit, surplus, etc. Whether or not one undertakes a road expansion project from two to four lanes depends on the magnitude of the net benefit pie, taking into account the welfare gain and loss triangles.

Optimal Pricing and Investment with Indivisibilities: An Example

51. Consider the demand as depicted in Fig. 8(b), with the actual consideration of indivisibilities. First, we follow the first-best pricing rule of tolling the difference between short-

[45/] A similar set of discontinuous curves is found in, for example, Bennathan and Walters (1979, Fig. 2.2). Note that the scalloped-like pattern is asymmetric because the unit cost curves for four and six lane roads are horizontal multiples of those of the two-lane road. In other words, under constant returns to scale, the AVC and AFC curves, and hence ATC curves fan out horizontally -- a mistake that is quite frequently made in the literature (see, for example, Hayutin (1984, Fig. 2.8 and 2.16)).

run marginal cost and short-run average variable cost, t*. This yields the optimal traffic level of Q* and a positive profit. Without the indivisibility constraint, the existence of profit would indicate that the road is underbuilt. With indivisibilities, the direct one-to-one correspondence between economic profit and road expansion is lost. One is therefore left with the binary choice of, say, expanding the road from a 2-lane road to a 4-lane road. Using the welfare apparatus that we have developed in the Appendix and Figs. 3 and 4, however, the net benefit of such a move is shown to be the sum of the welfare gain of going from Q_{2-4} to Q**, as indicated by the triangular areas *b+c*, and the welfare loss of moving from Q* to Q_{2-4}, as shown by the triangular area *a*. Such a move would clearly be desirable. With a 4-lane road, however, the optimal toll of t** would be insufficient to cover the fixed cost of this new road, resulting in a shortfall. Thus, even though the move from Q* to Q** is a beneficial one, it would mean that the road authority would switch from a profitable regime to a loss regime after incurring the investment cost.

52. Thus the optimal sequence of decision-making is first to establish the policy of implementing marginal cost pricing and then to plan future adjustments of the road network according to *expected* future demand and established pricing policies. When demand fluctuates, pursuing short-run marginal cost pricing at present would mean setting different prices, or tolls, in response to expected current conditions.

53. Suppose the government were faced with a tight budget constraint. It would understandably then be unable to undertake all public projects with positive net benefits. *If* this road authority were mainly concerned about cash flow and financial viability, it could opt not to expand a road, i.e., under-invest, but still charge a congestion toll on the built-up traffic and satisfy economic efficiency in the short run. This option however would not lead to the maximization of society’s welfare in the long run.

4) Returns to Scale

54. The issue of whether constant returns to scale exists or not in road transport is a controversial and important one. Ultimately, it can be answered only via careful statistical analysis. The available evidence in road transportation indicates that all three cases exist: decreasing, constant and increasing returns to scale (see Fig. 9) -- paralleling the case of a competitive private firm and industry -- with profit, zero profit and loss, respectively. (This is but a well-known result of economic theory applied with slight modification to the highway.) It is important to realize at the outset that the case of scale economies, or increasing returns to scale with fixed factor prices under least cost combinations, is merely a case of insufficient demand with respect to the market size in the long run -- a point that is sometimes neglected. The implication is that if traffic were to grow to a point where the capacity of a road is reached, congestion delay would set in, and congestion toll revenues could be collected. After all, the short-run marginal cost curve is always non-decreasing, as shown in Fig. 8. Profits may occur *despite* the fact that the long-run average and cost curve may be declining and the corresponding long-run marginal cost curve lies below the average cost curve.[46/] Then if traffic were to continue to grow as real incomes and auto ownership rise, concomitant with expressway expansion, the decreasing returns region would then be encountered (see Fig. 9). In the case of increasing returns with perfect divisibility (where natural monopoly arguments lie), the long-run marginal cost curve below the long-run average total cost curve pulls it downwards, resulting in losses, beckoning government subsidization. On the other hand, if travel demand is sufficiently high relative to the engineering capacities of roads, the money-making road enterprises would provide much sought-after funds which could be used to finance efficiently priced but money-losing roads -- only if these roads yield positive net benefits to society. We turn next to a discussion of the theory underlying the economies vs. diseconomies issue, together with perfect divisibility vs. indivisibilities, and end with a review of the empirical evidence and recent work.

[46/] This point will be established later in Fig. 12 and 13(a). Basically, with perfect or almost perfect divisibility and scale economies, only losses will occur, whereas with indivisibilities, either profits or losses will result depending on the level of travel demand.

A. Economies of Scale and Rural Roads

55. There is a preponderance of evidence -- geometric, engineering and otherwise -- supporting the case that there are significant economies of scale in the construction of *rural* roads (Walters (1968, pp. 180-82), Mohring (1976, pp. 140-42)).[47/] The illustrations below follow Mohring's scale economy analysis using the geometry of transport right-of-way. In particular, a two-lane road requires a minimum of a twelve-feet width for each lane and a few feet for shoulders and drainage ditches. What this means is that a non-trivial proportion of the provision of a road's right-of-way involves dead space. These indivisibilities -- required to contribute to the building of a minimum acceptable standard such as a given pavement thickness and road size -- help contribute to economies of scale as the large fixed cost of construction and invariate maintenance costs are shared over greater amounts of traffic. Thus doubling the width of a two-lane road more than doubles its traffic capacity, the so-called "shoulder-effect" (Hayutin (1984), pp. 106 and 154). Further, we know that the engineering or basic capacity of a two-lane road is about 2000 vehicles per hour. Since the standard four-lane road has an average engineering capacity of 1800 - 2000 vehicles per *lane-hour*, doubling the width of a two-lane road almost quadruples its capacity (see any Highway Capacity Manual, for example, Transportation Research Board (1985, Tables 2-1 and 2-2), yielding economies of scale associated with road use. However, capacity per lane remains constant beyond four-lane roads, resulting in zero economies of road use thereafter. Further, in order to level hilly terrain and/or fill valley for transportation purposes, the earth moving costs rise less than proportionately. In fairly flat or rolling country selected as sites for road building, doubling the width of a two-lane road generally involves less than doubling the earth moving costs. (On steep hillsides, however, the reverse may be true.) Hence, for these three reasons of: 1) the existence of large fixed costs due to indivisibilities, 2) the technology of road capacity, and 3) the possible reduction in earth moving costs, we can claim that there are economies of scale associated with the construction of a two-lane to a four-lane road. Nevertheless, despite the fact that typical four-lane highways possess two-thirds dead

[47/] Meyer, Kain and Wohl (1965, pp. 200-204) also found some evidence of scale economies of urban roads. However, Keeler and Small (1977, p. 5) query Meyer, Kain and Wohl's findings, stating that their results are in fact based on their initial assumptions.

space and eight-lane highways have only half the dead space, it is not clear from the geometry of highway rights-of-way that economies of scale in *urban* highway construction exist. This is because it is rather difficult to control econometrically for the effects of urbanization and separate it from the effects of size. For example, four-lane roads tend to be built in rural areas, where interchanges and overpasses are widely dispersed, and right-of-way costs are low. On the other hand, six-lane or eight-lane roads are built mainly near metropolitan conurbations, where expressway interchanges and overpasses are closely spaced together, and land acquisition costs are high. In practice, the road authority tends to trade off (and avoid) high right-of-way costs with increased tunnelling and flyover construction costs. Lane expansion from a six-lane to an eight-lane expressway at the margin, for example, would increasingly encounter alignment constraints associated with the terrain. (These constraints might explain why capacity per lane is reduced with six- to eight-lane expansion (Mohring and Harwitz (1962, p. 97)).) This argument is quite independent of whether the expressway is located near urban areas. Hence all three cases of returns to scale occur, resulting in the classic, U-shaped long-run average cost curve, paralleling that of a firm within a competitive industry as we have seen in Fig. 9.

B. Diseconomies of Scale and Urban Roads

56. The discussion thus far has centered on economies of scale to road width for single roads, as opposed to a *system* of roads. Strotz (1964a) conjectures, and Vickrey argues convincingly, that there are considerable diseconomies of scale associated with an urban road networks.[48/] The reasoning is based on the geometry of road network. Given a rectangular grid for an urban road network spaced 2 kilometers apart, as in Fig. 10(a), there are 9 sets of (space-intensive) intersections and traffic lights in a 6 kilometer-wide area. Suppose the number of streets is doubled in order to double road capacity, yielding a grid of a one kilometer-wide spacing (Fig. 10(b)). Quite apart from the possible increase of construction or land acquisitions

[48/] I am extremely grateful to William Vickrey for pointing out the subtleties of these arguments. See Vickrey (1965, pp. 287-288) and Mohring (1976, pp. 144-145) also for the basic line of reasoning.

costs, the number of intersections -- and the required land area and traffic light installations -- is quadrupled to 36. (If no traffic lights are installed, each intersection requires an even costlier overpass or perhaps even a full interchange, a case considered by Mohring (1976, pp. 144-145).) Moreover, despite the fact that trip length remains unchanged between the origin, O, and the destination, D, the number of traffic lights encountered, and hence waiting costs, would then double from 4 to 8 (unless overpasses or interchanges are constructed). Either the installation of traffic lights or the building of overpasses and interchanges would serve to bid up the opportunity cost of land (because less non-highway space would be left for business or other activity) as well as to increase substantially the sum total of the costs of undertaking a trip to the community, resulting in a rising long-run average cost and also a long-run marginal cost curve. The resultant long-run equilibrium for an urban road network in the presence of diseconomies of scale is depicted in Fig. 11. The analysis of the relationship between long- and short-run cost curves is similar to the case of constant returns to scale and is not repeated here. However, because the long-run average cost curve is rising, the short-run average total cost curve for a two-lane road, $SRATC_2$, is tangent to the long-run average cost curve to the right of the minimum $SRATC_2$ point. Because the $SRMC_2$ lies above the $SRATC_2$, $SRMC_2$ intersects the LRMC from below. Short-run marginal cost pricing is equivalent to long-run marginal cost pricing at the efficient output level Q_2*. Notice that the optimal toll, t_2*, follows from pricing at short-run marginal cost and tolling the difference between short-run marginal cost and short-run average variable cost. The difference between $SRATC_2$ and $SRAVC_2$ is, by definition, $SRAFC_2$. The optimal toll, t_2*, clearly exceeds the $SRAFC_2$ by the unit profit of π_2*, with the corresponding rectangular area of profit $\pi_2^* \cdot Q_2^*$. As the urban road network expands from an existing single two-lane road to a double two-lane road, say, substantially costlier construction, tunnelling and land acquisition costs are encountered.[49/] We have also established that time costs rise because of additional intersections and wait time required. Either of the above two factors would serve to push the SRAFC and SRAVC curves up, together with the SRATC curve. Just as the congestion toll filled in the wedge caused by the divergence of short-run marginal and average

[49/] Increasing financial costs of construction via tunnelling and/or flyovers, together with high land resumption cost, are consistent with the findings of Hau (1989) for Hong Kong.

variable cost curves, the divergence between long-run marginal and average total cost curves serves as an indicator of the unit profit (or loss). In competitive equilibrium, all economic profits are competed away in the long run, so the questions that follow are: in what sense is the case of diseconomies of scale a ‘long-run’ concept, and what is the interpretation of the profit areas of $\pi_2^* \cdot Q_2^*$ and $\pi_4^* \cdot Q_4^*$ (for demand curves Q_2^d and Q_4^d, respectively)? The existence of economic profits in the long run is attributable to the rents earned by an invaluable fixed factor of production -- land.[50/] Even though both the SRAFC and the SRAVC curves reflect the increase in costs mentioned above, the existence of the rising opportunity cost of the fixed factor of the remaining parcels of land is still left unaccounted for. Intuitively, just as the driver, in the short run, by imposing external congestion cost on others due to his presence on the road, is charged for it, so also should the urban community, in the long run, charge for the use of scarce urban land in a market economy. Putting it another way, if all factors of production -- including land -- were doubled, so that a scarcity value could be imputed to land, all economic profits would be competed away and vanish in the long run. Clearly, the supply of land cannot be doubled, so it is the existence of land rents which yields (the areas of) economic profits in Fig. 11. Notice that we could no longer use of the existence of profits as a surrogate market signal because of decreasing returns to scale. Since the urban road network is *supposed* to generate substantial sums of money because of high land values, relying solely on the profit mechanism and incautiously investing on urban roads until all economic profits are competed away would result in over-investment in road capacity. With diseconomies of scale and divisibility, roads generate positive profits. Performing proper project appraisal of roads cannot therefore be circumvented.

C. Diseconomies of Scale and Indivisibilities

57. The case of diseconomies of scale combined with indivisibilities is similar in spirit to the analysis of the constant returns case with indivisibilities in Fig. 8. For instance, at any moment

[50/] A purist might argue that these rents actually accrue to land owners and the usual story of long-run super-profits being zero still holds in the presence of rising long-run average costs.

in time, the regions to the right of Q_{2min} and Q_{4min} in Fig. 11 would yield profits. In the case of decreasing returns to scale, with or without indivisibilities, the correct recourse is to invest only if the road project in question passes a stringent cost-benefit test and only if rising costs of roads are compared to the quasi-rents generated in the variable demand periods diurnally. In other words short-run marginal cost pricing, or congestion tolling, as opposed to long-run marginal cost pricing, should be implemented for each period over the demand cycle and the quasi-rents summed up, *regardless* of whether the facility is optimally built. When demand is not known with certainty, probabilistic or expected demand could be used instead and *all* real benefits and costs over the economic life and time periods of a project should be properly measured, discounted and compared with the capital cost of implementing the change. If finely divisible projects are available, we have seen that the procedure is to invest until the marginal benefit of a project -- in the form of time savings -- equals the marginal cost of constructing the capital facility. If indivisibilities abound, then traditional public investment criteria using the net present value criterion or Mishan's (1988, Chapters 35-38) normalized internal rate of return procedure could then be complemented with the diagrammatic welfare analysis presented in this paper. In such a case, unfortunately, the neat linkage between profits and losses as a 'market' mechanism and guide to investment is severed. Thus it is possible for a project with high net benefit to yield a financial deficit. As mentioned before, if a road authority were faced with a severe fiscal constraint, then the road agency could choose to under-invest, rather than over-invest, when faced with an all-or-nothing situation of indivisibilities. Still, it should pursue marginal cost pricing in the short run, while fully taking into account the opportunity cost of congestion toll financing (see Hau (1992)). In this way, short-run marginal cost pricing would yield both economic efficiency and profit, even though an optimal investment strategy would generate an even higher level of net benefit for the community in the long run.

D. Economies of Scale and Indivisibilities

58. *Per contra*, the above analysis for the case of decreasing returns to scale of an urban road network carries over in reverse to the case of increasing returns to scale for rural roads. There the sum total of the quasi-rents captured via short-run marginal cost pricing is insufficient to

cover the entire fixed cost of a particular rural road. Scale economies abound in the construction of rural roads, so that unit losses, L_2* and L_4*, result in the case of an optimized two-lane and four-lane rural road, respectively (see Fig. 12). Standard neoclassical arguments then call for subsidization out of the public treasury for the case of scale economies. The presence of both indivisibilities and scale economies could alter the calculation of optimal tolls and subsidies substantially (Kraus (1981b)). It turns out, perhaps surprisingly, that the existence of indivisibilities serves to improve the state of affairs vis-a-vis the government because, as was shown in the constant and decreasing returns to scale cases with indivisibilities, both surpluses and deficits would occur alternately, depending on the level of travel demand. Similarly, in the case of rural roads with both scale economies and indivisibilities, there are regions (such as those to the right of Q_{2min} and Q_{4min} in Fig. 12) where short-run marginal cost pricing yields profits rather than losses. This is because, with indivisibilities, the long-run marginal cost curve -- composed of joined segments of the short-run marginal cost curves -- is no longer declining all the way but possesses a sawtoothed pattern, alternately exceeding its corresponding long-run (and short-run) average total cost curves and rising at an even faster rate. Thus, just as in the case of constant returns with indivisibilities, whenever a SRMC curve exceeds a SRATC curve, profit exists and vice versa. It is therefore quite conceivable to have a congested two-lane road which generates profits even when subject to *increasing* returns to scale for sufficiently large changes in capacity. The existence of losses does not mean that the road agency should cut back on the provision of highway services. On the contrary, it merely means that other sources of funds ought to be sought in order to finance a worthy project. The government authority’s decision to cut back services just because of losses in such a situation would yield under-investment and possibly stifle economic development and growth in the longer run. The neat linkage between road expenditures and toll revenues has disappeared. Again, the passage of a tough cost-benefit criterion is then a prerequisite for a project to result in maximizing society’s welfare. To repeat, *only if* congestion toll financing is all that is sought by the road agency, and not optimal investment, is it possible to let a two-lane road become congested in the face of rising urbanization and motorization, while differentially pricing it via congestion tolls. In this way, efficient use of a given road is enhanced via short-run marginal cost pricing, but the efficient level of road capacity is not being achieved in the long run.

E. The Extent of Indivisibilities vs. Divisibility and Their Effects on Scale (Dis)economies

59. How often do we encounter surpluses in the presence of scale economies and deficits in the presence of diseconomies? The answer depends on the extent of the presence or absence of indivisibilities. There are two views on this issue. The first perspective a la Keeler, Small and Starkie argues that the aggregate road network could be regarded as divisible. The other view, presented by Walters (1968, Chapter 3) and Kraus (1981b), contends that roads are indivisible because the main measure of highway capacity involves the discreetness of the number of lanes.

60. The construction of a road or an additional lane may not be finely divisible in and of itself, but taking the road network *as a whole*, a single newly constructed facility can be regarded as an incremental addition to the network, resulting in the applicability of the foregoing marginal analysis (Keeler, Small and Associates (1975, Chapter 2)). Also, often times, varying some dimensions of road features, other than the number of lanes, increases the capacity of the road network. For example, the lane width, the provision of auxiliary lanes, horizontal and vertical alignments, and the surfacing of road shoulders can all be varied incrementally (Starkie (1982)). One could characterize this view by treating the lane capacity as a continuous variable rather than a discrete one (Small, Winston and Evans (1989, p. 103)). If the road authority pursues the twin optimizing rules of pricing and investment in capacity, then the road network would be in long-run equilibrium. So with constant returns and a divisible road network, roads would break even.[51/] However, some individual roads would make money and some would lose money. On the whole, if the economies and diseconomies of scale are "probably roughly offsetting" as Meyer and Gómez-Ibáñez (1981, pp. 191-192) concluded, then the highway budget would be balanced. Whether or not scale economies and diseconomies are counter-balancing would depend on the degree of urbanization. With increasing urbanization, profits would tend to predominate even after charging land rents as a cost. With indivisibilities, the profit (or loss) regime occurs

[51/] Constant or even decreasing returns would be satisfied if two bi-directional roadways are built in proximity to one another.

about half the time but it is unclear what the relative weights would be when travel demand is reasonably assumed to grow over time.

61. Under decreasing returns and perfect divisibility, we have shown in Fig. 11 that profits always occur. The essence of that figure is combined with Figs. 13(a) through 13(c). Perfect (Fig. 11) and almost perfect divisibility (Fig. 13(a)) and an urban road network would mean that the marginal cost pricing of trips would always be profitable. Note that the continuous regime of profit will not be uniform once a threshold level of indivisibilities is reached, resulting in regions where losses also occur. If the extent of indivisibilities progresses from small but significant (Fig. 13(b)) to severe (Fig. 13(c)), the regions which yield potential losses become even larger. The symmetry carries over to the increasing returns to scale case. Again, the substance of Fig. 12 is culled and combined with Figs. 14(a) through 14(c). With perfect (Fig. 12) and almost perfect divisibility (Fig. 14(a)) in the presence of scale economies, losses would always occur. With scale economies and a 'significant' level of indivisibilities (Fig. 14(b)), smaller regions of profit would become available but would disappear when approaching the neighborhood of the limit (Fig. 14(a)). Nevertheless, even if one were to accept Walters' (1968, Chapter 6) argument that there are significant indivisibilities and scale economies in *rural* roads, we have demonstrated that profits (and losses, of course) would nevertheless arise under congestion tolling. Scale economies in the presence of indivisibilities and financial viability are *not* necessarily incompatible.

F. Empirical Evidence on the Scale Economy Issue

62. Fitch and Associates (1964, p. 131) give some numerical support for the case of scale diseconomies in the United States. Walters (1968, pp. 184-185), using Meyer, Kain and Wohl's (1965, p. 205) data, shows that there are *diseconomies* of scale in the construction of four-lane, six-lane and eight-lane *urban* road segments. (By employing Walters' straightforward approach, Hau (1989) demonstrates that there are increasing costs associated with a sample of four-lane, five-lane and six-lane roads in Hong Kong.) Without imposing any prior specifications about the extent of returns to scale -- Keeler and Small (1977) find evidence of constant returns to scale

for a sample of San Francisco Bay Area roads. Their often cited econometric study is important because of the balance budget implication for congestion pricing, a result which was also quoted by Newbery (1990).[52/] By contrast, using engineering specifications to estimate each of the cost components of an urban highway network model, Kraus (1981a) finds that there are increasing returns to scale in road construction in terms of length of freeway and interchanges but not for overpasses and length of arterials. He makes the crucial observation that factor prices (such as right-of-way prices) are to be held constant for making relevant comparisons of scale - specific (dis)economies. The reciprocal of his best "pseudo-empirical" estimate of returns to scale in urban highway network capital costs is 0.84, which translates to the economies of scale degree of 1.19. Meyer and Gomez-Ibáñez (1981, pp. 191-192), in assessing the available estimates in the conflicting literature, conclude that economies and diseconomies of scale are "probably roughly offsetting." Hayutin (1984), in an unpublished dissertation, refined the Keeler-Small model and applied it to a sample of U.S. Interstate Highways. By including both intercity (or rural) and urban routes in her data set, she estimates that there are clearly increasing returns with respect to the number of lanes. Her results bear out Mohring's scale economy implications

[52/] Keeler and Small's (1977) result is one of the more rigorous econometric analyses bearing on the scale economy issue, in that they are able to separate the confounding effects of size and urbanization. In particular, they regressed construction cost per lane-mile on the number of lanes and various discrete variables which capture the differences between urban, suburban and rural-suburban areas. It is the inclusion of the latter variables which enable them to econometrically control for the effects of expressway capacity on construction cost. The selected sample of 57 roads was based on all state-maintained roads for nine San Francisco Bay Area counties, including arterials, expressways and rural roads. They estimated both a non-linear and log-linear specification, with both yielding statistically insignificant statistics for the estimated degree of homogeneity. Based on the slightly better fit for the log-linear specification, the returns to scale parameter of (-)0.0305 translates itself into the case of "mildly" increasing returns to scale of 1 - 0.0305 = 0.9695. By taking the reciprocal of the estimated returns to scale parameter, the economies of scale degree of 1.0315 is obtained. Since 1.03 is "statistically indistinguishable" from 1, they conclude that there is no firm evidence for scale economies in road construction measured in terms of lanes. The far-reaching policy ramifications of this result could perhaps be buttressed if there were calculations or discussion of the power of their hypothesis tests. This is because not being able to reject a null hypothesis is not necessarily equivalent to accepting the null hypothesis. One hopes that the type II error is relatively small. However, to be fair to their important contribution, the power function for published papers are seldom to be found in the literature.

of highway geometry with respect to the "shoulder effect".[53/] More recently, in reassessing the earlier work of Keeler and Small (1977), Small, Winston and Evans (1989, Chapter 6) use the result of zero economies (corresponding to the scale economy parameter of 1.00) as well as the mildly increasing returns case of 0.97 (corresponding to the scale economy parameter of 1.03) (see Small (1992, Chapter 3) for a review also). Fully cognizant of the possibility of the existence of increasing returns to scale in urban highway travel a la Kraus, they presented their simulation results on highway finance based on three encompassing parameters for the degree of scale economy: namely 1.00, 1.03 and 1.19.

G. Recent Results on Cost Recovery

63. As part of a major World Bank research project, Newbery (1988abc, 1989, 1990) and Newbery, Hughes, Paterson and Bennathan (1988) extend Mohring's now classic result of highway finance by relaxing the assumption of an infinitely durable highway whose capacity can be continuously adjusted. It is common knowledge that pavements wear out after extended use, but it is less well known that the damage of a vehicle on the road pavement is related to the *weight per axle*, as opposed to the gross vehicle weight *per se*. Common Practice (Highway Research Board (1962)) is to measure this damaging power by the number of equivalent standard axle loads (ESAL), where one ESAL is equivalent to the load of 18,000 lbs. (8.2 tons or 80 kilo newtons) single axle. The American Association of State Highway and Transportation Officials' milestone road test indicates that the damaging power of a vehicle on pavement is proportional

[53/] Hayutin (1984, Chapter 4) regressed construction investment per mile on the number of lanes and other variables similar to those used by Keeler and Small (1977) (see previous footnote). Further, by including the paved width per lane variable, diseconomies to width measured in feet, rather than in lanes, result. Hayutin (1984, Chapter 5) concludes that the stronger effect of scale economies with respect to the number of lanes dominates the effect of diseconomies associated with building wider footage. In an excellent review of the literature surrounding the scale economy issue, Hayutin (1984, pp. 158-159) observes that Meyer, Kain and Wohl's (1965, pp. 200-214) conclusion that there are increasing returns to scale in freeway construction stems from their engineering assumptions about costs, as opposed to actual estimation. Her statistical results, however, are supportive of their conclusions. My comment regarding the power of the hypothesis tests applies to her study also.

to the fourth power of individual axle loads, acquiring the title of the "fourth-power law."[54/] For instance, the rear axle of a 13-ton light truck with two axles can result in over 300 to 2400 times more pavement damage than that of a large car weighing 2 tons.[55/] One inescapable conclusion is that almost all damage is caused by heavy vehicles such as trucks and buses: relatively little is due to automobiles. Recently, Newbery (1988a) characterized another type of externality called a road damage externality: the *external* damages imposed upon the pavement by heavy vehicles result in increased vehicle operating costs to subsequent motorists for the rest of the periodic life cycle of a road. (Thus, the various costs of undertaking a trip mentioned in the introduction of this paper (footnote 3) should now be more narrowly defined as *external* costs, namely: 1) marginal external congestion cost; 2) road wear cost *and* damage externalities (grouped under the rubric of road damage cost); 3) environmental costs; and 4) costs due to increased risk of accidents not borne by private parties.)[56/]

[54/] Formally, the damaging power (in EASLs) of an axle load, *l*, in tons, is approximately equal to: $(l/8.2)^4$, hence we say that the damaging power is proportional to the fourth power of the axle load. As a hypothetical example, a 24 ton truck whose weight is distributed evenly among 2 axles would cause more than 3 times as much damage as the same truck whose weight is distributed evenly among 3 axles: $(2/3) \cdot (12/8)^4 = 3.375$.

[55/] The example assumes that more of the weight -- 8 tons -- is distributed on the rear axle of the light truck whereas a fully loaded automobile has its weight uniformly distributed among the two axles. Small and Zhang (1988) and Small, Winston and Evans (1989, Chapter 2 Appendix) have recently disputed the fourth-power law and performed a statistical analysis of AASHO's road test data to show that the equivalence factor for an axle load rises steeply to the third power, thereby proposing a "third-power law" instead. Whether the power term is actually 3, 3.5 or 4 (as respectively claimed by Small, Paterson, or engineers conventionally) is not as crucial as the significant maintenance cost savings to be made if vehicle damage is efficiently charged for. Conceivably, an increase in the power term would enlarge the small deficit that Small, Winston and Evans obtained by charging for congestion and road damage costs. I argue that the charging of *all* externalities including environmental and accident costs would very likely close any remaining gap.

[56/] Based on empirical evidence for the whole road network of Tunisia, Newbery, Hughes, Paterson and Bennathan (1988, Table 7, 23 and pp. 58-59) and Newbery (1988c, Table 1) demonstrate that the external urban congestion costs alone account for the "overwhelming fraction" (about nine-tenths or more) of total external costs of road usage (excluding environmental and accident costs) for automobiles and utilities. For heavy vehicles, the reverse appears to apply in both Tunisia and Ghana, with road damage costs dominating instead (see Gronau (1991) also). (Heggie and Fon (1991, Annex 1) argue that some of Newbery's calculations of congestion costs for Tunisia are overestimates.) Congestion costs constitute a large share of total road costs for both the United Kingdom (nine-tenths) and the United States (one-fifth) in the long run, whereas road damage costs are about 3 to 3.5% in the United Kingdom and only 2% in the United States (Newbery (1988a), Small, Winston and Evans (1989), Chapter 6). Pollution costs appear

64. Newbery's (1988a) "Fundamental Theorem of Road User Charges" says that *if*: 1) the maintenance policy pursued by a highway authority is condition-responsive, i.e., road maintenance is carried out whenever a road's pre-determined roughness level is reached, without being optimally set (Paterson (1987)); 2) the age distribution of roads over the life cycle is uniformly distributed; 3) there is no traffic growth, i.e., traffic flow is constant over time; and 4) all road damage is caused solely by vehicles, *then* the external road damage cost is *nil* on average (Newbery (1988a), Proposition 1). The variable road maintenance user charge component, applied on a per ESAL basis, will fully cover the average cost of repair (Newbery (1988a), Proposition 2). When viewed diagrammatically, this result is analogous to charging for maintenance cost in another dimension, as opposed to charging solely for the congestive effect of automobiles on a per passenger car equivalent (PCE) basis in Fig. 3. Intuitively, the damaging force of an additional ESAL has the external effect of raising the vehicle operating costs of subsequent vehicles over time, just as the external congestive effect of an additional car has on other vehicles behind it within a traffic stream. In addition, the indirect impact of accumulating more ESALs is to reach the preset, maximum allowable roughness level of the road earlier than expected, thereby precipitating an overlay and a corresponding lowering of vehicle operating costs. It is unclear *a priori* which of these two magnitudes dominate. Hence Newbery's (1988a, 1989) breakthrough was to prove, in a variety of ways, that the additional vehicle operating costs attributed to road damage externalities just cancel out with the vehicle operating cost savings when averaging over roads of different vintages.

65. In the general case, the fundamental theorem relaxes Assumption 4) above by allowing for road damages that are independent of use, such as weather-related deterioration. In this case of a condition-responsive maintenance policy, the road damage externality is no longer zero but is quantitatively negligible. While the use-related part of maintenance costs is chargeable because of the road damage cost, the fraction due to weather is not allocable. Hence the presence of weather-related road damage results in only partial recovery of all maintenance costs when only

to account for less than one-tenth of total road costs in the United States, whereas accident costs are on the same order as external congestion costs for the United Kingdom. (Newbery (1988b, 1990)).

charging by road damage costs. Therefore, in order to close the highway budget deficit, Newbery (1989) needs to price for the marginal external congestion cost within a standard long run framework (see Newbery (1988b) also). That is, under the conditions of constant returns to scale and road use, the efficient congestion toll will yield revenues that cover the capital cost of the highway but *only* the invariant (or non-allocable) portion of road maintenance cost attributable to weather, see Fig.5 for definition of short-run fixed cost (Newbery (1989), Proposition 1). (Following Mohring (1965), we have seen how this is done in our simple investment rule of comparing optimal toll revenues with fixed costs, be they construction or maintenance costs, see Fig. 5 and the surrounding text.) It then naturally follows that the optimal road user charge, i.e., the optimal congestion toll-cum-road damage charge, will then yield revenues that cover the capital cost of the road and the *total* road maintenance cost in a constant returns world (Newbery (1989), Proposition 2). In other words, the optimal congestion toll covers both the capital cost of the road and the non-allocable fraction of road maintenance, whereas the allocable fraction of road maintenance is still chargeable to traffic loadings via the average variable road maintenance cost component per ESAL. This explains why the total expenditure on road maintenance is fully recoverable. *If* the same constant returns conditions and assumptions 1) through 4) again apply, and *if* we further accept that 5) heavy vehicles predominantly use the slow lanes and are confined there, *then* the optimal road user charge will recover the capital cost of the highway and *twice* the total maintenance costs of the road (Newbery (1989), Proposition 3). If all heavy vehicles are confined to the slow lane, then the damage costs are accumulated in a shorter time span than if they were to be spread out evenly over all lanes. This has the effect of raising the maintenance cost of the slow lane and the effective cost of widening the whole road, since all lanes are typically resurfaced together once the performance service index of the road dips below the trigger point. Cost-benefit analysis of lane expansion to reap time savings suggests that the marginal cost of investment in capacity would now have to account for both the increased road maintenance cost for road strengthening *and* the annuitized capital cost on a PCE basis. (This is a slightly modified version of our earlier optimal investment rule.) Moreover, heavy vehicles are charged for the traffic loadings they create. While stronger roads are cheaper to maintain, investment in road strengthening is costly, resulting in costlier upkeep and higher capital cost. The larger fixed costs translate themselves, in the long run, into a requirement for higher congestion charges, which help contribute to an overall budget surplus for the road authority, despite "massive" increasing returns to road strengthening (Newbery (1989)). Another way of looking at this result is that the congestion toll effectively covers the road's entire capital cost as before, whereas the maintenance cost is recovered twice -- once via the variable road maintenance user charge component applied on a PCE basis and the second time around via heavy vehicles per ESAL.

66. The most serious drawback in the propositions above is the condition of constant returns to scale.[57/] We have reviewed the empirical evidence and find that there are increasing as well as decreasing returns to scale operating on different parts of the road network, meaning that deficits as well as surpluses would most likely co-exist.

5) Variability of Road Thickness

67. We are now in a position to formally relax the implicitly used assumption of a given road thickness and to incorporate the latest work into our model. In Road Work: A New Highway Pricing and Investment Policy, Small, Winston and Evans recently provided a comprehensive exposition of their technical extension of Mohring and Harwitz's long-run result on highway economics with optimal durability, within a standard neoclassical welfare maximization framework.[58/] Instead of taking current highway design standards as given,

[57/] Newbery (1988a) regards the lack of the pursuit of a condition-responsive maintenance strategy as the most serious limitation. Further, Newbery (1989) cites Keeler and Small's (1977) result of constant returns to scale, which is based on a sample of San Francisco Bay Area roads. Newbery (1990) also cites Kraus (1981a) as evidence supporting constant returns when Kraus demonstrates slightly increasing returns (see previous section). Based on an earlier version of Newbery (1989), Heggie and Fon (1991) take issue with many of Newbery's assumptions.

[58/] Newbery (1989), for example, derives optimal durability (or strength, rather) and confines the extent of his analysis principally to the constant returns to scale world. By contrast, Small, Winston and Evans (1989, Chapter 6) explicitly explore the case of increasing returns to scale of road construction, as well as durability, and present simulation results of an urban expressway and arterial -- with sensitivity analysis -- within a long-run equilibrium framework. The first order conditions of their optimization problem yield the optimal pricing, investment and durability rules.

another characteristic of a road -- its thickness -- can be varied. A thicker pavement would serve to withstand the damaging power of trucks more and thereby prolong the life of a road. Despite the fact that there are tremendous economies associated with road durability, the additional strengthening of the road would still substantially increase the total investment cost of an overlay. Using the standard optimization technique of consumer's surplus and producer's surplus maximization, the intuition of the three allocation rules derived are again based on the simple notion of setting marginal benefit equals to marginal cost. That is:

1) *The first rule -- the optimal pricing rule* -- says that the traveller should undertake a trip only up to the point where the incremental benefit just offsets the incremental cost to the community. A vehicle's entry into a transport corridor results in two effects: the congestive effect which depends on the number of PCEs, and the damaging effect based on the number of ESALs. Thus our optimal pricing rule derived earlier is adjusted to include an extra component -- a road damage charge -- which balances out any short-fall.

2) *The second rule -- the optimal capacity rule* -- says that, with a condition-responsive maintenance strategy built in, optimal investment for highway capacity is then to expand the width of the road up to the point where the marginal cost of capacity *plus* maintenance cost is equivalent to the resultant time savings.

3) *The third rule -- the optimal durability rule* -- says that the road is optimally strengthened by investing up to the point where the marginal cost of durability just equals the savings in vehicle operating costs to other motorists *plus* the associated savings in maintenance costs due to a thicker pavement.

The two latter rules could be regarded as an investment rule extended into two different dimensions. Thus the three rules of optimal pricing, optimal capacity and optimal durability yield efficient prices and trips, as well as the optimal number of lanes and inches of pavement thickness.

68. There is evidence that there exist economies associated with road strengthening (Small, Winston and Evans (1989, Chapters 2-3), Winston (1991)). By reanalyzing AASHO's road test data within an economic optimization framework, Small and Winston report that the optimal thickness for a rigid pavement is an inch and a half higher than the current ten inch standard,

which follows AASHO's guidelines. The remarkable finding is that a mere increase in thickness of 15% would lead to a doubling of pavement life from 13 to 26 years (Small and Winston (1986, 1988), Small and Zhang (1988). The logical implication of massive economies to increasing road thickness -- a fact long known to highway engineers -- is losses for the road authority. We therefore ask: is there a theoretical reason that would allow us to still adhere to the impeccable marginal cost pricing principle and yet achieve the goal of cost recovery? To answer this question we turn to the notion of diseconomies of scope.

Economies of Scope vs. Diseconomies of Scope

69. The condition of constant, increasing or decreasing returns to scale yields the respective result of break even, loss or profit, respectively (see section XIV - 4 and Fig. 9). The empirical findings of returns to scale evaluated earlier suggest that decreasing returns to scale in road construction in urban areas may perhaps offset the increasing returns to scale in rural areas, but that the highway budget may still be in deficit due to economies of road strengthening. The answer can only be established using econometric analysis. Even if we were to accept Keeler and Small's (1977) careful finding of constant returns, would it not seem plausible to argue that the pavement deficit due to significant economies to road durability necessarily yield an overall deficit for the road authority? The answer is no.

70. Up till now, we have confined ourselves to the single product world of traffic volume. But road transport involves two products: namely traffic volume and loadings, which carries us into the literature of multiproduct industries and returns (Bailey and Friedlaender (1982)). The issue is whether a multiproduct firm can jointly manufacture the various products cheaper than if each firm were to produce an output separately. If it is cheaper to combine operations and share in the joint costs, there are economies of scope. The notion of diseconomies of scope can best be grasped by illustrating the design of a railway track (Kim (1987)). A railroad track that is built to withstand the axle loadings of freight would need greater strength and thickness, which conflict with the requirement of securing a smooth ride for train passengers. Ensuring both thickness and smoothness of tracks is more costly, resulting in diseconomies of scope. The

analogy carries over to roads, where it would cost more to produce a highway that is thick enough to handle heavy vehicles and yet wide enough to accommodate the considerably larger number of automobiles. Thus there are no economies of scope even though one product -- the number of loadings -- is clearly subject to increasing returns to durability, whereas the other product -- the traffic volume -- is potentially subject to increasing returns to scale due to road construction. In other words, it is the other characteristic of production -- the scope -- that would tip the budget balance away from a deficit (Small, Winston and Evans (1989), Chapter 6).

71. Automobiles cause virtually no road damages compared with heavy vehicles but trucks are fewer in number and therefore cause lesser amount of congestion. Hence one could separate, roughly speaking, traffic flow from loadings by identifying automobiles with the former and trucks with the latter. Traffic flow, in turn, requires road capacity whereas loadings require road strengthening. The empirical finding of "modest but significant" diseconomies of scope by Small, Winston and Evans (1989, Chapter 6) of about 6 to 10% tips the product-specific economies of scale closer to one -- the constant returns case. (The multi-product economies of scale range from 1.00 to 1.06.) Their bottom line simulation results of an urban expressway and arterial and sensitivity analysis demonstrates that a budget balance and hence cost recovery are achievable. The shortfall of a few percent of total road costs can still be recovered by maintaining some level of first registration fees, annual licenses and/or fuel taxes. Finally, since *all* externalities ought to be internalized in principle, air, noise pollution and accident costs should also be appropriately charged for (Carbajo (1990) and Cameron (1991)). In this way, the highway budget would most likely make a profit (see footnote 56).

72. The basic intuition behind this remarkable result is as follows: because most roads are currently built to accommodate both automobiles and heavy vehicles, a neat dichotomy of allocating pavement wear costs between automobiles and trucks cannot be achieved. Thus 1) the marginal cost pricing of traffic flow requires a congestion charge -- which effectively covers the capital cost of investment in the long run -- and 2) the marginal cost pricing of pavement wear -- associated with the investment cost of strengthening the road -- results in a relatively steep road

damage charge. The marginal cost-based road user charges combine to yield the double-charging of roads, depicting diseconomies of scope.[59/] One logical implication of the diseconomies of scope argument is that savings could be reaped from building a thinner autos-only road system (the savings of which are estimated to be on the order of 23% by Keeler and Small (1977)). With a universal road system, roads should then be built or resurfaced durably on the slow lane only and heavy vehicles should be confined to that lane. (This experiment is being carried out in California and Florida (Small, Winston and Evans (1989, p. 15)).

[59/] The theoretical finding of the double-charging of roads is also derived by Newbery (1989) for the case of constant returns to scale and road use. However, he neither cites nor employs specifically the concept of multiproduct returns in his work.

XV. Summary and Conclusions

73. One of the earliest contributions to the economic analysis of road pricing was from a French engineer, Jules Dupuit (1844). He was the one (and not Alfred Marshall) who introduced the concept of consumer's surplus -- the cornerstone of the welfare analysis of any public project -- and brought it to bear on the subject of toll roads. It is a similar tack that this paper has taken, in the belief that a picture would perhaps speak a thousand words. I have synthesized the dominant thinking to date on the topic of road pricing by the main protagonists and integrated them into a consolidated, analytical framework.

74. According to my analysis of the traditional road pricing arguments, it is hardly surprising that road pricing as advanced in the past encountered its share of difficulties. This is because the congestion toll has the effect of a tax (increase) on trip-makers, despite the fact that it is an externality-corrective tax. The gist of the argument is as follows: people are against congestion pricing because: 1) those who are tolled would face a higher price relative to a no tax situation on average;[60/] 2) those who are priced off the road in order to circumvent paying the toll are clearly worse off as a result of the 'forced' switch onto a different mode or time of day; and 3) the other road users who are not tolled -- the tolled on -- are no better off and, indeed, may even be worse off if congestion is encountered. My applied welfare analysis actually takes into account the increase in revenue to the government and the issue of transfer payment. If each of the parties is separated out as: 1) the tolled, 2) the tolled off, 3) the tolled on, and 4) the government, the group that stands to gain the most is the government (and the untolled -- the rest of society), unless the toll revenues are earmarked. The other party that is primarily better off are those with very high values of time. Only in the hypercongestion case could *all* groups be made better off on average.

[60/] Consider a more realistic situation where the existence of a uniform fuel tax effectively translates itself into a trip price that is higher (but not as high as the peak-period price), and exactly the same analysis follows.

75. It has been argued that the disposition of the revenues of externality corrective (toll-) taxes should accrue to the public treasury (Baumol and Oates (1988), p. 29).[61/] Conventional cost-benefit analysis treats a dollar as a dollar to whomsoever it accrues, and also implicitly assumes that only consumers derive satisfaction from revenues. Hence, *unless* toll revenues are channeled back through reduced transportation related taxes, user charges or improved public services, neither the tolled nor the tolled off would endorse road pricing.

76. In this paper, I have also shown step-by-step how to implement short-run marginal cost pricing in transport following Walters and others. In particular, I have established that implementing the *optimal pricing rule* -- the first rule -- is equivalent to setting an optimal road user charge, where: 1) a congestion toll on the *difference* between the marginal cost and the average variable cost of a trip is imposed, and 2) the maintenance cost of road use is also charged. Further, I have shown that the process of determining an optimally priced and invested road system is similar, albeit with a couple of important differences, to the process of achieving long-run equilibrium of a typical product within a competitive environment, inspired by the basic Mohring-Harwitz model. The issue of short-run vis-a-vis long-run marginal cost pricing is clarified *inter alia*. For instance, implementing short-run marginal cost pricing is equivalent to pursuing long-run marginal cost pricing in a steady-state world in the long run. However, only short-run marginal cost pricing would be able to capture the peak/off-peak nature of travel demand. In the absence of scale economies, the *optimal capacity rule* -- the second rule -- says that the existence of economic profit, i.e., toll revenue collection less the fixed and non-use-related costs of a road, would serve as a surrogate market mechanism indicating that the road ought to be expanded. Putting it another way, the investment rule says that a road ought to be expanded to the point where the additional cost of investment in capacity equals the additional savings in travel time. In the long run, toll revenues would cover the interest on the capital investment, invariate maintenance, depreciation and operating costs of the road. Maximizing society's welfare dictates that one should implement short-run marginal cost pricing *over the long*

[61/] This is because a Pigouvian tax is one which imposes a positive price on a producer of an externality and a zero price on a consumer of an externality.

run by varying the size of the highway capital stock. In this way, the pursuit of efficient pricing and a self-financing road system would be compatible with one another and no residual cost need be covered. In one stroke, the same congestion tolling mechanism solves the pricing-cum-investment problem, satisfying the conceptual guidelines of efficiency pricing, economic and financial viability as set out in the introductory section of this paper.

77. Relaxing and flushing out major assumptions indicate that the result derived here is robust and applicable to: 1) a multiplicity of roads, 2) a road which is subject to diurnal variation of demand and the peak-load problem, and 3) differences in values of time. If constant returns to scale can be shown to hold on average for a particular city with severe congestion, it could potentially aid in greatly simplifying the planning of highway investment. It could also be used as a yardstick, against which scale economies or diseconomies could be measured.

78. Economic efficiency would be enhanced if marginal cost pricing of a trip were done in the short run and optimal investment in capacity were pursued over the long run. I have established that, if governmental authorities were to charge correctly for congestion, it is possible for them to make money on a road while satisfying economic efficiency. Profitable roads arise in heavily utilized or urban areas because land rents of real estate are high and congestion tolls reflect the rising opportunity costs. Yet, it is possible that congestion pricing in the presence of both indivisibilities and diseconomies of scale in urban roads may curtail the extent of profitable undertakings. Similarly, pursuing marginal cost pricing under the restrictive conditions of both indivisibilities and scale economies of rural roads could also result in profits in the short run. Thus far the points I have made are based on first principles.

79. In the long run, the simple pricing and investment model implies that the marginal cost pricing of trips covers all the fixed costs of the road and the congestion toll (which captures the quasi-rent) behaves *as if* it is a capital charge. This analysis also means that an optimally invested road system is in fact one where a road should not always be uncongested during the peak period. An optimally congested road is akin to the commonly accepted notion of an optimal pollution level in the field of environmental externalities. An uncongested road for every time

period of the day would suggest that that road is over-invested, either because of indivisibilities or nonmarginal cost pricing. If a road is indeed overbuilt, abandoning or downsizing it in the long run may be unavoidable on narrow cost-benefit criteria. The act of downgrading lightly used roads in order to save on the costs of maintaining higher standards of road pavement is a form of disinvested.[62/] Alas, given that almost all existing road systems are non-optimally designed and that costs are considered sunk in the short run, the efficient usage of such a network would still call for marginal cost tolls. Any increase above the road user charge should then be regarded as a 'pure tax element' or surcharge, whose contribution to general revenues should perhaps be made on either fiscal or non-economic grounds.

80. One may ask: starting off with an overbuilt road system, say, is there a way in which road pricing based on the marginal cost concept can be implemented within an institutional context where severe fiscal constraints on public expenditures prevail? To answer this question requires that we go beyond first principles.

81. Recent extensions a la Newbery-Small-Winston have enriched the basic model developed diagrammatically here by incorporating the fact that heavy vehicles are the cause of road damage. Charging for both the external and variable cost of road damage on a vehicle weight *per axle* basis would help close the deficit that may arise from congestion tolling. Further, a road needs to be strengthened to the point where the additional cost of investing in durability just balances out the incremental savings from maintenance and vehicle operating costs. Thus the third rule -- the *optimal durability rule* -- is born. As a natural extension of pricing for externalities, air, noise pollution and accident costs ought to be charged for. Surely in this way the highway budget would more likely involve profits than losses *if* the issue of cost recovery of the sector cannot be ignored.

[62/] The common practice of downgrading roads is consistent with the findings of road deterioration in developing countries (Harral and Faiz (1988, p.32)).

82. Given point estimates (or preferably, functional specifications) of speed-flows, demands and the value of time, one can estimate and simulate some of the analytical results developed here. When combined with the associated optimal pricing and investment rules, the efficient level of prices, user charges, speed, volume-capacity ratios, and trips, as well as the optimal number of lanes and inches of pavement thicknesses can be obtained.

83. The fact that in the transport context, the consumer-producer is both a willing 'victim' as well as a 'beneficiary' has policy implications. As 'victims' of congestion externalities, perhaps travellers ought to be compensated. Note, however, that Pigouvian toll-tax revenues are not supposed to be used to compensate 'victims' of externalities (Baumol and Oates (1988), p. 23). Also, intuitively, motorists would be induced to drive more because the level of compensatory payments would depend on their car usage, so economic efficiency would be violated. In this context, a road fund would be consistent with first-best pricing only if the funds were used in an indirect manner. Travellers are also 'beneficiaries' of road transport by virtue of their being present on congested roads, and their contributions to the toll revenue component of 'user charges' reveal their willingness to pay. In the absence of lump sum transfers, earmarking of toll revenues could serve as a useful device in principle to approximating benefit taxation as a way of satisfying a commonly accepted notion of 'fairness.' Similarly, heavy vehicles ought to incur their 'fair' share of hefty pavement wear fees. Combining these plausible arguments and our earlier results of optimal pricing and investment principles, suggests that some form of dedicated funds is perhaps necessary -- either in the form of a road fund or a transport fund -- if road pricing is to gain political acceptance.[63/]

[63/] Recent developments in electronic toll collection and electronic road pricing in Norway, Sweden and Cambridge (England) point to the fact that travellers do not object to road pricing when the toll revenues are earmarked for both road construction and improvement and/or the provision of better public transport. (With optimal tolling, however, high purchase taxes and registration/license fees of vehicles ought to be reduced to a level sufficient to cover the administrative and enforcement costs of collection. If the road maintenance cost is constant with respect to the traffic level, an appropriate fuel tax could perhaps be used to approximate usage.)
Indeed, a recent national survey conducted in England indicates that when people were asked whether they are for or against road pricing, about 57% are against it. However, when the question was posed in a different way: would they be supportive of a *package* approach to road pricing, with the revenues from road pricing used only to finance public transport, 57% of the *same* surveyed population were in favor of

A. The Role of a Road Fund

84. If a road fund were to be set up, compensation would need to be made sufficiently indirect to satisfy Pareto efficiency. Thus the funds for road construction should be used both to maintain existing roads and to finance new roads, and the profits generated from urban roads could be used to finance the fixed capital cost of *worthwhile* rural roads in a nondistortionary manner.[64/] To what extent can the profits collected from heavily used roads offset the losses arising from the construction of lightly used roads? The answer depends on the extent of the interaction of both scale economies and indivisibilities and should be examined on a country by country and case by case basis using econometric analysis.

85. A road fund is attractive because of the high marginal cost of raising a tax dollar. Moreover, a road fund run by an autonomous authority would increase the *linkage* between revenues and expenditures, currently lacking in a politically-based budgeting process, thereby improving managerial efficiency. Without the setting up of such a fund, however, deficits from lightly used rural roads (with increasing returns to scale) would demand subsidization from the

road pricing rather than against it (Jones (1991), Goodwin (1989)). (May and Gardner's (1990) simulation results also confirm the case for an integrated approach to road pricing, and buttress the analytical results presented here.) Using an equilibrium travel demand model of modal choice applied to the San Francisco Bay Area, Small (1983) concludes that congestion pricing combined with the redistribution of toll revenues would result in benefits to *all* income groups.

As an illustration, the Oslo Toll Ring, currently in operation, charges private cars (light vehicles) going into the city center a toll of 10 kronos (approximately US$1.50). Roughly 60% of motorists opt for the subscriber lanes in 1991, which are operated via electronic toll collection, rather than manually-operated toll lanes. About 80% of the earmarked funds are used to finance road construction and 20% for busways, buses and trams. The optional nature of the successful Oslo Toll Ring raises the intriguing idea that congestion tolling should perhaps be implemented on a voluntary basis: with the choice of a combination registration fee/fuel tax based on 'average' usage, or a subscription to the use of electronic devices.

[64/] The common sense notion of using the profits from heavily used roads to finance lightly used roads has its intellectual roots in Alfred Marshall's 'tax and bounty' system of pricing:

> "One simple plan would be the levying of a tax by the community on their own incomes, or on the production of goods which obey the law of diminishing return, and devoting the tax to a bounty on the production of those goods with regard to which the law of increasing return acts sharply."
>
> Alfred Marshall, Principles of Economics, 1920, p. 392

Marshall also observes that it is necessary to consider the cost of administering such a tax-subsidy system.

public treasury, and would thus compete for tax money valued at a high opportunity cost. By symmetry, surpluses that accrue in heavily utilized urban areas (with decreasing returns to scale) should then be priced at a *premium*.[65/] These welfare losses and premiums would presumably offset one another if viewed within the same (transport) sector -- with a nominal value of a dollar being treated at its face value -- so that we are back to the case of pure efficiency concerns.

86. Even if a certain city in a developing country, say, is found to be faced with mainly increasing returns to scale, the deficit could be closed, in principle, by appealing to the notion of diseconomies of scope. Meeting the requirement that a road network be both large enough, in terms of capacity, and strong enough in terms of pavement thickness, can be quite costly. Scope diseconomies in highways mean that a road network that accommodates both loading and traffic volume found universally is more costly than the sum of an autos-only *and* a tailored trucks-only road system. Hence, the surplus associated with diseconomies of scope offsets the potential deficits associated with scale-specific economies of road construction or use. The viability of the fund is enhanced by the fact that the maintenance cost of the road pavement is recovered twice: once when traffic flow creates congestion, and the second time when traffic loadings cause road damage. Thus, the idea of a trust fund administered by an independent agency according to strict cost-benefit principles is likely to be feasible.

B. The Role of a Transport Fund

87. Alternatively, taking the transport sector as a whole, a transportation fund ought to be set up.[66/] If dedicated funds are set up in this way, indirect 'compensatory' payments can be achieved and would not depart far from optimality. I recommend this both because the

[65/] I am indebted to Sir Alan Walters for this insight.

[66/] Using an entirely different model than the one used here, Vickrey (1977) establishes the result that cities should use land rent tax revenues arising from agglomeration economies to finance mass transit and public transportation which are subject to increasing returns. Indeed, he argues forcefully and proves the case that *not* subsidizing these fixed costs would be inefficient.

problem of highway congestion is tied intrinsically to the provision of poor transit alternatives and because public transport encompasses a substantial if not the lion's share of trips undertaken in both newly industrializing and developing countries (Deaton (1987)).[67/] Typically, the production of bus services is subject to consumer-side bus route and frequency economies of scale. Hence additional funds in the form of *user-side subsidies* are required to meet the financial shortfall arising from the capital equipment, if bus usage is priced at marginal cost. Road pricing would result in more crowded and inferior public transport services unless bus companies were to offer more bus services (and hence lower generalized prices) as a supply response. Then 'untolled' public transport users or captive riders would be made better off.[68/] Here the double charging of automobiles via traffic volume and heavy vehicles via loadings would help to close the deficit gap. Increasing by popular rapid mass transit and light rail systems -- both of which are subject to significant scale economies -- also require capital funds, the construction of which should be based on economic viability. Unless a global view is taken of the congestion problem and more rational time-of-day pricing practiced in *all* modes (in contrast to tackling individual, non-optimally priced modes), the urban transportation problem will continue to be pervasive.

88. Even without dedicated funds, it is essential to pursue efficient pricing and stringent benefit-cost analysis link by link and mode by mode on both a volume and loading dimension. Thereafter, the results can be presented for public scrutiny, thereby improving managerial efficiency and public accountability. The competitive tendering and private provision of certain transport services could also serve to enhance managerial efficiency in the public sector. Issues warranting further investigation include the corporatization of certain transport agencies.

89. Subject to further research, the idea of setting up a transportation or road fund and the pursuit of marginal cost pricing in all its dimensions would enable us to satisfy the quinpartite

[67/] The provision of public transport mentioned in Section VII, assumed to operate under constant returns was used merely as an illustrative convenience but this assumption does not result in loss of generality.

[68/] Notably, captive bus passengers would benefit from road pricing if equilibrium transit travel times are lowered.

principles of the World Bank's general guidelines, as stated at the outset of this paper, namely to: 1) implement efficiency pricing, 2) meet economic viability, 3) meet (to a considerable extent) financial viability, 4) achieve (some degree of) 'fairness' among beneficiaries, and 5) attain (somewhat) managerial efficiency of the public authority. The conception of a fund passes many of the criteria for a 'good' earmarking arrangement as presented in McCleary (1991). The implementation of marginal cost pricing in both the traffic and loading dimension could be done with the advent of recent technological breakthroughs in automatic road user charging utilizing automatic vehicle identification and classification, all of which are subject to remarkable scale economies (Hau (1992)). Alternatively, less powerful road pricing instruments such as area licensing, simple cordon pricing schemes and the monitoring of vehicle and axle loading via weigh-in motion scales can be used.

Timothy D. Hau

APPENDIX:

Measurement of the Welfare Impact of Road Pricing

1. A brief analysis of the measurement of the welfare impact of road pricing would help explain why road pricing is unpopular. We will consider several acceptable approaches to measuring the net benefits offered by the introduction of road pricing on a non-perturbed equilibrium. Each casts different light and insights on the controversy surrounding road pricing.

A. Quantity Approach

2. The first approach, which is more popular in the U.S. literature, is to measure the so-called welfare gain or loss areas (see Kraus, Mohring and Pinfold (1976), for example). This standard method is labelled the quantity approach or the 'American' approach (see Fig. 3). The loss in valuation to the consumer-traveller from a reduction in trips from Q^o to Q' as a result of increasing the generalized travel cost to him from P^o to P' is the vertical, trapezoidal area $d+g+k$. The saving in resource cost to travellers from the reduction in traffic, together with the saving of congestion in the form of externality reduced, is the vertical area $l+d+g+k$. The net benefit to society of the introduction of road pricing is given by the triangular area l.

Net Benefit Approach

3. A variant of this approach is the net benefit approach (see Fig. 3). The net benefit in the case of the optimal traffic level of Q' is typically a large triangle (the pie area $a+b+e+h$ between the demand function and the marginal cost curve), with the pie triangle emanating from the point of optimum. Similarly, the net benefit in the case of the non-optimal level of Q^o is given by the difference of the pie area $a+b+e+h$ and the small triangular area l. The latter area is of course the welfare cost saved when the traffic level is induced to be lowered from Q^o to Q'. This variant is intuitively appealing as it graphically illustrates that net benefit is maximized with

marginal cost pricing. Any departure from the point Q′, either in a positive or negative direction, would slice into this maximal net benefit pie. To the left [or right] of Q′, travellers' marginal valuation would exceed [or be less than] the marginal cost.

B. Change in Total Benefits and Total Costs Approach

4. The above procedure, and its variant, is an impeccable one. However, there is an alternative intuitive method to calculating the net benefit of introducing road pricing. This approach is widely used in the British literature (Ministry of Transport (1964), Tanner (1963, p. 318); Gwilliam and Mackie (1975, pp. 105-106), Thomson (1970) and Thomson (1974, pp. 142-145)). The findings of the Ministry of Transport, known as the Smeed Report, present a different calculation of the areas of gains and losses indicated above, yielding different insights into the problem (see their Appendix 3). The 'British' approach uses the change in total benefits and change in total costs. The change in total benefits accruing to those who are tolled off the road are negative because they suffer a loss in valuation equivalent to the vertical area *d+g+k*. The change in total costs -- expressed as the reduction in the total expenditure on travel in the form of savings in time cost -- accrues to all motorists and is given by AVC^oQ^o - $AVC''Q'$ (or P^oQ^o - $P''Q'$), that is, the area *e+f+g+k*. The net gain to society is the area *e+f-d*. Heuristically, the remaining users find that they derive satisfaction from the savings in time cost of the area *e+f*. The losers -- those tolled off the road -- would clearly experience a welfare loss of the area *d*.

5. The discussion thus far gives the conclusion and mistaken impression that those who remain behind are in fact better off by the *entire* savings in time cost of area *e+f*. In fact, drivers who remain on the road have to make toll payments of the area *b+c+e+f*, which in turn become a gain to the government in the form of toll revenues. (This is the notion of a transfer payment excluded in cost-benefit calculations using the British approach, see Gwilliam and Mackie (1975, pp. 105-106).) Yet, paradoxically, it is precisely the imposition of this tax -- resulting in a transfer payment -- which enables those who remain on the road to benefit the time savings of

an additional area *e+f*.[69/] Without the tax, motorists are not properly induced to save valuable time resources: the time is completely lost. The ones who remain on the road, however, actually suffer a loss of consumer's surplus of the rectangular area *b+c*. It is as if a discriminating monopolist -- in the guise of the government's tax department -- carves away part of the users' consumer surplus. Also, to 'benefit' from time savings of the area *e+f*, drivers are in fact trading a dollar of money for a dollar's worth of time, implying that both a standard and constant value of time and efficiency analysis are assumed. (The rest of the transfer payment of area *e+f* also accrues to the government in the form of tax revenues.)

6. *Prima facie*, whether or not the net benefits of introducing road pricing using this latter approach (i.e., area *e+f-d*) and the former approach (i.e., area *l*) are equal is not at all obvious. The latter procedure gives less indication of the notion of optimality when compared to the first approach, especially with regard to its variant. In the quantity approach, one could move to the left or right of Q′ and observe that the net benefit pie to society would clearly be eroded, suggesting that Q′ yields maximal net benefit. Using the latter approach, however, as Q increases past Q′, a welfare loss area would increase. This would have to be offset with a new rectangular area of saving in resource cost. The point is that it is unclear whether Q′ can be shown to be optimal, at least diagrammatically, because the new rectangular area *may* not offset the new (trapezoidal) welfare loss area. Formally, the proof is as follows: the move from Q′ to Q^o yields a change in cost to society of area *l+d+g+k* because the vertical area below the marginal cost curve is a proper measure of cost. Equivalently, the change in variable cost of

[69/] The rectangular area *b+c+e+f* should be counted as accruing *either* to the government in terms of toll revenues *or* returned to consumers (via a hypothetical lump sum transfer mechanism). A lesson to be learned regarding the issue of transfer payment is to avoid double-counting. *If* a dollar is treated as a dollar to whomsoever it accrues *and* the transfer mechanism is implemented, *then* the move to road pricing results in positive net benefit to society of area *e+f-d*. Putting it another way, the standard notion of a transfer payment of the area *b+c+e+f* says that money goes from the consumer's pocket into the government's. It is important to view the time savings of area *e+f* as an additional layer on top of the transfer payment itself. The bottom layer goes from the motorist's pocket to the road agency's; the top layer is obtained because the motorists who are tolled are forced to trade money with time. Remarkably, it is this coerced payment of the tax revenue area *e+f* which brings about a real saving of time of an area of equal size.

going from Q′ to Q^o is the inverted L-shaped area *e+f+g+k*. By definition, these two areas must be equal, implying that *e+f=l+d* or *l=e+f-d*. Diagrammatically, it may *appear* as if the change in total benefits and total costs approach yields larger net benefit area-use. However, it needs to be clearly shown here.[70/]

C. Consumer's Surplus and Producer's Surplus Approach

7. The third approach using the summation of changes in consumer's surplus and producer's surplus involves the term quasi-rent.[71/] The traveller is both a consumer (in the sense that he derives benefits from purchasing a transport service) and a producer (in the sense that he himself has to purchase the inputs with both his own time and operating costs). In the absence of road pricing, because drivers travel up to the point where average variable cost intersects the demand (at output level Q^o), the entire receipt (from the consumer traveller's expenditure) goes to cover the 'payment' of user-supplied factor inputs, so zero quasi-rent is thereby generated. However, it could be equivalently stated that the nil area can be expressed as the difference of two triangles, i.e., area *(e+h)* - *(c+d+l)*, by simply exploiting the meaning and geometric relationship of average and marginal cost curves. In the advent of road pricing, the quasi-rent -- the return to a fixed factor of production -- is essentially the amount which the traveller-as-producer 'receives' over and above his total variable costs. This quasi-rent, instead of accruing to the drivers as such, is captured by the government in the form of toll revenue or a user charge, and hence should be accounted for properly in benefit-cost calculus. Note that the quasi-rent of area *b+c+e+f* can be re-expressed as the area *b+e+h*. Clearly, the change in the quasi-rent would be equal to the area *b+c+d+l*. Coupled with the loss in consumer's surplus of

[70/] Lee (1982) claims that the two areas based on the different methodologies are the same but does not prove it. The spirit of the analysis I show here underlies Gwilliam and Nash's (1972) comment on Beesley and Walters' (1970) evaluation of urban road investments.

[71/] The notion of rent is a slippery one and warrants clarification. Rent is the analog of producer's surplus in the input market. Rent is a *permanent* payment to a factor over and above that which is required to draw forth its resources. Quasi-rent is a *temporary* payment and would continue only until the capital asset is depreciated or possibly transferred to another use (see footnote 29 and Mohring (1976, Chapter 2) also). Note that a high price and willingness-to-pay yields high quasi-rent, and not the reverse.

area *b+c+d*, the net benefit area *l* emerges. Hence, we have shown in different ways that the three approaches are identical.[72] Note, also, however, the second approach is used and extended because it graphically illustrates the broad distributional implications of road pricing.

8. Perhaps one reason why there is confusion regarding the two approaches above is because of Walters' (1961a, 1961b) treatment of the MC and AVC curves as marginal social cost and *marginal* private cost curves respectively. (My interpretation here is at variance with the common use of the latter term since I regard it as somewhat of a misnomer.) Walters' use of the AVC as MC curve immediately brings to mind standard diagrams of an externality such as the classic economics text example of a polluting factory, with the consequent changes in consumer's surplus, producer's surplus and externality valuation. As carefully shown above, this example is not valid in our analysis because the marginal private cost curve is only marginal with respect to the driver himself. The individual perceives and bears the average variable cost only: it is merely a decision curve and no more. Since the area below the AVC curve is not the total private cost, only an incorrect interpretation can be drawn by mathematically integrating the area under the marginal private cost curve, which turns out to be an *average* variable cost curve. (Further, producer's surplus should really be interpreted as quasi-rent to avoid possible confusion, especially in understanding the relationship between pricing and investment.) In the absence of the optimal pricing of trips, average (variable) cost pricing prevails with the associated inefficiencies. This illustrates strongly the need to reserve the term 'short-run marginal cost pricing' to be consistent with the World Bank's policy guideline (World Bank Operational

[72] In fact, if Q^o is very close to Q', the marginal cost is equivalent to the change in total variable cost. By evaluating the difference of the change in variable cost with the marginal valuation, the two methods discussed above (the quantity approach and the *change* in total benefits and total costs approach) are seen to be equal.

Manual Statement (1977)). I employ the term marginal *social* cost pricing when accounting for all the other externalities such as environmental pollution and accident costs.[73/]

9. A technical paper by consultants hired by the Hong Kong Government indicates that the net benefit due to introducing road pricing corresponds to the area $e+f+d+g+k$ (Transpotech, (1983, Fig. 4)). The consultants' explanation of the extra area $g+k$ is either based on a possible misunderstanding of the first two approaches, or simply a matter of double-counting areas $g+k$. The authors state that "this money [referring to the areas $g+k$] is available to be spent in other ways, perhaps on other modes of travelling". Having already included the resource saving as a reduction in the expenditure on travel, P^oQ^o - $P''Q'$, the resource saving from the tolled off drivers of the vertical area $g+k$ should *not* be counted twice. The point here is that unless care is taken to ensure rigorous cost-benefit analysis, the benefit (or cost) figures would be biased, as has been the case with the evaluation of the electronic road pricing experiment in Hong Kong.[74/]

[73/] For example, Glaister's (1981, Chapter 5) use of the term 'marginal social cost pricing' is synonymous with the marginal cost pricing concept employed here.

[74/] Based on the numbers presented for an illustrative case, the bias is 40% upwards. It should be stressed, however, that it is unclear from a reading of the Hong Kong Government's Main Report on the Electronic Roading Pricing Pilot Scheme whether the final report followed the methodology outlined in Technical Paper 1 (Transpotech (1983, 1985)).

FIGURES

Economic Fundamentals of Road Pricing: A Diagrammatic Analysis

by

Timothy D. Hau

Transport Division
Infrastructure and Urban Development Department
The World Bank

Figure 1
Derivation of a Travel Time-Flow Curve of an Urban Highway

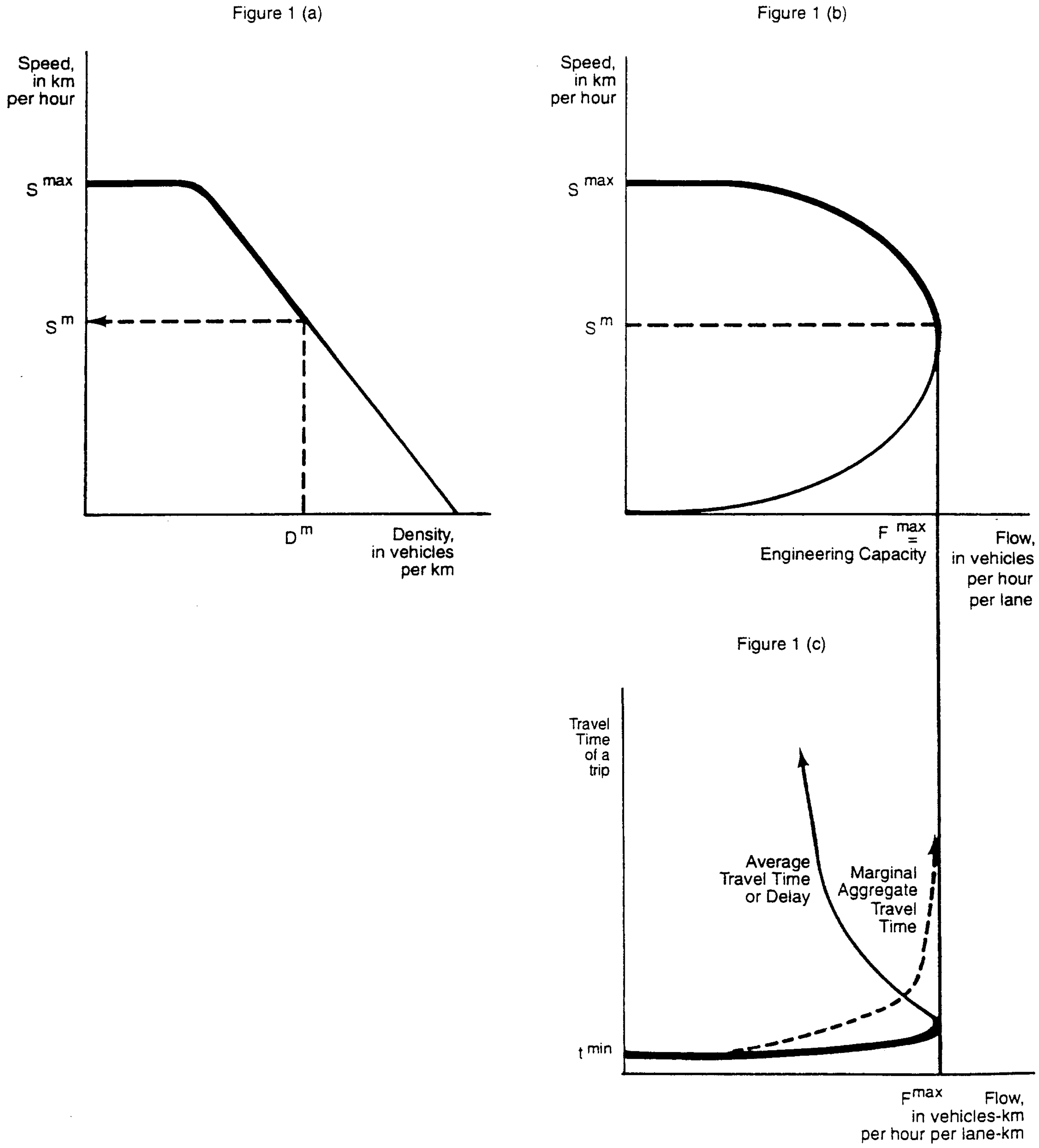


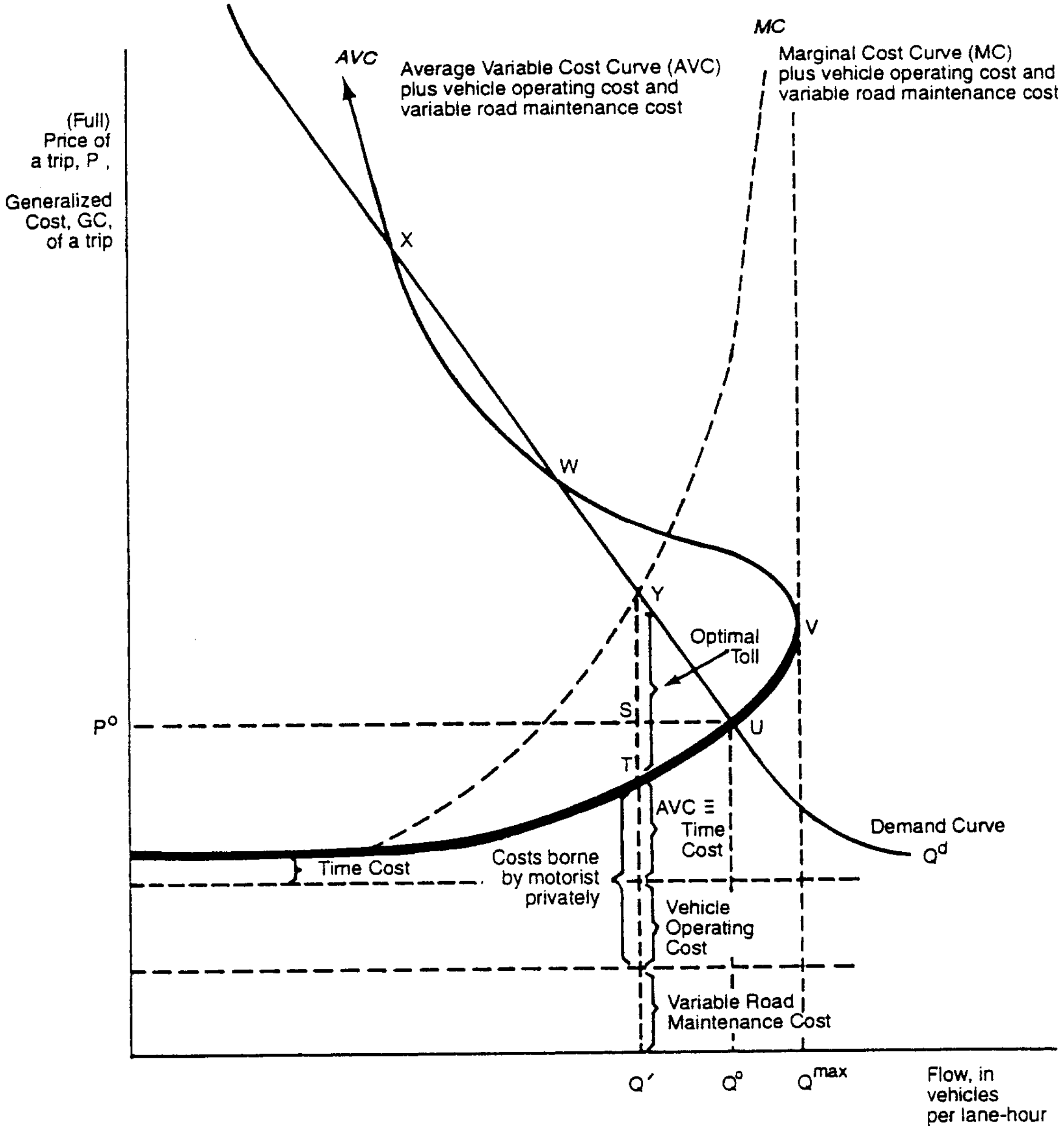
Figure 1 (a)

Figure 1 (b)

Figure 1 (c)

Speed, in km per hour
S^{max}
S^{m}
D^{m}
Density, in vehicles per km
Speed, in km per hour
S^{max}
S^{m}
F^{max} = Engineering Capacity
Flow, in vehicles per hour per lane
Travel Time of a trip
Average Travel Time or Delay
Marginal Aggregate Travel Time
t^{min}
F^{max}
Flow, in vehicles-km per hour per lane-km

Figure 2
Derivation of the Marginal Cost Curve and Congestion Toll


MC
AVC
Average Variable Cost Curve (AVC) plus vehicle operating cost and variable road maintenance cost
Marginal Cost Curve (MC) plus vehicle operating cost and variable road maintenance cost
(Full) Price of a trip, P ,
Generalized Cost, GC, of a trip
X
W
Y
V
Optimal Toll
S
U
T
P°
AVC ≡ Time Cost
Demand Curve
Q^d
Time Cost
Costs borne by motorist privately
Vehicle Operating Cost
Variable Road Maintenance Cost
Q′
Q°
Q^{max}
Flow, in vehicles per lane-hour

Optimal User Charge = Costs Imposed on Other Motorists + Road Authority
= External Congestion Cost + Variable Road Maintenance Cost
= Optimal Toll + Variable Road Maintenance Cost User Charge Component

Figure 2(a)
'Dynamic' Phenomenon of Traffic Growth: The Relaxation Effect

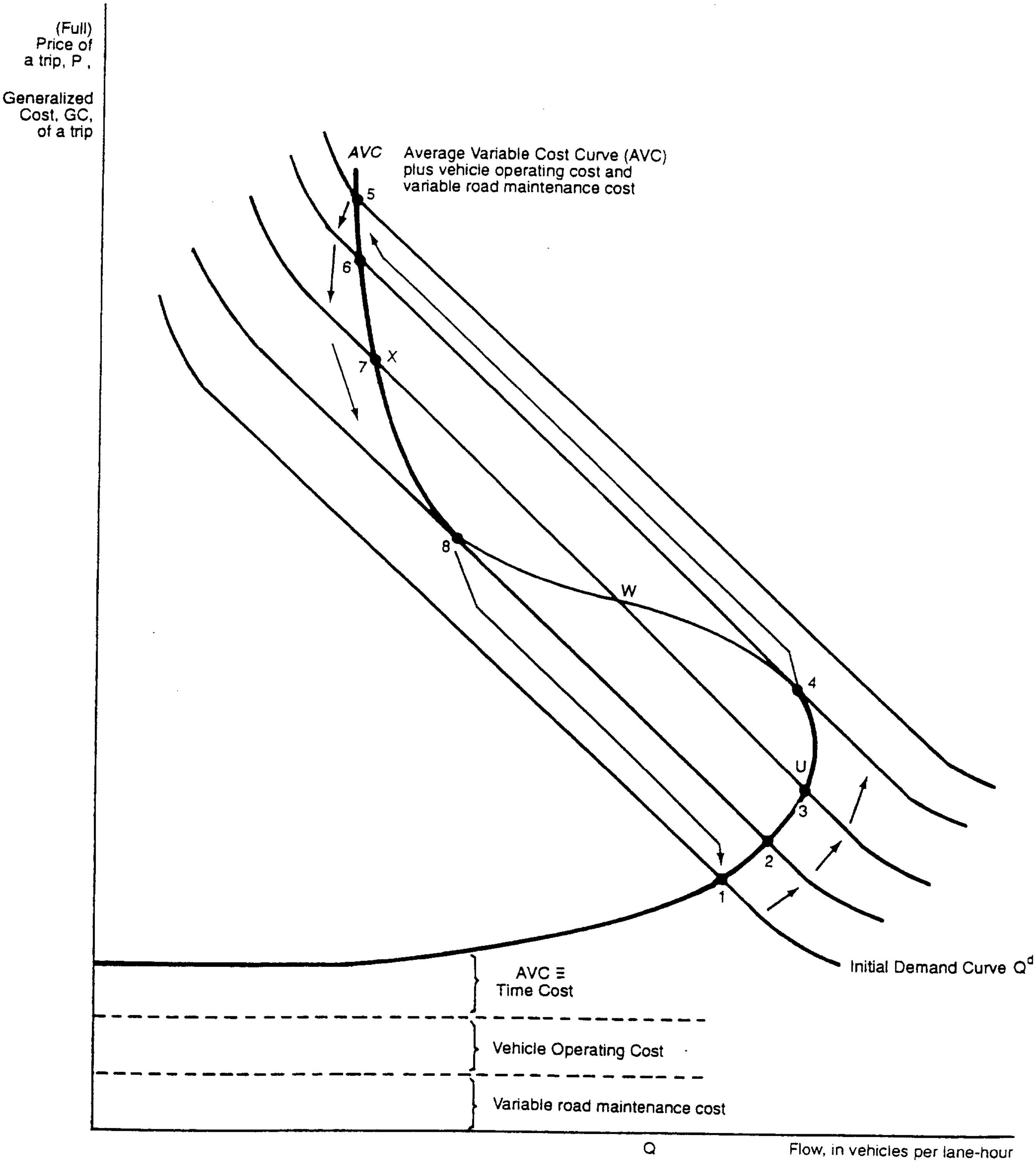


Figure 3
Welfare Impact due to the Introduction of Road Pricing in the Peak Period: Short-Run Marginal Cost Pricing

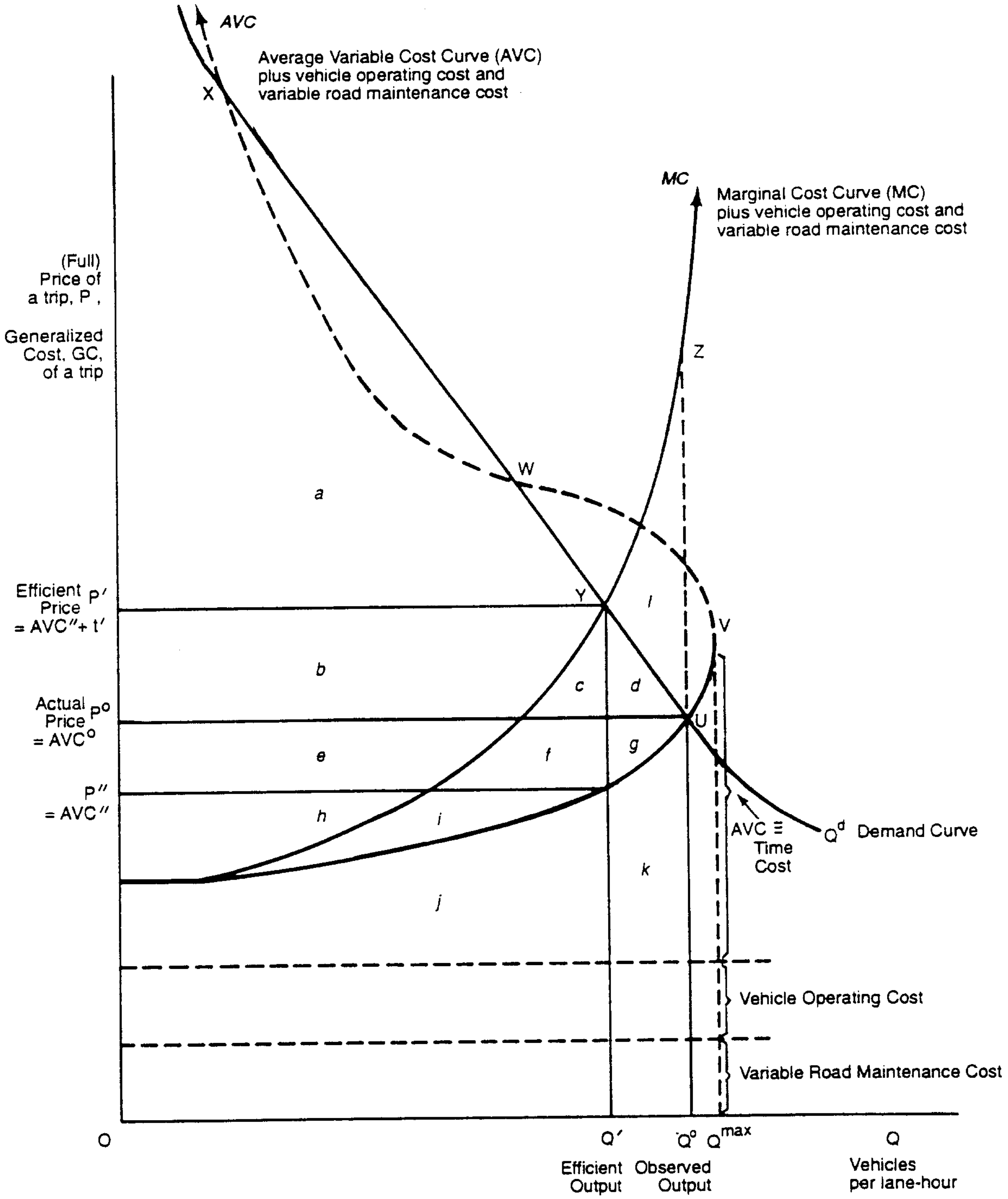


Figure 3(a)
Welfare Impact due to the Introduction of Road Pricing in the Peak Period:
Short-Run Marginal Cost Pricing
'Hypercongestion' Case

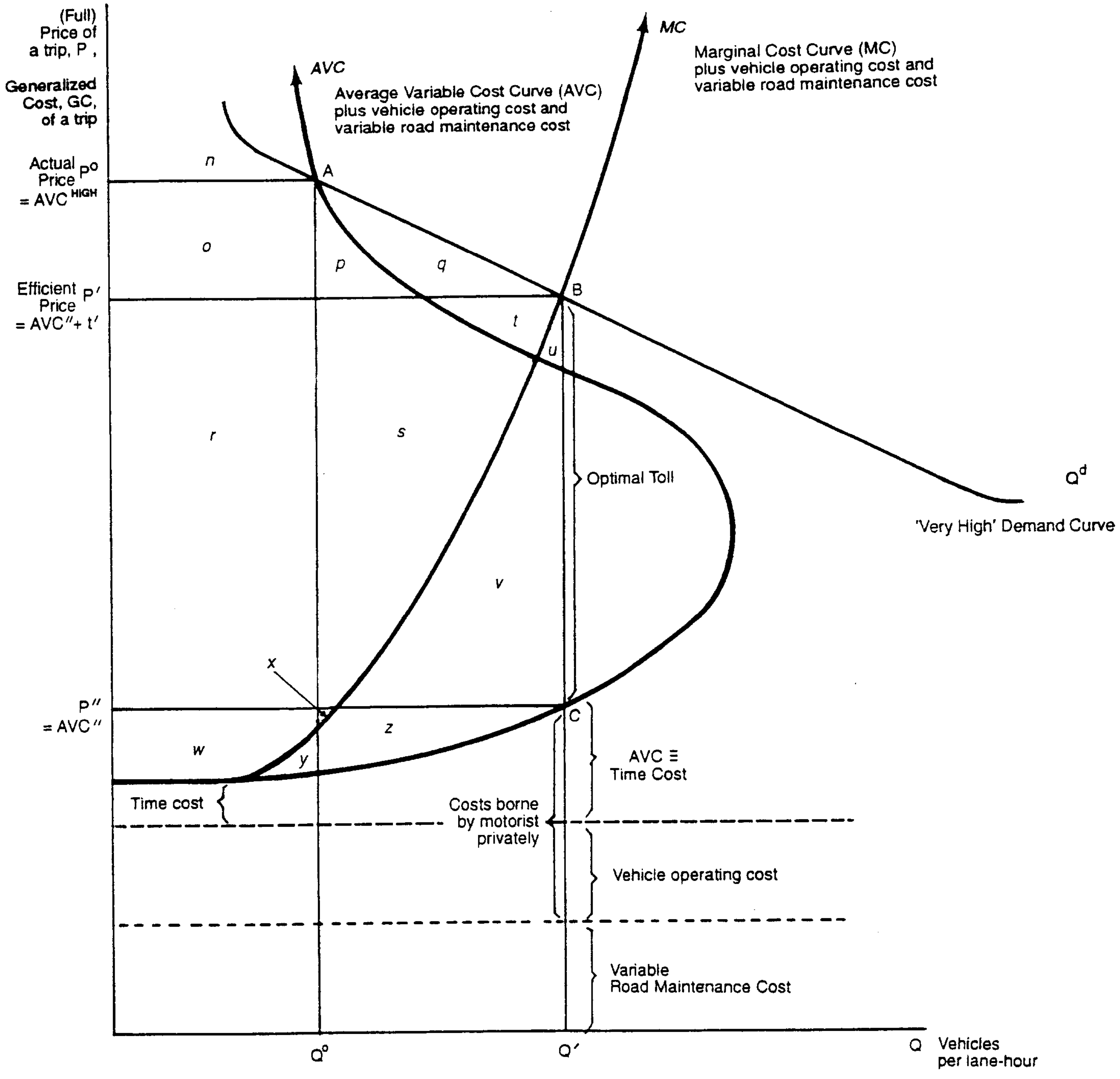

(Full) Price of a trip, P ,
Generalized Cost, GC, of a trip
Actual Price P^o = AVC HIGH
Efficient Price P′ = AVC″+ t′
P″ = AVC″
AVC
Average Variable Cost Curve (AVC) plus vehicle operating cost and variable road maintenance cost
MC
Marginal Cost Curve (MC) plus vehicle operating cost and variable road maintenance cost
Q^d
'Very High' Demand Curve
A
B
C
n
o
p
q
r
s
t
u
v
w
x
y
z
Optimal Toll
AVC ≡ Time Cost
Time cost
Costs borne by motorist privately
Vehicle operating cost
Variable Road Maintenance Cost
Q^o
Q′
Q
Vehicles per lane-hour

Figure 4
Effect of the Introduction of Road Pricing in the Peak Period on the Off-Peak Period

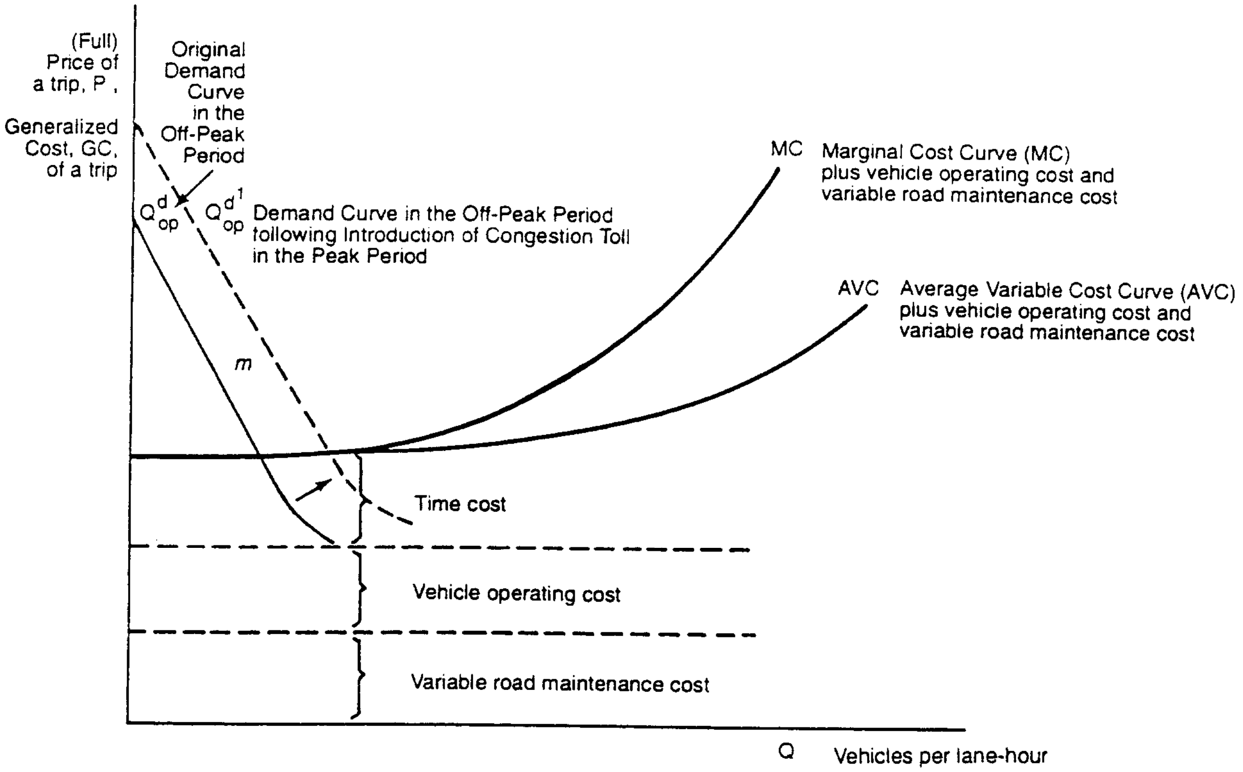
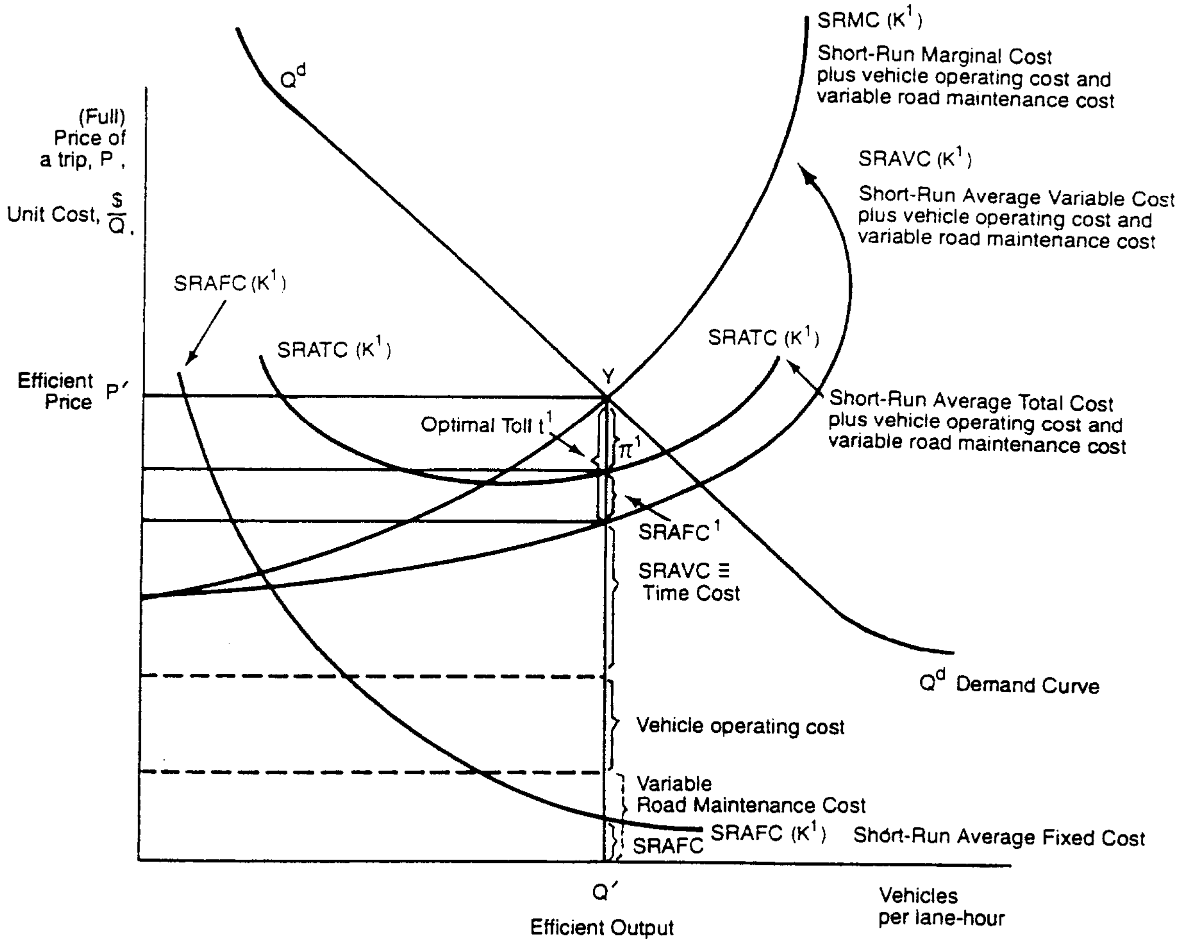

(Full) Price of a trip, P ,
Generalized Cost, GC, of a trip
Original Demand Curve in the Off-Peak Period
Q^d_{op}
$Q^{d^1}_{op}$
Demand Curve in the Off-Peak Period following Introduction of Congestion Toll in the Peak Period
MC
Marginal Cost Curve (MC) plus vehicle operating cost and variable road maintenance cost
AVC
Average Variable Cost Curve (AVC) plus vehicle operating cost and variable road maintenance cost
m
Time cost
Vehicle operating cost
Variable road maintenance cost
Q
Vehicles per lane-hour

Figure 5
Introducing the (Short-Run Average) Fixed Cost, SRAFC, of a Road, Short-Run Optimal Toll with Economic Profit

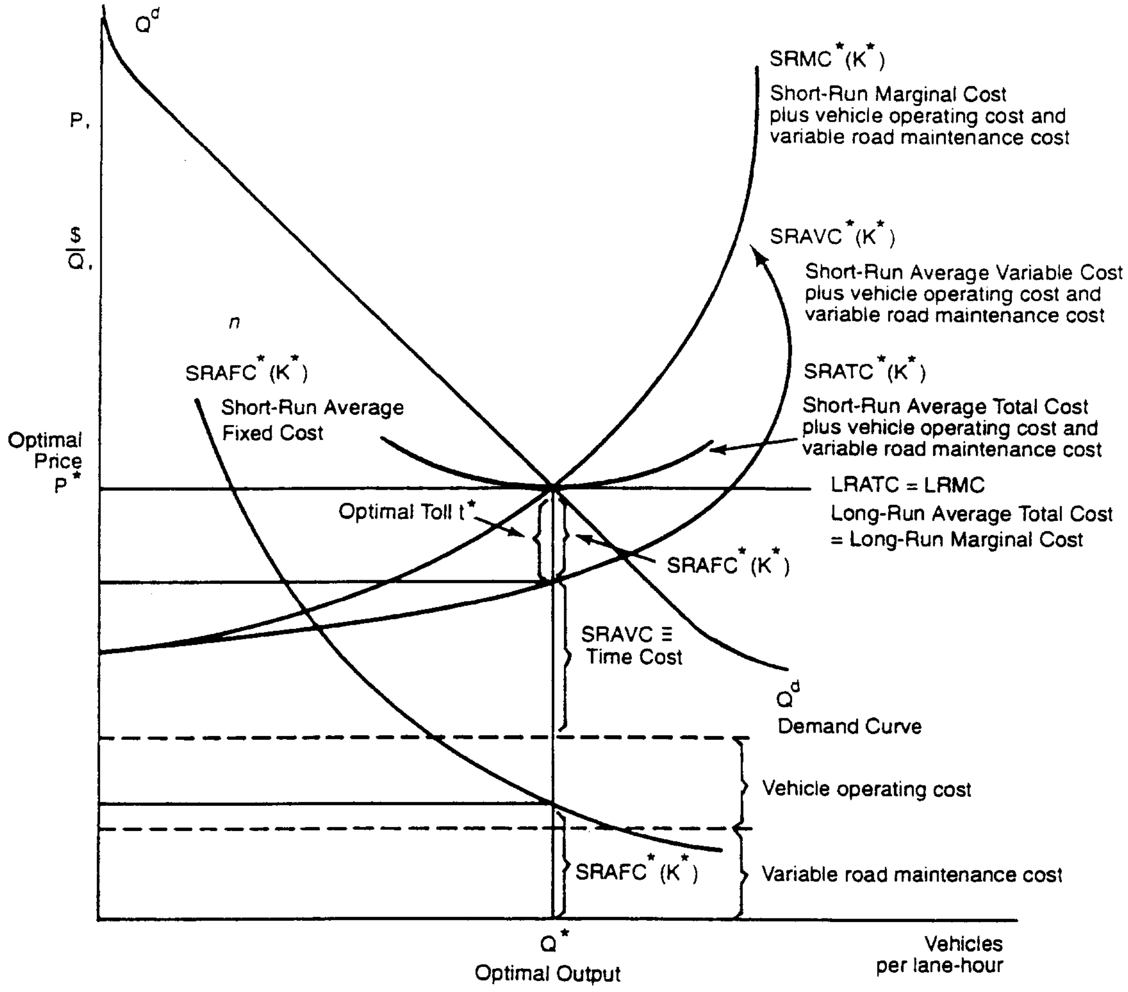


K^1 = (Non-optimal) Road capacity

t^1 = Optimal toll for a non-optimally built road

π^1 = Economic profit for a non-optimally built road

Figure 6
Long-Run Equilibrium of an Optimally Designed Road With Both Optimal Pricing and Optimal Investment


Q^d
SRMC*(K*)
Short-Run Marginal Cost plus vehicle operating cost and variable road maintenance cost
P,
$/Q,
SRAVC*(K*)
Short-Run Average Variable Cost plus vehicle operating cost and variable road maintenance cost
n
SRAFC*(K*)
Short-Run Average Fixed Cost
SRATC*(K*)
Short-Run Average Total Cost plus vehicle operating cost and variable road maintenance cost
Optimal Price P*
LRATC = LRMC
Long-Run Average Total Cost = Long-Run Marginal Cost
Optimal Toll t*
SRAFC*(K*)
SRAVC ≡ Time Cost
Q^d
Demand Curve
Vehicle operating cost
SRAFC*(K*)
Variable road maintenance cost
Q*
Optimal Output
Vehicles per lane-hour

t* = Optimal toll K* = Optimal road capacity

Figure 7(a)
Constant Returns to Scale with Road Divisibility: Doubling Optimal Road Capacity (K) and Traffic (Q) Result in Doubling Fixed Cost, Variable Cost and Total Cost (FC, VC, TC) and Toll Revenues ($t^*_L \cdot Q^*_L$)

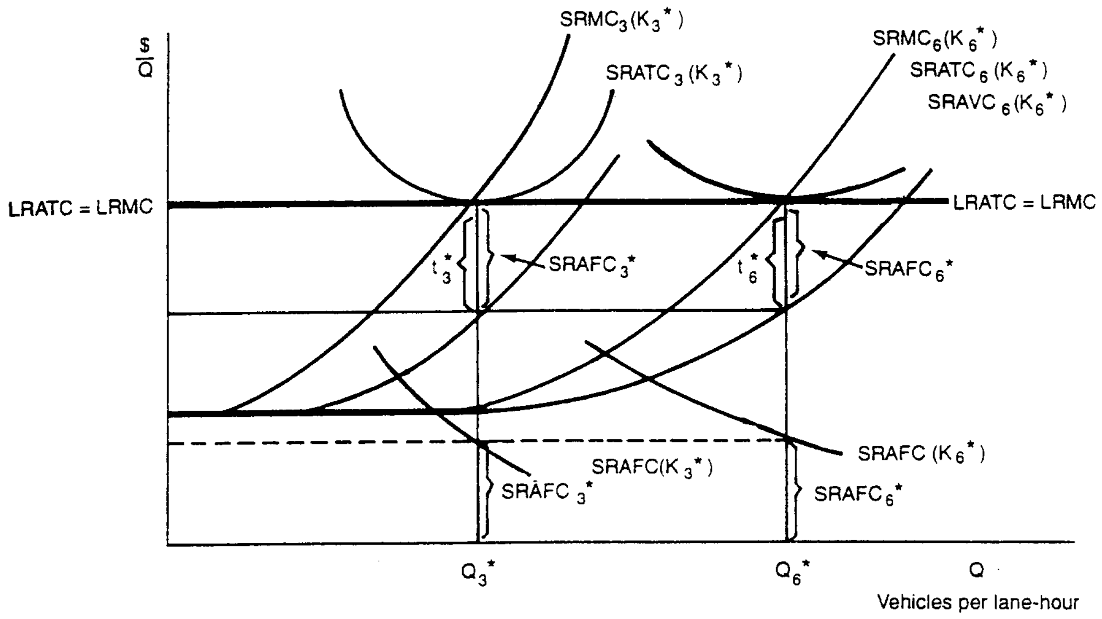
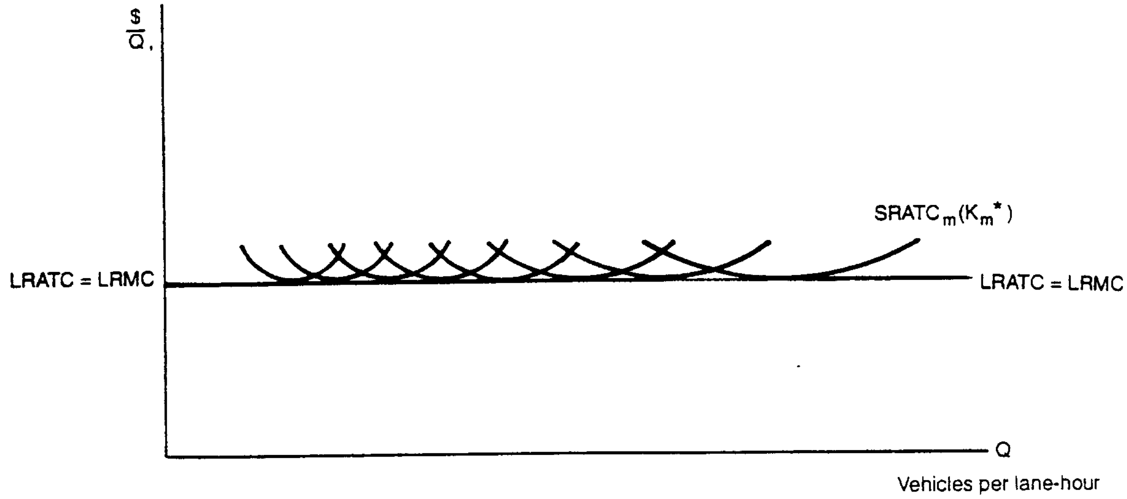


Figure 7(b)
The Relationship between Short-Run Average Total Cost and Long-Run Average Total Cost and Marginal Cost with Perfect Road Divisibility and Constant Returns to Scale



(*) Denotes that that variable is optimized
($_L$) Denotes an L-lane road
For example, K_3^* denotes the optimal capacity of an optimally-built 3-lane road
t_3^* denotes the optimal toll associated with an optimally-built 3-lane road

Figure 8(a)
Road Indivisibilities under Constant Returns

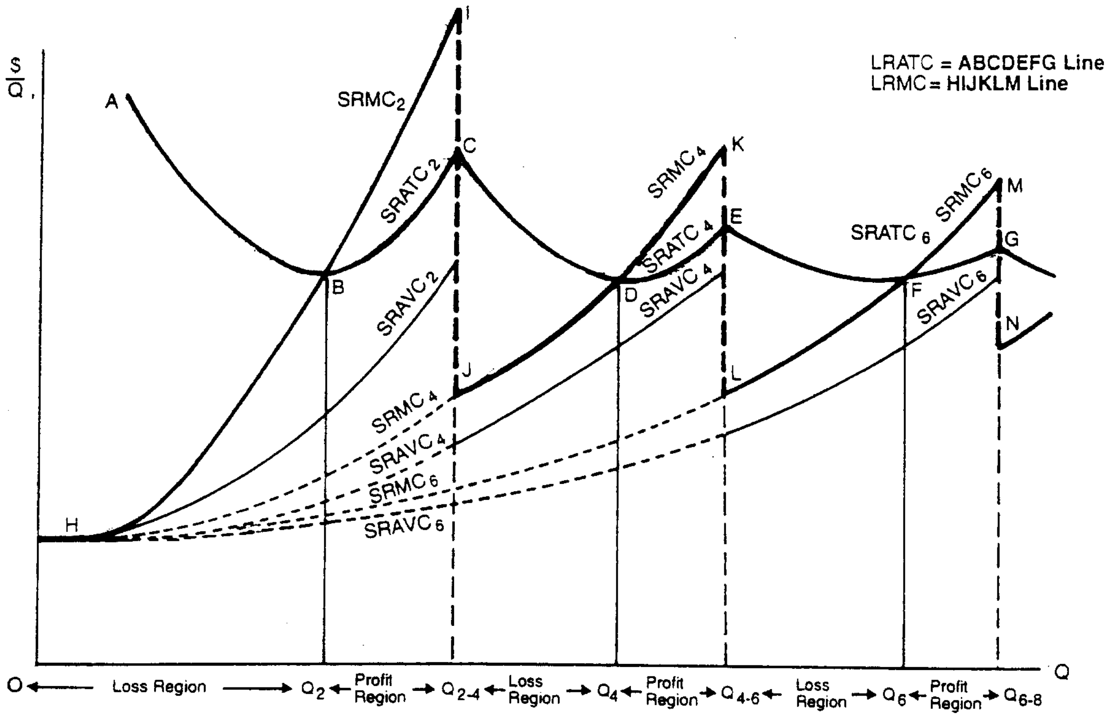


Figure 8(b)
Optimal Pricing and Investment with Indivisiblities:
Expansion from a 2-Lane Road to 4-Lane Road

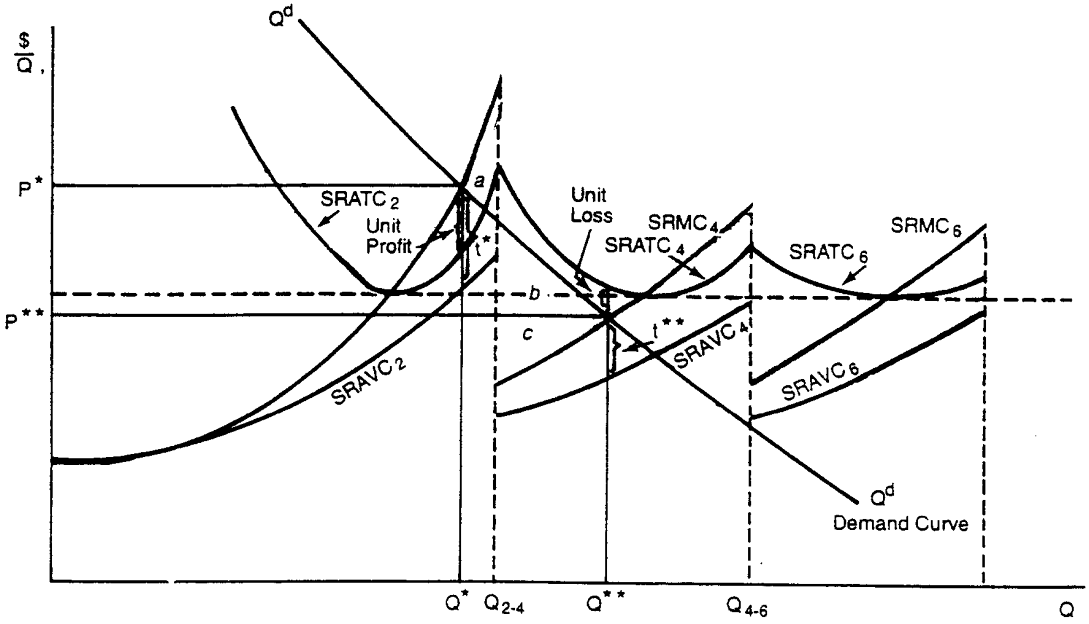


Figure 9
Economies and Diseconomies of Scale in the Provision of Road Capacity with the Growth of Travel Demand

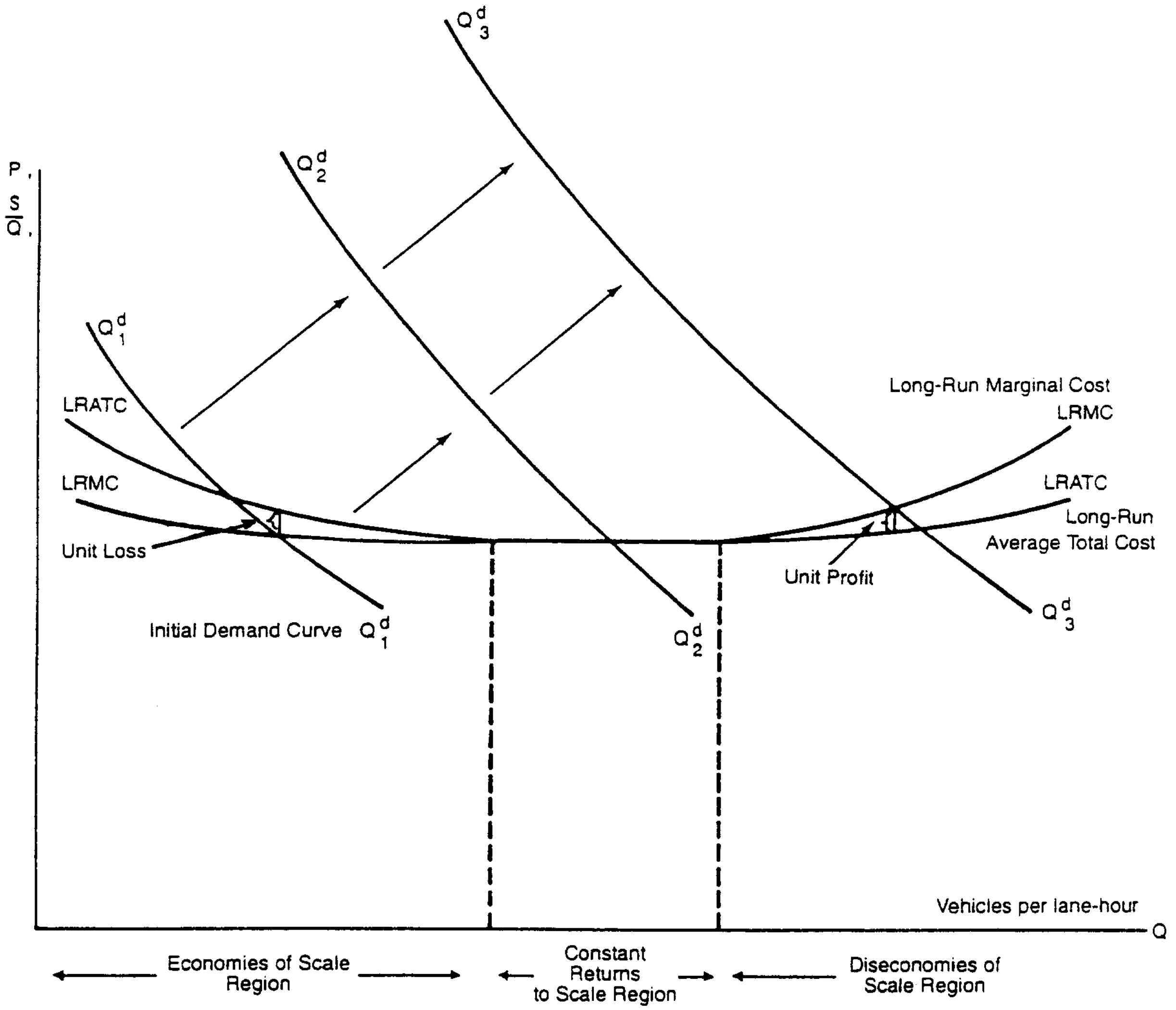


Figure 10
Doubling the Number of Streets--Road Capacity--Results in Quadrupling the Number of Intersections and Traffic Lights and Doubling Waiting Time

Figure 10(a) Original Scenario - Existing Street Configuration

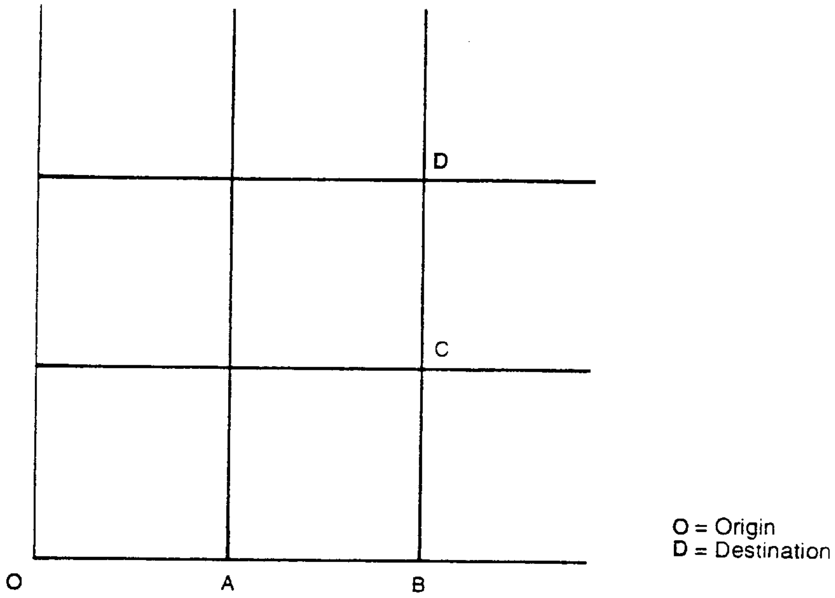


Figure 10(b) Final Scenario - Number of Streets are Doubled

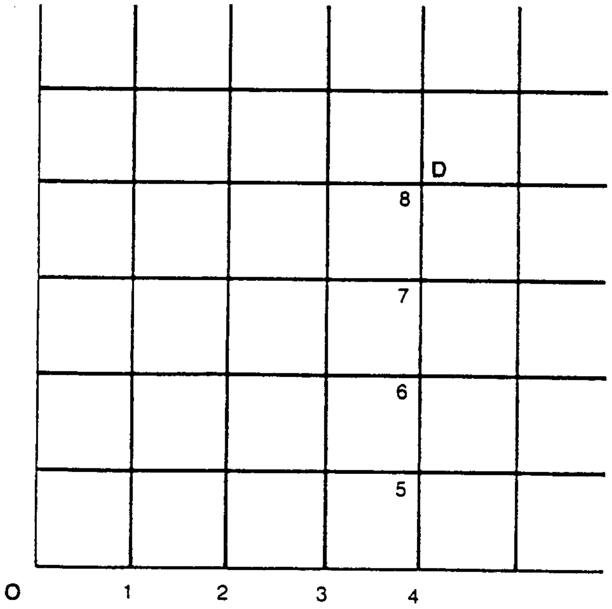


Figure 11
Diseconomies of Scale: Urban Roads Network with Perfect Divisibility
π = Economic Profit

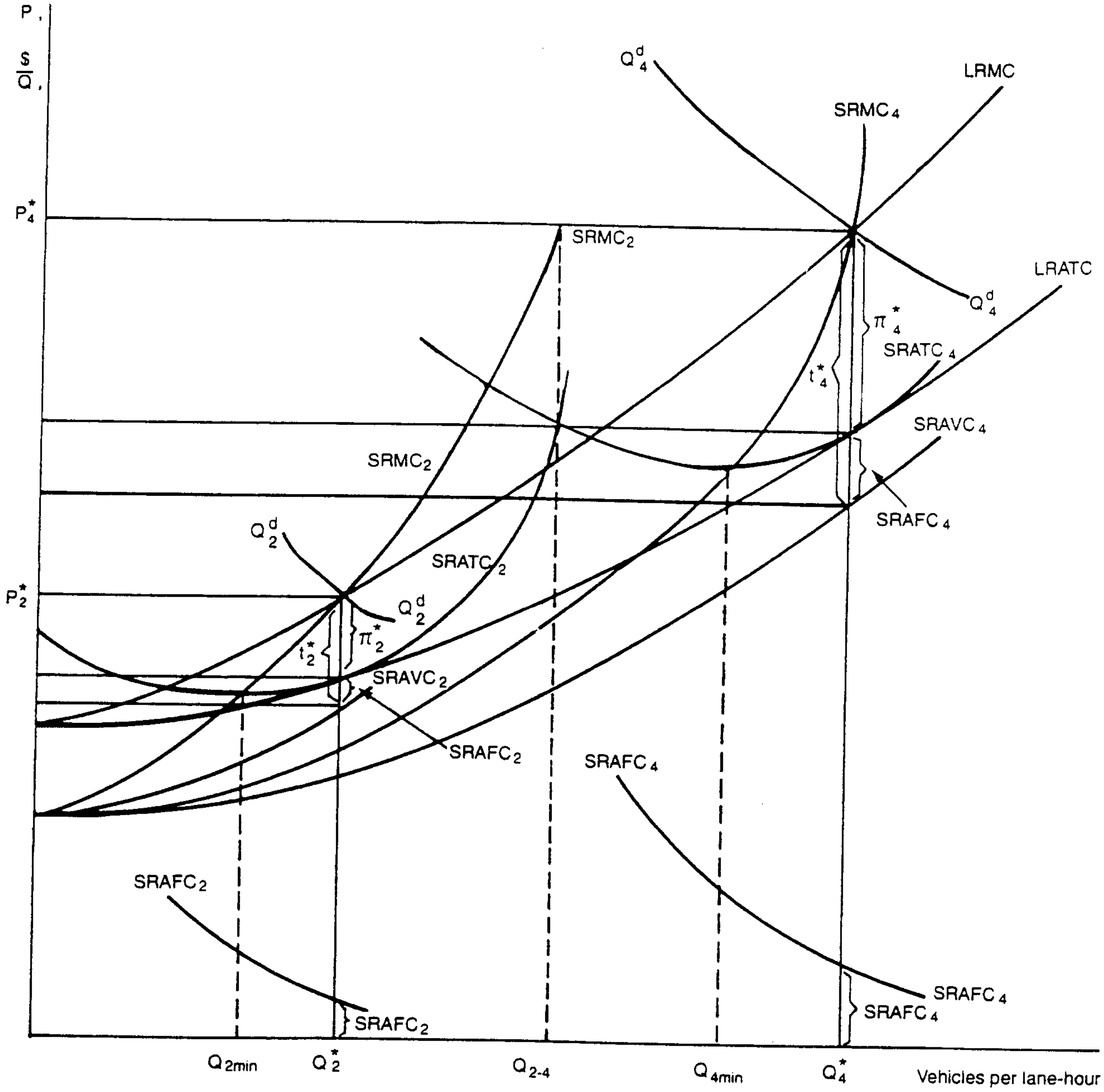


Figure 12
Economies of Scale: Rural Roads with Perfect Divisibility

L = Loss

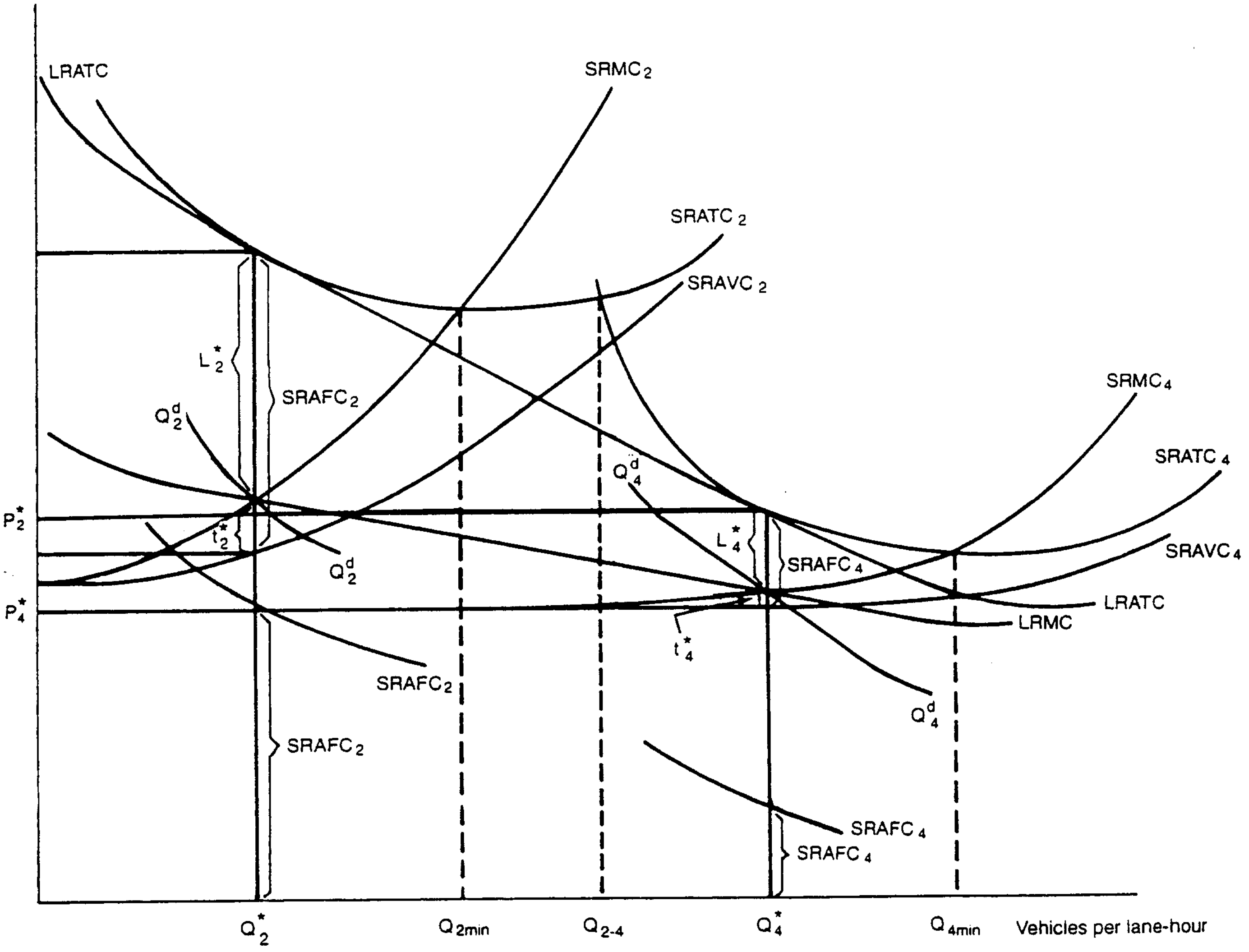


Figure 13(a)
Decreasing Returns to Scale and Extent of Indivisibilities

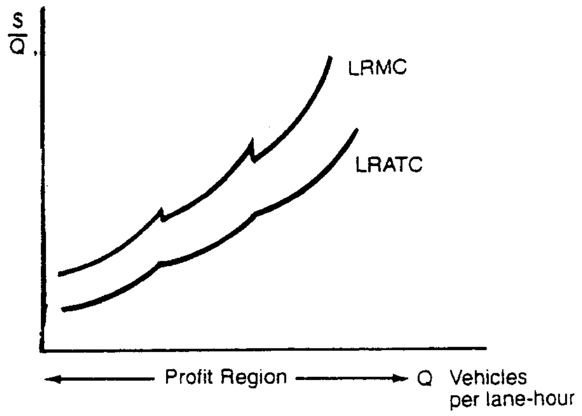


Figure 13(b)

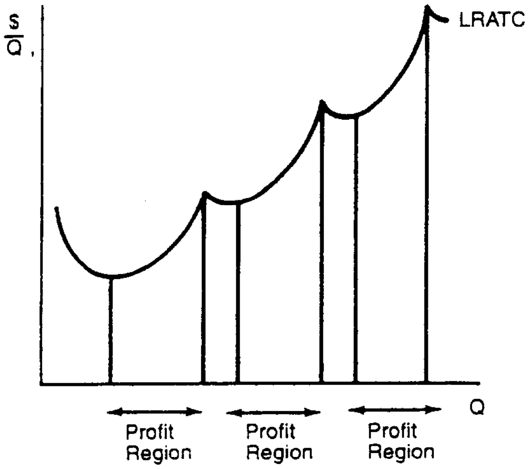


Figure 13(c)

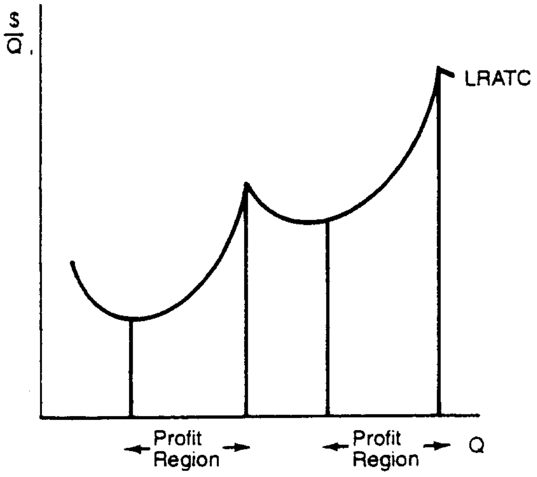


Figure 14(a)
Increasing Returns to Scale and Extent of Indivisibilities

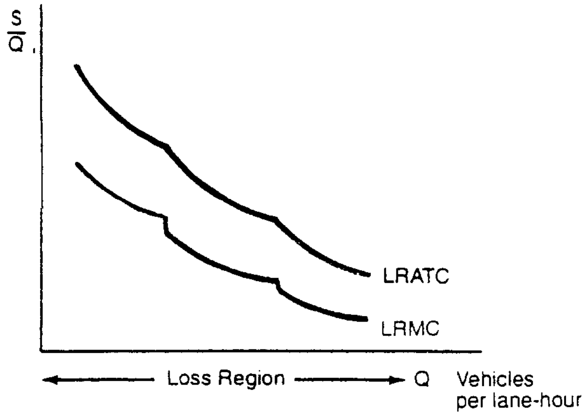


Figure 14(b)

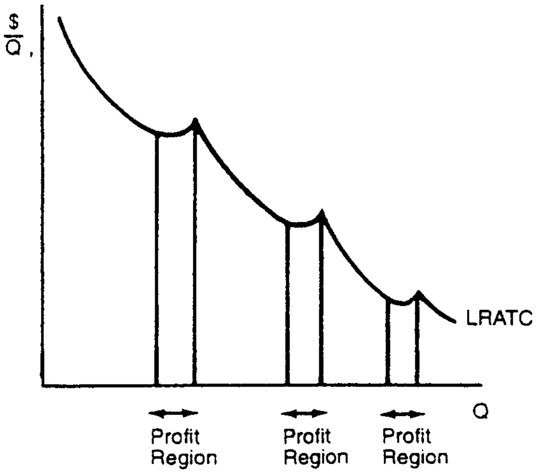
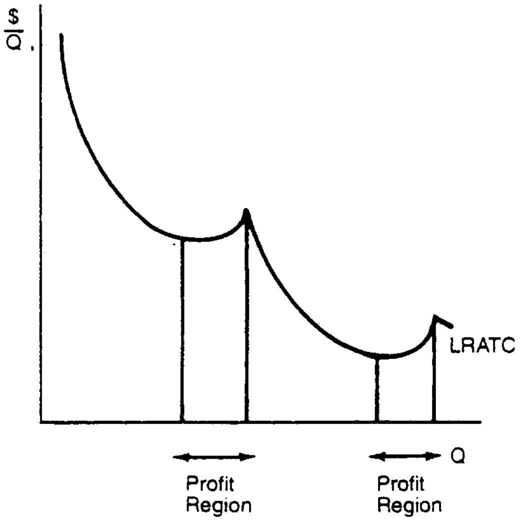


Figure 14(c)



REFERENCES

American Association of State Highway Officials (AASHO) (1960), “Road User Benefits for Highway Improvements,” a report by the Committee on Planning and Design Policies, American Association of State Highway Officials, Washington, D.C., pp. 1-152.

Bailey, Elizabeth E., and Ann F. Friedlaender (1982), “Market Structure and Multiproduct Industries,” Journal of Economic Literature, Vol. 20, No. 3, September, pp. 1024-1048.

Baumol, William J., and Kyu Sik Lee (1991), “Contestable Markets, Trade, and Development,” The World Bank Research Observer, Vol. 6, No. 1, January, pp. 1-17.

Baumol, William J., and Wallace E. Oates (1975,1988), The Theory of Environmental Policy, first edition, 1975, and second edition, 1988, Cambridge University Press, New York.

Baumol, William J., John C. Panzar and Robert D. Willig (1982,1988), Contestable Markets and the Theory of Industry Structure, first edition, 1982, and revised edition, 1988, Harcourt Brace Jovanovich and Academic Press, San Diego, California.

Beesley, Michael E., and Alan A. Walters (1970), “‘Some Problems in the Evaluation of Urban Road Investments’ and ‘A Reply’,” Applied Economics, Vol. 1, pp. 241-259, and Vol. 4, pp. 317-320.

Behbehani, Redha, V. Setty Pendakur and Alan T. Armstrong-Wright (1984), “Singapore Area Licensing Scheme: A Review of the Impact,” July, Water Supply and Urban Development Department, The World Bank, Washington, D.C., pp. 1-55.

Bennathan, Esra, and Alan A. Walters (1979), Port Pricing and Investment Policy for Developing Countries, published for the World Bank, Oxford University Press, Washington, D.C.

Bird, Richard M. (1976), “Charging for Public Services: A New Look at an Old Idea,” Canadian Tax Papers No. 59, December, Canadian Tax Foundation, Toronto. Excerpt reprinted as “Principles of Public Pricing,” in Pricing Policy for Development Management, Gerald M. Meier (1983), pp. 171-176, ed., EDI Series in Economic Development, Johns Hopkins University Press, published for The World Bank, Baltimore, Maryland.

Boadway, Robin W., and David E. Wildasin (1984), Public Sector Economics, second edition, Little, Brown and Company, Boston, Massachusetts.

Boiteux, Marcel (1960), “Peak-Load Pricing,” Journal of Business, April, Vol. 33, No. 2, pp. 157-179, translated from the French article “La tarification des demandes en pointe: application de la theorie de la vente au coût marginal,” by H. W. Izzard, as revised by the author, in the Revue Générale de l’Electricité, August 1949, Vol. 58, No. 8, pp. 321-340. Reprinted in Marginal Cost Pricing in Practice, ed. James R. Nelson, Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1964, pp.59-89.

Buchanan, James M. (1965), “An Economic Theory of Clubs,” Economica, Vol. 32, pp. 1-14.

Button, Kenneth J. (1982), Transport Economics, Heinemann Educational Books Limited, London.

Cameron, Michael (1991), “Transportation Efficiency: Tackling Southern California’s Air Pollution and Congestion,” Environmental Defense Fund, Regional Institute of Southern California, March, pp. 1-103.

Carbajo, José (1990), “Accident and Pollution Externalities in a System of Road User Charges,” unpublished Working Paper, December, The World Bank, Washington, D.C., pp. 1-22.

Catling, Ian, and Gabriel Roth (1987), “Electronic Road Pricing in Hong Kong: An Opportunity for Road Privatization?”, paper presented to the Transportation Research Board conference on ‘Roles of Private Enterprise and Market Processes in the Financing and Provision of Roads’, Transportation Research Record, No. 1107 of the Transportation Research Board, National Research Council, Washington, D.C., pp. 51-55.

Churchill, Anthony A. (1972), Road User Charges in Central America, with contributions by K. Huber, E. Meldau and Alan A. Walters, World Bank Staff Occasional Paper No. 15, Johns Hopkins University Press, Baltimore, Maryland.

Coase, Ronald H. (1960), “The Problem of Social Cost,” Journal of Law and Economics, Vol. 3, No. 1, October, pp. 1-44.

Coase, Ronald H. (1988), The Firm, the Market and the Law, University of Chicago Press, Chicago.

Cooter, Robert D. (1982), “The Cost of Coase,” Journal of Legal Studies, Vol. 11, No. 1, January, pp. 1-34.

Deaton, Angus S., (1987), “The Demand for Personal Travel in Developing Countries,” assisted by Duncan Thomas, Janet Neelin and Nikhilesh Bhattacharya, Discussion Paper No. INU-1, Infrastructure and Urban Development Department, The World Bank, Washington, D.C., pp. 1.1 - 9.26.

Downs, Anthony (1962), “The Law of Peak-Hour Expressway Congestion,” Traffic Quarterly, Vol. 16, July, pp. 393-409.

Dupuit, Jules (1844), “On the Measurement of the Utility of Public Works,” Annales des Ponts et Chaussées, 2nd series, Vol. 8, 1844, translated from the French essay “De la mesure de l’utilité des travaux publics,” by R. H. Barback for International Economic Papers, No. 2, 1952, pp. 83-110. Reprinted in Transport, ed., Denys Munby, Penguin Modern Economics, London, 1968, pp. 19-57.

Dupuit, Jules (1849), “On Tolls and Transport Charges,” Annales des Ponts et Chaussees, 2nd series, 4th part, 1849, pp. 207-248, translated from the French essay “De l’influence des pearges sur l’utilite des voies de communication,” by Elizabeth Henderson for International Economic Papers, No. 11, 1962, pp. 7-31.

Fitch, Lyle C., and Associates (1964), Urban Transportation and Public Policy, Chandler Publishing Company, San Francisco.

Gerlough, Daniel L., and Matthew J. Huber (1975), Traffic Flow Theory, A Monograph, Special Report 165, Transportation Research Board, National Research Council, Washington, D.C.

Glaister, Stephen (1981), Fundamentals of Transport Economics, Basil Blackwell, Oxford.

Goodwin, Phil B. (1989), “The Rule of Three: A Possible Solution to the Political Problem of Competing Objectives for Road Pricing,” Traffic Engineering and Control, Vol. 29, No. 10, October, pp. 495-497.

Gronau, Reuben (1991), “Are Ghana’s Roads Paying Their Way? Assessing Road Use Cost and User Charges in Ghana”, Policy Research and External Affairs Working Paper series, WPS 773. The World Bank, September 1991, pp. 1-44.

Gwilliam, Kenneth M., and Christopher A. Nash (1972), “Evaluation of Urban Road Investments -- ‘A Comment’ and ‘A Rejoinder’,” Applied Economics, Vol. 4, pp. 307-315 and p. 321.

Gwilliam, Kenneth M., and Peter J. Mackie (1975), Economics and Transport Policy, George Allen and Unwin, London.

Haight, Frank A. (1963), Mathematical Theories of Traffic Flow, Academic Press, New York.

Harral, Clell G., and Asif Faiz (1988), “Road Deterioration in Developing Countries: Causes and Remedies,” A World Bank Study Policy, with contributions by Esra Bennathan, Graham Smith, and Anil Bhandari, June, The World Bank, Washington, D.C., pp. 1-61.

Hau, Timothy D. (1986), “Distributional Cost-Benefit Analysis in Discrete Choice,” Journal of Transport Economics and Policy, Vol. 20, No. 3, 1986, pp. 313-338.

Hau, Timothy D. (1987), “Using a Hicksian Approach to Cost-Benefit Analysis in Discrete Choice: An Empirical Analysis of a Transportation Corridor Simulation Model,” Transportation Research, Vol. 21B, No. 5, 1987, pp. 339-357.

Hau, Timothy D. (1988), “Urban Transport,” in The Economic System of Hong Kong, eds., Henry C.Y. Ho and Laurence L.C. Chau, Asian Research Service, pp. 190-226.

Hau, Timothy D. (1989), “Road Pricing in Hong Kong: A Viable Proposal,” Built Environment, Vol. 15, Nos. 3/4, pp. 195-214.

Hau, Timothy D. (1990), “Electronic Road Pricing: Developments in Hong Kong 1983-89,” Journal of Transport Economics and Policy, Vol. 24, No. 2, May, pp. 203-214.

Hau, Timothy D. (1992), “Congestion Charging Mechanisms for Roads: An Evaluation of Current Practice,” World Bank Policy Research Working Paper Series WPS 1071, December, The World Bank, Washington, D.C., pp. 1-99.

Hayutin, Adele M. (1984), “Scale Economies in Highway Capacity: Empirical Evidence and Policy Implications,” unpublished Ph.D. dissertation, Department of Economics, University of California, Berkeley, California, pp. 1-182. Also available from University Microfilms International, Dissertation Information Service, Ann Arbor, Michigan.

Heggie, Ian G., and Vincy Fon (1991), “Optimal User Charges and Cost Recovery for Roads in Developing Countries,” Policy Research Working Paper Series, WPS 780. The World Bank, Washington, D.C., pp. 1-50.

Highway Research Board (1962), The AASHTO Road Test: Report 5, Pavement Research, HRB Special Report 61E, National Academy of Sciences, National Research Council, Washington, D.C., pp. 1-352.

Hoban, Christopher J. (1987), “Evaluating Traffic Capacity and Improvements to Road Geometry,” World Bank Technical Paper No. 74, October, The World Bank, Washington, D.C., October, pp. 1-145.

Johansen, Frida (1989), ed., "Earmarking, Road Funds and Toll Roads," A World Bank Symposium, Infrastructure and Urban Development Department, Report INU45, The World Bank, Washington, D.C., pp. 1-196.

Jones, Peter M. (1991), "Gaining Public Support for Road Pricing Through a Package Approach," Traffic Engineering and Control, Vol. 32, No. 4, April, pp. 194-196.

Jordan, W. John (1983a), "Heterogeneous Users and the Peak Load Pricing Model," The Quarterly Journal of Economics, Vol. 93, No. 1, February, pp. 127-138.

Jordan, W. John (1983b), "The Theory of Optimal Highway Pricing and Investment," Southern Economic Journal, Vol. 50, No. 2, October, pp. 560-564.

Jordan, W. John (1985), "Capacity Costs, Heterogeneous Users, and Peak-Load Pricing," The Quarterly Journal of Economics, Vol. 95, November, pp. 1335-1337.

Julius, DeAnne, and Adelaida P. Alicbusan (1989), "Public Sector Pricing Policies: A Review of Bank Policy and Practice," April, Policy, Planning, and Research, Working Paper Series 49, The World Bank, Washington, D.C.

Keeler, Theodore E., Kenneth A. Small and Associates (1975), with contributions by George S. Cluff, Jeffrey K. Finke, Leonard A. Merewitz, Randall J. Pozdena and Philip A. Viton, The Full Costs of Urban Transport: Part III, Automobile Costs and Final Intermodal Cost Comparisons, Monograph No. 21, Institute of Urban and Regional Development, University of California, Berkeley, pp. 1-154.

Keeler, Theodore E., and Kenneth A. Small (1977), "Optimal Peak-Load Pricing, Investment, and Service Levels on Urban Expressways," Journal of Political Economy, Vol. 85, No. 1, January, pp. 1-25.

Kim, H. Youn (1987), "Economies of Scale and Scope in Multiproduct Firms: Evidence from U.S. Railroads," Applied Economics, Vol. 19, June, pp. 733-741.

Knight, Frank H. (1924), "Some Fallacies in the Interpretation of Social Cost," Quarterly Journal of Economics, Vol. 38, August, pp. 582-606. Reprinted in Readings in Welfare Economics, eds., Kenneth J. Arrow and Tibor Scitovsky, American Economic Association Series, Richard D. Irwin, Inc., 1969, pp. 213-227.

Kraus, Marvin C., Herbert D. Mohring and Thomas Pinfold (1976), "The Welfare Costs of Non-Optimum Pricing and Investment Policies for Freeway Transportation," American Economic Review, Vol. 66, No. 4, September, pp. 532-547.

Kraus, Marvin C. (1981a) "Scale Economies Analysis for Urban Highway Networks," Journal of Urban Economics, Vol. 9, No. 1, January, pp. 1-22.

Kraus, Marvin C. (1981b), "Indivisibilities, Economies of Scale, and Optimal Subsidy Policy for Freeways," Land Economics, Vol. 57, No. 1, February, pp. 115-121.

Lee, Douglass B. (1982), "Net Benefits from Efficient Highway User Charges," Transportation Research Record, Vol. 858, Highway Research Board, National Research Council, Washington, D.C., pp. 14-20.

Marshall, Alfred (1920), Principles of Economics, 8th edition, McMillan Press Limited, London.

May, Adolf D. (1990), Traffic Flow Fundamentals, Prentice-Hall, Englewood Cliffs, New Jersey.

May, Anthony D., and Keith E. Gardner (1990), "Transport Policy for London in 2001: The Case for an Integrated Approach," Transportation, Vol. 16, pp. 257-277.

McCleary, William A. (1991), “The Earmarking of Government Revenue: A Review of Some World Bank Experience,” The World Bank Research Observer, Vol. 6. No. 1, January, pp. 81-104.

Meyer, John R., John F. Kain and Martin Wohl (1965), The Urban Transportation Problem, Harvard University Press, Cambridge, Massachusetts.

Meyer, John R., and José A. Gómez-Ibáñez (1981), Autos, Transit, and Cities, Harvard University Press, Cambridge, Massachusetts.

Mills, David E. (1981), “Ownership Arrangements and Congestion-Prone Facilities,” American Economic Review, Vol. 71, No. 3, June, pp. 493-502.

Ministry of Transport (1964), Road Pricing: The Economic and Technical Possibilities, Her Majesty’s Stationery Office, London, pp. 1-61.

Mishan, Ezra J. (1988), Cost-Benefit Analysis, fourth edition, Unwin and Hyman, London.

Mohring, Herbert D., and Mitchell Harwitz (1962), Highway Benefits: An Analytical Framework, Northwestern University Press, Evanston, Illinois.

Mohring, Herbert D. (1965), “Urban Highway Investments,” in Measuring Benefits of Government Investments, ed., Robert Dorfman, papers presented at a Conference of Experts held November 7-9, 1963, The Brookings Institution, Washington, D.C., pp. 231-291.

Mohring, Herbert D. (1970), “The Peak Load Problem with Increasing Returns and Pricing Constraints,” American Economic Review, Vol. 60, No. 4, September, pp. 693-705.

Mohring, Herbert D. (1975), “Pricing and Transportation Capacity,” in ‘Better Use of Existing Transportation Facilities,’ Special Report 153, Transportation Research Board, National Research Council, Washington, D.C., pp. 183-195.

Mohring, Herbert D. (1976), Transportation Economics, Ballinger Press, Cambridge, Massachusetts.

Morrison, Steven A. (1986), “A Survey of Road Pricing,” Transportation Research A, Vol. 20A, No. 2, March, pp. 87-98.

Musgrave, Richard A., and Peggy B. Musgrave (1989), Public Finance in Theory and Practice, fifth edition, McGraw-Hill, New York.

Nash, Christopher A. (1976), Public versus Private Transport, Macmillan Studies in Economics, The Macmillan Press Limited, London.

Neutze, G.M. (1966), “Investment Criteria and Road Pricing,” Manchester School of Economics and Social Studies, Vol. 34, No. 1, January, pp. 63-73.

Newbery, David M. G. (1986), “Estimating Urban Congestion Costs,” mimeo, August 12, Fiscal Affairs Department, International Monetary Fund, Washington, D.C., pp. 1-39.

Newbery, David M. G. (1988a), “Road Damage Externalities and Road User Charges,” Econometrica, Vol. 56, No. 2, March, pp. 295-316.

Newbery, David M. G. (1988b), “Road User Charges in Britain,” The Economic Journal Supplement (Conference 1987), Vol. 98, No. 390, pp. 161-176.

Newbery, David M. G. (1988c), “Charging for Roads,” The World Bank Research Observer, Vol. 3, No. 2, July, pp. 119-138.

Newbery, David M.G. (1989), “Cost Recovery from Optimally Designed Roads,” Economica, Vol. 56, May, pp. 165-185.

Newbery, David M. G. (1990), "Pricing and Congestion: Economic Principles Relevant to Pricing Roads," Oxford Review of Economic Policy, Special Issue on Transport, Vol. 6, No. 2, Summer, pp. 22-38.

Newbery, David M. G., Gordon A. Hughes, William D.O. Paterson and Esra Bennathan (1988), "Road Transport Taxation in Developing Countries: The Design of User Charges and Taxes for Tunisia," World Bank Discussion Paper No. 26, April, The World Bank, Washington, D.C., pp. 1-94.

Oum, Tae Hoon, and Michael W. Trethaway (1988), "Ramsey Pricing in the Presence of Externality Costs," Journal of Transport Economics and Policy, Vol. 22, No. 3, September, pp. 307-318.

Oum, Tae Hoon, and Yimin Zhang (1990), "Airport Pricing: Congestion Tolls, Lumpy Investment and Cost Recovery," Journal of Public Economics, Vol. 43, No. 3, December, pp. 353-374

Paterson, William D.O. (1987), Road Deterioration and Maintenance Effects: Models for Planning and Management, Vol. III, Highway Design and Maintenance (HDM) Standards, Johns Hopkins University Press for the World Bank, Baltimore, Maryland.

Pigou, Arthur C. (1920), The Economics of Welfare, first edition, Macmillan Company, London.

Pozdena, Randall J., Ronald Schmidt and Deborah Martin (1990), "Market-Based Solutions to the Transportation Crisis: The Concept," A Two-Part Report, Bay Area Economic Forum, May, pp. 1-17.

Prest, Alan R. (1969), Transport Economics in Developing Countries: Pricing and Financing Aspects, The Trinity Press, Weidenfeld and Nicholson, London. Excerpt reprinted as "Public Transport Pricing and Cost Recovery," in Pricing Policy for Development Management, ed., Gerald M. Meier (1983), pp. 216-222, EDI Series in Economic Development, Johns Hopkins University Press, published for The World Bank, Baltimore, Maryland.

Roth, Gabriel J. (1987), The Private Provision of Public Services in Developing Countries, EDI Series in Economic Development, Oxford University Press, published for The World Bank, Washington, D.C.

Schiff, Maurice (1991), "New Findings in the Theory of Optimal Congestion Taxes, With an Application to Road Transportation", Revista de Análisis Económico, Vol. 6, No. 1, June, pp. 81-92.

Small, Kenneth A. (1983), "The Incidence of Congestion Tolls on Urban Highways", Journal of Urban Economics, Vol. 13, No. 1, January, pp. 90-111.

Small, Kenneth A. (1992), Urban Transportation Economics, Fundamentals of Pure and Applied Economics 51, Harwood Academic Publishers, Chur, Switzerland.

Small, Kenneth A., and Clifford M. Winston (1986), "Efficient Pricing and Investment Solutions to Highway Infrastructure Needs," American Economic Review: AEA Papers and Proceedings, 1985, Vol. 76, No. 2, May, pp. 165-169.

Small, Kenneth A., and Clifford M. Winston (1988), "Optimal Highway Durability," American Economic Review, Vol. 78, No. 3, June, pp. 560-569.

Small, Kenneth A., and Feng Zhang (1988), "A Reanalysis of the AASHO Road Test Data: Rigid Pavements," unpublished paper, December, pp. 1-33.

Small, Kenneth A., Clifford M. Winston and Carol A. Evans (1989), Road Work: A New Highway Pricing and Investment Policy, The Brookings Institution, Washington, D.C.

Starkie, David N. M. (1982), “Road Indivisibilities: Some Observations,” Journal of Transport Economics and Policy, Vol. 16, No. 3, September, pp. 259-266.

Strotz, Robert H. (1964a), “Principles of Urban Transportation Pricing,” Highway Research Record, No. 47, paper presented at the 43rd Annual Meeting of the Highway Research Board, January 13-17, 1964, National Research Council, Washington, D.C., pp. 113-121.

Strotz, Robert H. (1964b), “Urban Transportation Parables,” in The Public Economy of Urban Communities, ed., Julius Margolis, papers presented at the Second Conference on Urban Public Expenditures held February 21-22, 1964, Resources for the Future, Johns Hopkins University Press, Baltimore, Maryland, 1965, pp. 127-169.

Talvitie, Antti P., and Associates (1978), “Policy Analysis of a Transportation Corridor,” ITS Research Report UCB-ITS-RR-78-5, Urban Travel Demand Forecasting Project, Final Report Series, Vol. 10, Institute of Transportation Studies, University of California, Berkeley, California, pp. 1-306.

Tanner, J.C. (1963), “Pricing the Use of Roads -- a Mathematical and Numerical Study,” Road Research Laboratory Note No. L/N 319, Department of Scientific and Industrial Research, Hammondsworth, England, pp. 317-345.

Thomson, J. Michael (1970), “Some Aspects of Evaluating Road Improvements in Congested Areas,” Econometrica, Vol. 38, No. 2, March, pp. 298-310.

Thomson, J. Michael (1974), Modern Transport Economics, Richard Clay (The Chaucer Press) Limited, Penguin Modern Economics, Bungay, Suffolk.

Transportation Research Board (1985), Highway Capacity Manual, Special Report 209, Transportation Research Board, National Research Council, Washington, D.C., p. 1-1 to 14-4.

Transpotech (1983), “Electronic Road Pricing in Hong Kong, Pilot Stage: Economic Background and An Example,” Technical Paper 1, October, background paper prepared for the Hong Kong Government, pp. 1-14.

Transpotech (1985), “Electronic Road Pricing Pilot Scheme: Main Report”, May, Main Report prepared for the Hong Kong Government, pp. 1.1-4.40.

Vickrey, William S. (1965), “Pricing as a Tool in Coordination of Local Transportation,” in Transportation Economics, Proceedings of a Universities-National Bureau Committee Conference on Transportation Economics, National Bureau of Economic Research, New York, pp. 275-296.

Vickrey, William S. (1968), “Congestion Charges and Welfare: Some Answers to Sharp’s Doubts,” Journal of Transport Economics and Policy, Vol. 2, No. 1, January, pp. 107-118.

Vickrey, William S. (1977), “The City as a Firm,” in The Economics of Public Services, ed., Martin S. Feldstein and Robert P. Inman, Proceedings of a Conference held by the International Economic Association at Turin, Italy, published by the Macmillan Press, Ltd., London, pp. 334-343.

Vickrey, William S. (1985), “The Fallacy of Using Long-Run Cost for Peak-Load Pricing,” The Quarterly Journal of Economics, Vol. 95, November, pp. 1331-1334.

Walters, Alan A. (1954), “Track Costs and Motor Taxation,” Journal of Industrial Economics, Vol. 2, No. 2, April, pp. 135-146.

Walters, Alan A. (1961a), “The Theory & Measurement of Private & Social Cost of Highway Congestion,” Econometrica, Vol. 29, No. 4, October, pp. 676-699. Reprinted in Readings in Urban Economics, eds., Matthew Edel and Jerome Rothenberg, (1973), Macmillan, New York, pp. 417-437.

Walters, Alan A. (1961b), “Empirical Evidence on Optimum Motor Taxes for the United Kingdom,” Applied Statistics, Vol. 10, No. 3, November, pp. 157-169.

Walters, Alan A. (1968), The Economics of Road User Charges, World Bank Occasional Paper Number 5, International Bank for Reconstruction and Development, Johns Hopkins University Press, Baltimore, Maryland.

Walters, Alan A. (1987), “Congestion,” in The New Palgrave: A Dictionary of Economics, ed., John Eatwell, Murray Milgate and Peter Newman, The Macmillan Press Limited, London, Vol. 1, pp. 570-573.

Watson, Peter L., and Edward P. Holland (1976), “Congestion Pricing: The Example of Singapore”, Finance and Development, Vol. 13, No. 1, March, pp. 20-23.

Watson, Peter L., and Edward P. Holland (1978), “Relieving Traffic Congestion: The Singapore Area License Scheme,” World Bank Staff Working Paper No. 281, June, The World Bank, Washington, D.C., pp. 1-286.

Winston, Clifford M. (1985), “Conceptual Developments in the Economics of Transportation: An Interpretive Survey,” Journal of Economic Literature, Vol. 23, No. 1, March, pp. 54-87.

Winston, Clifford M. (1991), “Efficient Transportation Infrastructure Policy,” Journal of Economic Perspectives, Vol. 5, No. 1, Winter, pp. 113-127.

Wohl, Martin, and Chris Hendrickson (1984), Transportation Investment and Pricing Principles: An Introduction for Engineers, Planners and Economists, Wiley Construction and Engineering Series, Wiley-Interscience, New York.

World Bank Operational Manual Statement (1977), “Cost Recovery Policies for Public Sector Projects: General Aspects,” No. 2.25, March, The World Bank, Washington, D.C., pp. 1-7.

World Bank (1986), “Urban Transport: A World Bank Policy Study,” Water Supply and Urban Development Department, The International Bank for Reconstruction and Development, Washington, D.C., pp. 1-61.

World Bank (1991), “Urban Policy and Economic Development: An Agenda for the 1990’s,” Urban Division, Infrastructure and Urban Development Department, The International Bank for Reconstruction and Development, Washington, D.C., April, pp. 1-87.

Zettel, Richard M., and Richard R. Carll (1964), “The Basic Theory of Efficiency Tolls: The Tolled, The Tolled-Off and The Un-Tolled,” Highway Research Record, No. 47, paper presented at the 43rd Annual Meeting of the Highway Research Board, January 13-17, 1964, National Research Council, Washington, D.C., pp. 46-65.

Preliminary Draft: April 19, 1991

Revised Draft: October 7, 1991

Final Draft: December 23, 1992