"Alleviating Traffic Congestion:
Alternatives to Road Pricing"

by

Richard Arnott

Working Paper No. 282
September 1994

ALLEVIATING TRAFFIC CONGESTION:

ALTERNATIVES TO ROAD PRICING

Richard Arnott†

September 1994

Revised version

Preliminary draft: Please do not cite or quote without the permission of the author

For presentation at the Taxation, Resources and Economic Development (TRED) conference on "Alternative Strategies for Managing Externalities," Lincoln Institute of Land Policy, Cambridge, Massachusetts, September 30/October 1, 1994.

† Department of Economics
Boston College
Chestnut Hill, MA 02167

Alleviating Traffic Congestion: Alternatives to Road Pricing

I am a believer but I am also a realist. In an ideal world, full efficiency would be achieved by marginal (social) cost pricing, and equity through lump–sum redistribution. In such an ideal world, travelers would make efficient choices since they would face the true social costs of their actions on every margin of travel choice -- there would be perfect congestion pricing. And since prices would equal social costs, governments would be able to determine precisely the optimal capacity of all transport facilities.

But we do not live in an ideal world. Whatever its virtues, congestion pricing will not be introduced soon in the United States, at least on a significant scale. When and where it is first introduced, it will take a crude form, which falls a long way short of having auto drivers face the true social costs of their actions on every margin of choice. Even with reasonably sophisticated electronic road pricing, many margins of driver choice will still be distorted.

Thus, while full marginal cost pricing — of which perfect congestion pricing is one element — is the ideal to be strived for, realistically we must muddle through as best we can in the untidy world of the second best. In the context of urban travel, this entails a messy mix of policies, including some forms of road pricing.

In this paper I do not argue against congestion pricing — after all, I am a believer. But I do think there is the danger that striving for the first best may distract us from the second best, that in researching and advocating congestion pricing we take time away from consideration of the other policies that may be at least as practical and practicable.

In section 1, I elaborate on my views on congestion pricing. The bottom line is that wherever, whenever, and in whatever form congestion pricing is introduced, a broad array of other policies are and will be needed for efficient urban traffic management. In section 2, I run through alternative urban traffic management policies. And in section 3, I provide an overview.

1. Congestion Pricing

That road travel entails a congestion externality which can be internalized by means of tolling was recognized by Dupuit (1844) and has been widely–known and well–understood by economists since Knight (1924) and Pigou (1912). Today, almost all economists favor congestion pricing in principle, though some have doubts about its practicality.

Almost forty years ago, Vickrey (1959) drew up a detailed scheme for electronic auto congestion pricing in Washington, D. C. While the scheme he proposed probably seemed futuristic at the time, today the technology for electronic auto congestion pricing is available and reliable, and with widespread use should be inexpensive. The technological and economic feasibility of electronic road pricing was persuasively demonstrated in a well–designed experiment in Hong Kong in the early 1980's (Catling and Harbord (1985)).

The wisdom of congestion pricing is accepted, at least among economists, and the technology is there to implement it. Why then has not a single jurisdiction introduced electronic road pricing?

1.1 Political opposition to road pricing

The answer is that politicians have regarded it, with justification, as political poison.

The Hong Kong experiment was sufficiently successful that serious consideration was given to the implementation of electronic road pricing in Hong Kong. But political opposition was mounted on the grounds that electronic road pricing constitutes an invasion of privacy that can be abused by the government, and resulted in the proposal being withdrawn (Borins (1986)). Under the proposed Hong Kong system, the government would have collected information on the travel history of all its residents. The potential for abuse is obvious — the government could, for instance, discredit an opposition politician who travels extensively or parks frequently in the red light district. This argument against electronic road pricing is hardly a damning one, since the system can easily be designed not to record travel histories.

But the opposition to congestion pricing is more deep–seated than the Hong Kong example suggests. Congestion pricing entails the pricing of a service — travel on city streets — that was previously provided free. Relatedly, people tend to view free travel on city streets as a right, or at least as a right earned from paying local taxes. As well, citizens are justifiably cynical about political motives, and tend to view proposals for congestion pricing as another tax grab, another incursion by Leviathan. Also, the promised benefits of congestion pricing are intangible. Road pricing would reduce congestion but, given the substantial day–to–day variations in traffic, the reduction might not be obvious. Finally, opponents of congestion pricing argue, probably correctly, that the policy is regressive. Professionals can alter their travel schedules to avoid paying peak tolls; clerical and factory workers cannot. As well, congestion pricing entails paying with money rather than with time.

A successful campaign for congestion pricing will have to overcome all these political obstacles. The proponents of congestion pricing are developing more political savvy. They are attempting to educate the public concerning the virtues of congestion pricing. Also, they are proposing to earmark the revenues raised from congestion tolls for the expansion and improvement of transportation infrastructure, or to rebate the revenue collected, perhaps via the income tax, in such a way that all major groups in the population benefit (Jones (1991), Small (1992)). There are indications that this campaign is having at least some success. A recent poll found that the majority of commuters in London would favor congestion pricing if the revenues collected were used to upgrade the transport system (Bayliss (1992)).[1]

I forecast that the political opposition to congestion pricing will be stronger in the U.S. than in any other industrialized country, and therefore that the U.S. will be among the last industrialized countries to adopt urban auto congestion pricing on a large scale. There are three reasons for my pessimism. First, Americans have a more individualistic and libertarian political culture; second, U.S. cities, with the exception of Manhattan, are significantly less congested that cities of similar size in Europe and Asia; and third, the extreme jurisdictional fragmentation in the U.S. is going to make congestion pricing significantly harder to implement.

---

[1]The adoption of congestion pricing in Oslo, Bergen, and Trondheim provides only weak evidence that the resistance to congestion pricing is weakening. Those cities did not introduce congestion pricing because they were persuaded of its effectiveness in reducing the volume of traffic or in spreading out the rush hour. Rather, the federal government refused to provide funding for the access roads, and the only way the cities had to raise revenue to finance the roads' expansion and improvement was tolling (Ramjerdi (1994)).

The advocates of congestion pricing have recently shown more political awareness. They have not, however, shown the same sensitivity with respect to practical problems of implementation. Unless these are successfully addressed, a system could be so seriously flawed or be so unpopular that it would be withdrawn. The congestion pricing schemes in the Norwegian cities of Oslo, Bergen, and Trondheim are relatively crude, since they cover only major access roads and entail little variation in the toll by time of day. I would guess that these tolling systems have been (apparently) successful because the topography is such that major access roads are hard to avoid and because Norwegians have a collective spirit. But if such a tolling system were applied in the United States, at least a significant minority of commuters would take delight in taking detours on secondary roads around the tolling points and community residents would be on the warpath. This problem would be avoided if there were electronic road pricing on all roads. But comprehensive electronic road pricing will have problems of its own. First, there is the problem of who will pay for the installation of the metering hardware for a car. A sure way to guarantee unpopularity would be to require that car owners do so;[2] all expenses associated with the metering hardware should come out of government revenue, and when the system is operational, out of toll revenue. Second, it will be vital that any system be user–friendly. Big city life is stressful, largely due to the many petty frustrations encountered. Bureaucratic hassles with billing, including overbilling, bureaucratic delays in getting the metering hardware installed and maintained, and unjust fines due to malfunctioning equipment, could condemn any system, however sound its basic design, to an early demise. Third, the billing system will have to be well–thought–out; prepayment cards which can be bought at convenience stores — like lottery tickets — and which would display how much of the card has been used up would seem a sensible option. Fourth, the metering hardware will have to be tamper–proof.

One of the thorniest problems in the implementation of any system of congestion pricing is how to deal with non–residents. If the system were to entail tolling on only major access roads, then there could be toll booths where non–residents would purchase day passes. But a large proportion of non–residents might then choose to enter the city on secondary roads, reasoning that the jurisdiction would not bother to follow up on fine collection — just as today many non–residents do not pay parking tickets. This problem could be mitigated by cooperative arrangements between jurisdictions or through regionalization of tolling. With electronic road pricing, the problem would be trickier.

[2] Kingston, Canada, finally adopted taxi meters only about five years ago. One of the contentious issues in moving from a zone to a meter system was who should bear the cost of the meters.

Presumably non–residents would have to purchase some gizmo — a decal, a smart card, etc. — that would identify the car as non–resident and which would be valid for a specified period. Such a system would be less than ideal since the toll paid by non–residents would be unrelated to their patterns of travel within the jurisdiction. In small jurisdictions, where a large proportion of road traffic is non–resident, this could seriously undermine the effectiveness of the tolling system.

Another problem is that jurisdictions will have an incentive to delay implementation of congestion pricing because of uncertainty concerning which system to adopt — the costs of irreversibility. Each jurisdiction will have an incentive to learn from the mistakes of other jurisdictions, and then to adopt the system that is the most cost–effective and hassle–free, particularly since the adoption of any system will entail sizable fixed costs. Smaller jurisdictions have an additional incentive to delay because of adoption externalities. The inner cities of Boston would, for instance, realize that it was in their mutual interests to adopt a common system, but there would be disagreement concerning what system to adopt, how to share costs and administration, etc. These problems would be less severe if tolling were managed at the metropolitan level.

I do not want to paint too gloomy a picture. There are messy, practical problems of implementation with any policy, particularly when there is little experience to draw upon. But with a policy as politically sensitive as congestion pricing, it is imperative that these problems be thought through and resolved prior to implementation.

Reformists tend to be dreamers. I remember a discussion with a group of transportation scientists in the early days of research on vehicle information systems. They envisaged widespread adoption of vehicle information systems by 1995. I was highly skeptical. I am equally skeptical of the rapid adoption of congestion pricing, and think that electronic road pricing will not be implemented in the United States for decades. I envisage the East Asian countries introducing electronic road pricing first, because of their receptiveness to technological innovation, their collectivist culture, their relatively high degree of political centralization, and the severity of traffic congestion in their cities. Northern Europe and France will follow. And major American cities will start adopting only after systems elsewhere have proved their effectiveness.

Congestion pricing, in a crude form, followed by electronic road pricing, will come. It has to, since it is the only sensible, long–run solution to the urban traffic

problem. But, especially in North America, it will not come soon, and in the interim we must come up with alternative policies to deal with traffic problems.

1.2 Perfect versus imperfect congestion tolling

We tend to speak rather carelessly about congestion pricing. On one hand, at a theoretical level, we use the term to refer to Pigouvian pricing — imposing a tax or toll equal to the congestion externality, evaluated at the optimum — so that price equals marginal social cost. On the other hand, at a policy level, we use the term to refer to any policy that raises the monetary cost of traveling in congested traffic. Let me refer to the former as perfect congestion tolling, and the latter as imperfect congestion tolling when price does not equal marginal social cost. The main point I want to make in this subsection is that all congestion pricing schemes that have been implemented or that are under study (Small and Gomez–Ibanez (1994)) are highly imperfect, and as a result additional policies to relieve traffic congestion are potentially welfare–improving. Thus, even if congestion pricing, in the policy sense, is in place, other policies are called for.

Perfect road pricing would entail variation of the toll so that at each location and at each point in time, each driver would face the marginal social cost of all his travel and driving decisions.

i) temporal variation

Ignore stochasticity for the moment, so that the temporal variation of traffic is fully predictable. Then, at a specific location, a perfect toll would vary continuously over time to reflect the temporal variation of traffic. Current technology permits arbitrarily fine temporal variation of congestion tolls. Such fine variation would, however, be annoying since it would make the informal calculation of when to travel complex. Thus, one expects that electronic road pricing will entail step tolls, with several steps. An interesting issue is what proportion of the potential efficiency gains from tolling can be had with step tolls. In a model of bottleneck congestion (Laih (1994)), a toll that is optimal conditional on having n steps over the rush hour achieves a fraction $\frac{n}{n+1}$ of the efficiency gains from a fully–flexible, time–dependent toll; for example, an optimized toll with three steps achieves three–quarters of the potential gains from tolling.

Step tolls provide a good example of how complementary traffic management policies can be beneficial when tolling is less than perfect. Step tolls tend to create congestion peaks after step falls. The freeway congestion caused by these peaks can be

smoothed by limiting the flow of cars that enter the freeways through "on–ramp metering."

ii) spatial variation

Just as continuously varying the congestion toll over time would be annoying, so too would be continuously varying the toll over space. Thus, one expects that electronic road pricing will entail a zone toll system. This entails some inefficiency. But there is a far more important source of efficiency loss associated with the spatial variation of tolls, which has already been alluded to.

Before sophisticated electronic tolling schemes à la Vickrey are introduced, it is likely that cities will start by tolling freeways and later freeways and major access roads to downtown. The problem with imposing congestion tolls on only some city streets is that drivers will tend to switch to those city streets without a toll — the tolling system will create traffic diversion. Similarly, imposing congestion tolls only on urban freeways will induce drivers to switch to city streets. In both cases, the implementation of a congestion toll may actually worsen congestion. There are two effects. The toll will reduce the amount of traffic, which by itself reduces the amount of congestion, but it will also divert traffic to untolled links, and if the diversion is to "more congestible" links, congestion may increase. This is demonstrated in Figure 1 below, which depicts the

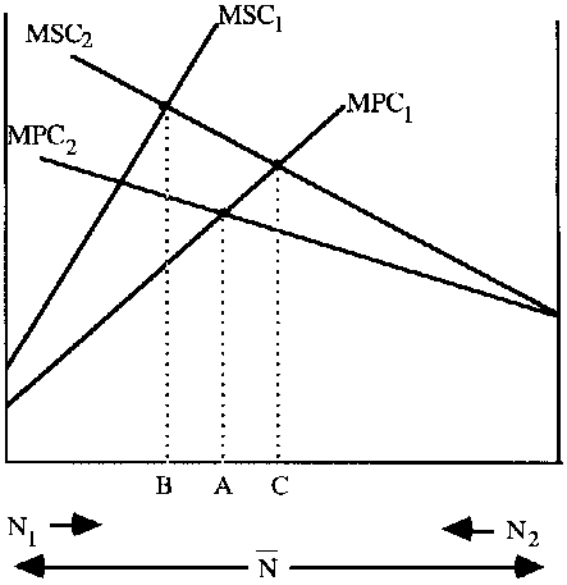

Figure 1: Tolling on only some streets can worsen congestion

extreme situation where demand is completely inelastic. There are $\overline{N}$ cars which travel between a given origin and destination along either of two routes, 1 and 2. In the absence of congestion tolling, the division of traffic between the two routes is shown by A, where the private costs are equalized. With perfect congestion tolling on both routes, the division of traffic is given by B, where the social costs are equalized. And with perfect congestion tolling on only the less congestible route — route 2 — the division of traffic is given by C. The important point is that the efficiency loss, relative to the first–best division of traffic, is larger when a perfect congestion toll is imposed on only the less congestible route that when there is no congestion tolling. Unfortunately, this situation is more than just a curiosum since it is likely that secondary urban streets are more congestible than major urban arterials which are in turn more congestible than urban freeways. Thus, it appears that a tolling scheme which covers only urban freeways, or only urban freeways and major arterials, may be highly inefficient.

Complementary policies are needed to discourage cars from switching from tolled to untolled roads, or from making detours around tolling stations. To discourage cars from detouring around freeway tolling stations, tolling stations could be installed on entry and exit ramps. And commuting on secondary streets can be discouraged by introducing a complex system of one–way streets and by blocking some streets — but this may give rise to more delay than it alleviates.

ii) stochastic variation

The efficiency loss due to stochastic variation in congestion depends on the magnitude of the stochasticity, the responsiveness of the congestion tolls to the stochasticity, and the responsiveness of drivers to both the stochasticity and the tolls. Consider a link with a time–varying toll that is not adjusted for stochastic variation in traffic density. When density is lower than average, the toll will be higher than optimal, and when density is higher than average, the toll will be lower than optimal. A completely responsive toll adjusts to the instantaneous realization of congestion. Vickrey has advocated responsive pricing in many contexts (Vickrey (1971)). An important point he has sometimes overlooked is that the effectiveness of responsive pricing is crucially dependent on the information that drivers have when making their travel decisions. Charging drivers for getting blocked behind an accident if they are unaware of either it or of the increase in the toll it induces is not only ineffective but also adds needless insult to injury. Thus, the effectiveness of a responsive tolling scheme in dealing with stochastic

variation depends on the quality and timeliness of information provided travelers about traffic conditions and the level of the toll.[3]

In recent years, there has been much discussion among transport engineers of vehicle information systems (Ref.?), and a number of experimental systems are being tested.(Ref.?) Congestion tolling will be more effective if it is combined with a vehicle information system which provides current information of toll levels and traffic conditions. Conversely, vehicle information systems will generate greater efficiency gains if they are combined with congestion pricing.[4]

iv) individual variation

Individuals vary considerably in their quality as drivers. To charge each for the "average" congestion caused by a driver is to undercharge bad drivers and to overcharge good drivers, which entails an efficiency loss relative to the first best. This inefficiency calls for complementary policies which penalize bad drivers. To some extent traffic ticketing serves this purpose. But traffic ticketing as practiced is largely ineffective, and covers only gross violations such as not stopping at stop signs, speeding, and running red lights. There are many other forms of bad driving that exacerbate congestion but are not ticketable. The hardware that will be used in electronic congestion pricing could perhaps be adapted to impose an instantaneous fine for committing some of these offenses — passing on the right, cutting in when the headway between cars is unsafely small, moving into an intersection so many seconds after the light has turned amber, not letting a car in when traffic is merging, accelerating rapidly, etc.

v) accidents and breakdowns

"U.S. studies in the Los Angeles conurbation show that more vehicle hours of delay results from extraordinarily and accidentally occurring traffic disturbances than from regularly occurring network overload during daily peak hours." (Busch, 1991).

---

[3]The proposed tolling scheme in Cambridge, England is responsive. Under the proposal, a driver will be charged on the basis of how long it takes her to travel between tolling stations. Unless drivers are informed of current traffic conditions, it would be preferable to charge them on the expected (ex ante) rather than realized (ex post) time.

[4]This theme is developed in Arnott, de Palma, and Lindsey (1991), Ben-Akiva, de Palma, and Kaysi (1991), and de Palma and Lindsey (1992). The basic idea is that, in the absence of congestion pricing, improved information will induce individuals to make travel decisions that are closer to privately optimal. But, with inefficient pricing, these need not be closer to socially optimal — better information with distorted prices can make drivers worse off. Electronic congestion pricing and vehicle information systems exhibit strong technological complementarities.

When we talk about congestion, we tend to think of regularly occurring network overload. But evidently traffic disturbances are at least as important contributors to congestion. Traffic disturbances are due to not only accidents and breakdowns but also to road repair, expansion, improvement, and maintenance. I do not know the proportion of delays due to accidents and breakdowns, but it could be calculated since the number of accidents and breakdowns are recorded, and traffic engineers have studied the average number of vehicle–hours of delay engendered by a two–lane accident, one–lane accident, lane–blocking breakdown, and a shoulder breakdown[5] — 4600, 2900, 1600, and 200, respectively (Goolsby (1971) and Grenzeback and Woodle (1992)). In any event, each accident and breakdown in rush–hour traffic causes a huge amount of delay. If accidents and breakdowns were random, it would be appropriate to lump them with congestion and price them via congestion pricing. But they are not random; the cautious, defensive driver who maintains a good car well will get into an accident and break down only once or twice in a lifetime. Accordingly, pricing vis–à–vis traffic accidents and breakdowns should entail insurance companies paying for the delay due to their clients' accidents and breakdowns and adjusting premia accordingly.

To sum up: One can expect that the forms of congestion pricing that will be employed over the next two decades will be quite crude. Even sophisticated forms of electronic road pricing fall a long way short of perfect congestion pricing.

1.3 Other distortions

Even if first–best road pricing were feasible, it would not be desirable because of other distortions which impact road travel. For the same reason, other traffic management policies would be potentially welfare–improving.

i) parking

Other than not perfectly road pricing, probably the largest distortion vis–à–vis urban auto travel is the mispricing of parking. Gillen (1977) found that, for commuters to downtown Toronto who had to park privately, the cost of parking was about the same as the sum of the costs of all other components of the auto commute. Thus, parking pricing has potentially a very substantial impact on urban auto travel. Perhaps the largest distortion with respect to parking is the provision of free, or heavily subsidized, parking

[5]The delay due to even a shoulder breakdown is remarkably large. Part of this delay is due to vehicles changing from the slow lane to avoid the broken–down car, but apparently the major cause is "rubber–necking."

by employers,[6] with no subsidy given to travel by mass transit.[7] A related distortion is the free provision of parking at suburban shopping malls. Meter parking is typically grossly underpriced during peak periods. Why more flexible temporal parking meter pricing is not employed is a puzzle since the technology is there. Responsive meter pricing which reflects the occupancy rate of parking meters is employed nowhere, though forty years ago Vickrey (1954) outlined and advocated such a scheme. A related distortion is the provision of free parking on city streets throughout much of the downtown area, especially in residential districts.

There are obvious interactions between road pricing and parking pricing. If roads are mispriced, parking pricing should be adjusted to compensate, and if parking is mispriced, road pricing should be adjusted to compensate.

ii) work start times

If urban travel is efficiently priced, the collectivity of firms together will tend to choose work start times that are excessively dispersed, since each firm will ignore the benefits that it confers on other firms from having similar working hours unpriced. If urban auto travel is underpriced, there is a countervailing effect (Henderson (1981)) since firms will ignore the congestion externality that its employees exert on the employees of other firms.

iii) transit

Transit is rarely priced at marginal cost. Deficit constraints and incentive considerations (Laffont and Tirole (1993)) push prices above marginal cost. Distributional considerations, meanwhile, in particular that mass transit is disproportionately used by the poor, push transit prices in the opposite direction. Unionization and patronage inflate costs, while grants–in–aid for capital equipment deflate them.

iv) auto insurance

---

[6]I do not have recent figures on the proportion of downtown employees who have subsidized, employer–provided parking. The proportion is lower in large cities and has been falling over time, but is still substantial.

[7]Employers presumably provide such parking free because it constitutes an untaxed fringe benefit, and do not subsidize travel by public transit since this would be taxable as income.

The marginal cost of auto insurance — that is, the increase in cost for traveling an extra mile — is probably substantially underpriced. An obvious reason is that auto insurance premia tend to be set independently of distance traveled. A more subtle reason, pointed out by Vickrey (1968), is that auto accident costs are like congestion — they increase with the density of traffic. Accordingly, efficiency requires that individuals pay marginal social accident costs, whereas with actuarially–fair insurance they pay only average accident costs. A practical way to deal with this distortion is to tax auto insurance.

v) other distortions

There are a host of other distortions which have some impact on urban transport — the incompleteness of insurance markets, inequities in income distribution, the underprovision of public infrastructure by local jurisdictions because of positive spillovers, and so on. It is difficult to know how wide to cast the net in considering non–transport distortions that should be taken into account in the design of transport policy.

To sum up: The title of this paper is somewhat misleading since it suggests that if pricing is employed, other traffic management policies are unnecessary or of secondary importance. This would be largely true with first–best pricing. But the crude forms of congestion pricing that will be employed in the next couple of decades, and even sophisticated forms of electronic road pricing, fall far short of perfect road congestion pricing. And even if perfect road pricing were employed, there would be a host of other transport–related distortions that should be considered in the design of transport policy. Thus, whatever form of congestion pricing is employed — and whether it is employed — a wide range of other transport management policies will be needed to alleviate traffic congestion efficiently.

2. Alternative Traffic Management Policies

There is a vast literature on urban transport policy. Rather than attempt a survey, I shall give a broad brush and idiosyncratic overview of literature, focusing on issues that I find particularly interesting or that I feel have been unjustifiably neglected.

I consider six categories of alternative or supplementary policies to road pricing as means to alleviate traffic congestion: i) Expansion or upgrading of existing roads, ii)

expansion or upgrading of mass transit, iii) regulation, iv) information, v) non–road transport pricing, and vi) changing driver behavior. To keep the paper's length down, I shall ignore land use and growth management policies. These are well–discussed in Downs (1992).

2.1 Expansion or upgrading of existing roads

The cost of expanding existing roads and freeways in the downtown of most U.S. cities is extremely expensive. The major components of the Central Artery Project in Boston are the construction of a new bridge and a new tunnel, and the replacement of an antiquated flyover system of freeways with an underground network of freeways. The estimated increase in capacity — in terms of the number of cars that can be delivered to the CBD per hour — is __ and the estimated cost is __, a figure that is sure to increase (Ref.?).

The capacity of existing roads and freeways can be increased through improvements in the road surface, street lighting, intersection design, traffic lighting, on–ramp metering, improved accident detection and management, grading, and ramp design. But such improvements together can only increase capacity by so much — perhaps by a factor of two — are expensive, and exacerbate congestion when being installed.

Suburban congestion has become a significant problem in recent years. Here there is scope for widening roads and the costs are not nearly as high as downtown. But major expansions are likely to generate considerable community opposition, as occurred with freeway construction in older, North American cities in the late sixties.

Small (1991) and Downs (1962), among others, have argued that, even if it were not prohibitively expensive, in the absence of congestion pricing building our way out of congestion is not the answer because any benefits from capacity expansion will be largely neutralized by latent demand. In what follows, I shall present and evaluate this argument.

Figure 2, adapted from Gronau (1994), illustrates the benefits from a transport improvement on an isolated road with and without congestion pricing. The improvement has no effect on free flow travel cost (OA) but makes the road (or road network) less congestible, lowering the private cost curve from PC^b to PC^a and the social cost curve from MSC^b to MSC^a. To simplify, linear demand and cost curves are assumed. The demand curve is $p = a - bq (a > 0, b \geq 0)$ and the private cost curve is $c = e + fq (e, f > 0)$.

Consider first the benefit from the improvement with congestion pricing in place. Before the improvement, social surplus equals ABC. Travel occurs to the point where $D(=MSB)=MSC^b$ so that social benefit is $OBCJ$ and social cost $OACJ$. After the improvement, social surplus equals ABE, so that the benefit from the transport improvement is ACE.

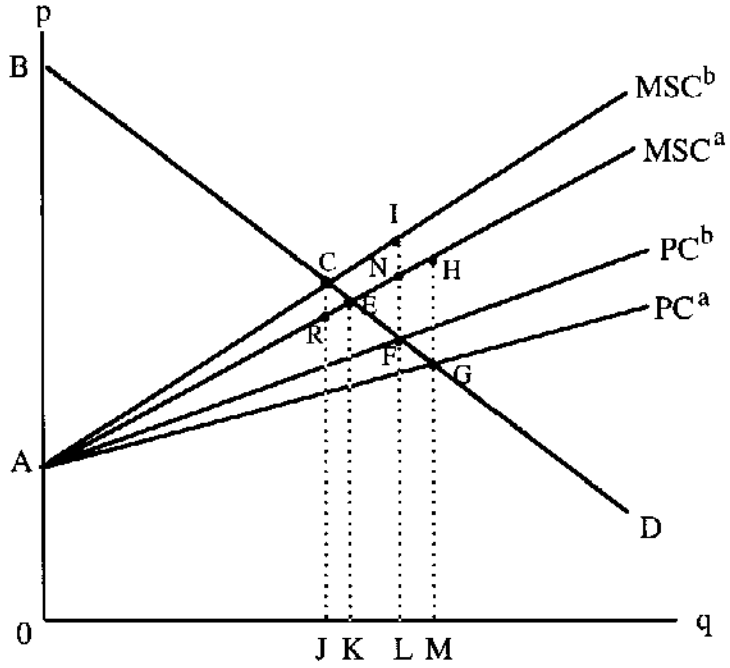

Figure 2: The benefits of a transport improvement with and without congestion pricing

Consider next the benefit from the improvement without congestion pricing. Before the improvement, social surplus is $ABC-CIF$. Travel occurs to the point where $D(=MSB)=PC^b$, so that social benefit is $OBFL$ and social cost $OAIL$. After the improvement, social surplus equals $ABE-EHG$, so that the benefit from the transport improvement is $ACE-(EHG-CIF)$.

There has been considerable discussion in the literature concerning the circumstances under which optimal capacity is lower when congestion is underpriced than when it is efficiently priced (Vickrey (1967), Mohring (1970), Wheaton (1978), Wilson (1983), d'Ouville and MacDonald (1990), and Gronau (1994)). In terms of the diagrams, the issue reduces to the circumstances under which the net benefit from a transport improvement is lower with underpriced congestion than with efficiently–priced congestion, *viz.* $EHG-CIF>0$. Simple algebra (see Appendix 1) gives that this inequality is equivalent to the inequality

$$-b^2 + bf + f^2 > 0.$$

Thus, for the case of linear demand and cost curves, optimal capacity is smaller with underpriced congestion when the elasticity of demand is sufficiently high (b sufficiently low) and the congestibility of the facility is sufficiently high (f sufficiently high).[8]

The intuition is as follows. The benefit from the transport improvement can be decomposed into the benefit with the number of trips fixed, B_1, and the change in the benefit due to the increase in trips due to the transport improvement, B_2 — the latent demand. Now $B_1^c = ACR$ and $B_1^{nc} = AIN$, where superscript c denotes with congestion pricing and superscript nc denotes without, so that $B_1^{nc} - B_1^c = RCIN > 0$. B_1 is larger in the absence of congestion pricing simply because more people travel on the road when congestion is underpriced. Also, $B_2^c = CER$ and $B_2^{nc} = -NHGF$. B_2 is small and positive with congestion pricing.[9] With unpriced congestion, it is negative since the marginal social cost of the extra trips induced by the transport improvement exceeds the marginal social benefit, while with congestion pricing, the marginal social cost of the extra trip is approximately equal to the marginal social benefit. Thus, latent demand at least partially neutralizes the benefit from a transport improvement when congestion is underpriced, but not when it is efficiently–priced.

An extreme illustration of this principle is provided by the Downs–Thompson–Vickrey Paradox. A fixed number of commuters, N, travel from A to B, which is connected by two roads, a direct one which is congestible and a circuitous one which is not. Now consider the effect of expanding the congestible road, on the assumption that some commuters travel on the uncongestible road both before and after the transport improvement. Commuters divide themselves between the two roads such that trip costs are equalized. Since the trip cost on the uncongestible road is unaffected by the improvement, so too are aggregate trip costs. The expansion of the congestible road generates zero benefits. The Paradox can be illustrated with Figure 2, with the diagram portraying the congestible road. Over the relevant range, where commuters travel on both roads, the demand for the congestible road is perfectly elastic at the trip cost on the uncongestible road, c. The increase in social benefit from travel on the congestible road is $c\Delta q$, where Δq is the increase in traffic on the road induced by the capacity expansion. The increase in users' costs is also $c\Delta q$ since private cost remains unchanged at c. Many

[8]The analysis is clearly considerably more complex when the demand and cost curves are non–linear.
[9]In fact, due to the Envelope Theorem, the ratio of CER to ACR goes to zero for infinitesimal capacity expansions.

real–world examples of the Paradox have been given. To them I add one more, provided by a group of students and assistents at the University of Munich: Hamburg has a relatively uncongested, circumferential highway which crosses the Elbe, as well as tunnels under the Elbe. Building more tunnels has not reduced travel time through them.

The central issue, to which I now turn, is whether the above arguments imply that, in the absence of congestion pricing, the benefits from capacity expansion and improvement will be largely neutralized by latent demand.

First, I shall turn the Downs–Thompson–Vickrey Paradox on its head. Consider an improvement of the uncongestible road, which reduces the trip cost from c^a to c^b. Perhaps some corners are taken out so that the road is shortened; perhaps the road surface, the grading, or the lane markings are improved so that cars can drive faster. In the absence of congestion pricing, the benefits from the improvement are $\left(c^b - c^a\right)N$; trip cost is reduced by $c^b - c^a$ not only for those who travel on the uncongestible road but also for those who travel on the congestible road. With congestion pricing, however, the benefits from the improvement are only $n\left(c^b - c^a\right)$, where n is the number of commuters on the congestible road. Thus, the Paradox illustrates the relative advantage of improving less congestible roads when congestion is unpriced, rather than, as has commonly been argued, the futility of road expansions when congestion is unpriced.

Second, Figure 2 can be interpreted as describing a group of identical arterial streets joining housing in the suburbs to work downtown during the rush hour. According to this interpretation, the relevant demand curve is that for commuting by car during the rush hour. Since this demand is likely highly inelastic, at least for the majority of U.S. cities with poorly–developed public transit, optimal capacity with unpriced congestion may in fact be higher than with efficiently–priced congestion.

To summarize: Relative to the situation where congestion is efficiently–priced, with underpriced congestion

i) optimal capacity is higher on less congestible roads, and lower on more congestible roads.

ii) capacity is higher in the aggregate when the aggregate demand for auto travel is highly inelastic, and lower when the aggregate demand is highly elastic.

On the basis of the above analysis, there would appear to be no presumption that the underpricing of congestion reduces the benefits from expanding capacity. We should therefore think twice before dismissing road expansion as a means of alleviating congestion when congestion is unpriced. It may be that road expansion is not a viable option in downtown areas because it is so expensive, but this applies whether or not congestion is underpriced. But it may also be that the benefits from expanding capacity downtown are extremely high, offsetting the extremely high costs. Road expansion in the suburbs merits serious consideration whatever the nature and level of congestion pricing.

2.2 Upgrading and expansion of mass transit

Auto travel in U.S. cities was heavily subsidized in the decades immediately following World War II. Gasoline was priced far below the world price and there was massive expenditure on highways and freeways which was financed out of general revenue. Expenditure on mass transit was insignificant by comparison. One result was a bias towards low–density urban development. Another was a faster transition from CBD–oriented cities with CBD–oriented arterials serviced by mass transit to a hierarchy of subcenters than would have occurred with efficient pricing. The combination of the two speeded up the transition to the "automobile city" in which mass transit is not cost–effective, at least in the newly–developed parts of the cities. Densities are too low to justify anything other than low trip frequency on the major C.B.D.–oriented arterials, and the increased spatial dispersion of workplaces and retail outlets makes line–haul mass transit more and more inconvenient.

Unfortunately, because of the durability of housing and urban infrastructure, especially roads, the spatial pattern of urban development is very slow to change. We are stuck with the automobile city, whether we like it or not. Residents of Los Angeles, Phoenix, Denver, etc. might wish that their downtowns had the vibrancy of Boston, New York, Paris, and London, and that they did not have to spend so much time in their cars in congested traffic. But wishing will not make it happen. For the same reason, I can envisage no scenario under which mass transit will make a major comeback. In most U.S. cities, public transit is and will remain the transport mode of the very poor who do not own a car.

Mass transit has, however, made a significant comeback in at least the inner areas of the older cities, that were constructed prior to the automobile era. This has come about due to a combination of increased road and parking congestion, and the revitalization of

downtown, at least in Boston and New York. In Boston, ridership has been steadily increasing over the last fifteen years. This has permitted improved rolling stock and more frequent service on an expanded set of routes. This process can be expected to continue.

What should public policy be vis–à–vis mass transit? In a first–best world, the answer is straightforward, with one qualification. Set prices equal to marginal costs and build capacity according to standard cost–benefit principles. The qualification is that mass transit is characterized by decreasing, long–run average costs, which may give rise to multiple local optima. With distortions, however, the answer is far from straightforward. To illustrate, suppose that there is only one distortion –– that congestion pricing is not applied to cars. Remarkably, second–best transit policy with underpriced auto congestion has not, to my knowledge, been analyzed in the literature. Here is not the appropriate place to analyze it. But let me make a couple of comments. The first is that with constant long–run costs to auto travel and decreasing long–run costs to mass transit, a representative individual, and perfect substitutability between auto trips and mass transit trips, the unrealistic solution is obtained that either all trips should be by car or all trips should be by mass transit. Thus, to obtain a sensible solution, imperfect substitutability between car and mass transit trips should be assumed, which rules out simple geometric analysis. The second is that the problem is intrinsically complex. In the previous section, we analyzed how the planner's choice of optimal road capacity is affected by the constraint that congestion is unpriced. That problem was complex enough. But now we have two imperfectly substitutable modes and three instruments –– road capacity, the transit fare, and transit capacity — that the planner can adjust to mitigate the distortion associated with auto congestion being unpriced or underpriced. Thus, it is fair to say that, given the current state of the theory, little can be said about optimal transit policy in the presence of unpriced auto congestion.

While I do not have an overarching model, I do have a few disconnected comments vis–à–vis mass transit and the alleviation of traffic congestion.

1. Discussions of urban transport tend to be framed in terms of auto versus mass transit. As a result, policies which combine the two modes often get overlooked. The two modes can be combined by providing parking at mass transit stops (park 'n ride). The desirability of expanding parking at specific transit stops should be considered.

2. Studies (see references in Giuliano and Small (1994)) have found that, in considering whether to switch to mass transit on commuting trips, drivers are more sensitive to the

quality and convenience aspects of mass transit travel — density of routes, frequency of service, comfort, and reliability of service — than to the price of mass transit. This stylized fact prompted three thoughts.

The first echoes a question posed by Vickrey(1979): What to do about those gregarious buses (and in Boston subway trains as well)? I commute by subway. Not infrequently, five subway trains go by in the opposite direction before one comes in my direction. To some extent this results because I reverse commute. But the main reason is bunching. Consider a situation where all subway trains are evenly spread on a circle. Stations are equally spaced too, and demand is the same at all stations, except for stochastic variability. Now suppose, by chance, that an unusually large number of passengers get on at a particular station. Boarding takes longer than normal and as a result the train is delayed. Because of this delay, a larger than average number of passengers are waiting for the subway at the next station, which causes further delay, and so on. Now consider the train behind. As a result of the train ahead being delayed, a smaller number of passengers are waiting for the train behind. Boarding takes less time than normal, and the train behind finds itself ahead of schedule and catching up the train ahead. The system is inherently unstable and results eventually in all the subway cars being bunched together. I have not worked out an answer analytically, but have at least a partial intuitive solution. Imagine the trains being attached by springs. If a train is delayed, the spring attaching it to the train ahead becomes tauter and slows that train down, while the spring attaching it to the train behind becomes looser and slows down that train too. The trains will lurch back and forth, but will not get bunched up. The real world analog to the effect of the spring is adjusting speed (and time stopped at stations) according to the headway between a train and the trains ahead and behind.

Another thought concerns the capital intensity of mass transit, buses in particular. It is generally acknowledged that there is a bias towards capital intensity in mass transit (Frankena (1987)) due to high union wages and grants–in–aid for capital but not labor expenses. This capital intensity obviously causes factor misallocation. Less obviously, it lowers the quality of mass transit by causing the frequency of service and density of routes to be reduced. Thus, policies to reduce its capital intensity would make mass transit more attractive.

A final thought concerns jitneys and taxis. Back twenty years ago, there was the expectation that urban decentralization would stimulate more flexible forms of public transit such as jitneys and mini–buses. As far as I know, all the experimental systems that

were tried have been discontinued. Waiting times were too long because passenger volume was too low. With the development of IVHS, these systems may make a comeback. In any event, a related policy is the expansion of taxi service. Mohring (1972) has pointed out that mass transit is characterized by increasing returns to scale. The same is true of taxi service since a doubling of passengers and cars results in decreased waiting time. This implies that, in the first best at least, taxi travel should be subsidized. The difficulty lies in developing a system of subsidization which cannot be abused.

2.3 Regulation

Regulation can be categorized according to whether it is designed to improve the efficiency of traffic flow, to decrease the amount of traffic, or to spread traffic over the course of the day.

i) regulation to improve the efficiency of traffic flow

Realistically, congestion pricing cannot be so finely tuned as to cause drivers to face marginal social cost on every margin of behavior. For example, it is absurd to think of a driver at an intersection weighing the benefit to him of advancing against the marginal social cost. For this reason, drivers' behavior is regulated — speed limits are imposed, stop signs and traffic lights are employed at intersections, and on–ramp metering is employed to regulate the merging of traffic entering a freeway. I have been impressed by the quality of traffic engineering. Innovations are not introduced without considerable experimentation and simulation. My criticisms of traffic engineering concern what it doesn't do rather than what it does.[10] First, it treats traffic flow as governed by physical laws, rather than as the outcome of individual maximizing decisions. Consequently, it has paid little attention to incentives to alter individual driver behavior. I shall have more to say about this later. Suffice it for the moment to point out an important margin of individual choice that traffic engineers apparently neglect — car, truck, and bus size. Second, traffic engineers appear to have given little attention to highway expansion, maintenance, and repair procedures. They study the congestion caused by these procedures, which is immense, and how to inform drivers of road works and to guide drivers around them. But they seem to take the highway procedures as given. Let me throw out a serious suggestion, but one that is bound to be contentious —

[10] An apparent exception to this is highway safety design standards, which appear to be based on minimal and dated experimental evidence (Hauser ( )).

all work on urban freeways should be done at night. I suggest, more generally, that highway maintenance and repair procedures should be evaluated by economists,[11] as should traffic accident clearance procedures.

Three other regulatory policies that have been widely used to improve the efficiency of traffic flow are car–pooling–only lanes, bus–and–taxi–only lanes, and HVO (high vehicle occupancy — which includes buses and carpoolers) –only lanes. Car pooling has become less popular over the past decade. I am not surprised. Car pooling among strangers would not work in the United States. And car pooling among acquaintances requires a highly inconvenient coordination of schedule. For these reasons, I anticipate that car pooling will decline in popularity and that many car–pooling–only lanes will be discontinued. Bus–and–taxi–only lanes are used extensively in Europe but are rare in the United States. If there were a large number of lanes of traffic, bus–and–taxi–only lanes would be very sensible. The number of lanes allocated to buses and taxis could be regulated to ensure that the lanes were well–utilized but at the same time were sufficiently less occupied than car lanes to ensure higher speeds and hence to induce switching from car to bus and taxi. As well, buses and cars do not mix well in traffic, so that segregating them improves the efficiency of traffic flow. Unfortunately, except for Manhattan, nowhere is the utilization of buses and taxis sufficiently high to justify the allocation of one lane of a two– or three–lane street to them.

One area of traffic regulation in downtown areas where substantial improvement is possible is trucking. It amazes me that large trucks are allowed to drive and park freely around downtown Boston at all hours of the day, blocking intersections and narrow streets. The cost of the congestion caused must be at least an order of magnitude larger than any direct savings in costs that derive from having larger trucks. In Paris, trucks are allowed to deliver in the Ville de Paris only during certain hours of the day and their size is regulated.

ii) regulation to decrease the amount of traffic

Congestion pricing is by far the most flexible and least cumbersome way to regulate the volume of traffic. But if congestion pricing is not applied or is crude — for example, is applied to urban freeways but not to city streets — then regulation merits consideration. In some European cities, regulation limits car travel downtown to certain days of the week according to license plate number. At first glance, this policy appears

[11]Newbery (1988) provides a seminal paper on this topic.

very clumsy and inefficient. But it would force auto commuters to experiment with mass transit options and would encourage work at home some days of the week. As well, if the regulation were to limit travel downtown only during rush hours, it would encourage employers to offer more flexible work schedules. Even though such regulation might be beneficial, I cannot contemplate Americans accepting it since the right to drive wherever and whenever one pleases is regarded as an essential element of personal liberty.

In recent years, there has been considerable discussion of tradable pollution rights. Pollution rights are allocated free and then trade in the pollution rights permitted. There is much to commend this resource allocation mechanism since it gives the government control over the total amount of pollution and gets each firm to face the shadow price of pollution, but without the government extracting revenue. Is such a resource allocation mechanism — tradable congestion rights — a viable option for traffic management? It would, in principle, allow the government to regulate the amount of congestion without the extraction of revenue, which, by itself, would increase the policy's popularity. Unfortunately, congestion is such an amorphous commodity that no market in congestion rights could be established. Thinking along these lines does, however, suggest a way that the government could make congestion pricing more palatable. It could grant everyone the right to, say, $2000 worth of congestion externality per year. Those who generated more than $2000 worth of congestion externality per year would pay the difference; those who generated less would receive a check from the government. This scheme might be practicable in countries with more centralized political systems, but it is hard to see how it could be made operational in the United States.

Boston has succeeded in limiting the amount of downtown traffic by regulation — by limiting the number of downtown parking spaces. This form of regulation is, however, highly inefficient since it entails rationing by queuing or, more precisely, rationing by cruising for parking. Not only is a huge amount of time wasted by those cruising for parking, but also cruisers–for–parking significantly impede the flow of downtown traffic. A local newspaper estimated that one–half of the cars traveling downtown Boston are cruising for parking. That figure seems to me high, but does suggest the order of magnitude of the inefficiency created by rationing parking by cruising for parking rather than by price.

iii) regulation to spread traffic over the course of a day

In recent years urban economists have changed the way in which they perceive rush–hour auto congestion because of a switch away from the naive flow model of congestion towards the bottleneck model of congestion, first developed by Vickrey (1969). According to the naive flow congestion model, the severity of congestion is related to the volume of traffic over the entire rush hour relative to capacity; and, since the demand for commuting trips is highly inelastic, the benefits from congestion tolling are predicted to be modest.

Let me describe the simplest bottleneck model of rush–hour traffic congestion with fixed demand, as presented in Arnott, de Palma, and Lindsey (1990). The assumption of fixed demand not only simplifies the exposition, but also puts into strong relief the difference between the bottleneck model and the naive flow congestion model, since, according to the flow congestion model, if demand is completely inelastic, congestion pricing is ineffective.

N commuters travel between a single origin (home) and a single destination (work) in the morning rush hour along a single route which has a bottleneck with flow capacity s. If the flow rate of cars arriving at the bottleneck exceeds s, a queue develops. In the absence of a queue, travel time from home to work is zero. All commuters have the same desired arrival time at work, t^*. The trip cost of a commuter is $C = \alpha$ (travel time) + β (time early) + γ (time late), where α is the shadow cost of travel time, β the shadow cost of time early, and γ the shadow cost of time late. The cost of early or late arrival at work is referred to as schedule delay cost. Each commuter decides when to leave home so as to minimize trip price which equals trip cost plus the toll.

Consider first equilibrium in the absence of a toll. The essential point is the trip cost must be the same for all commuters; otherwise, some commuters would have an incentive to change their departure times. This requires that travel time and hence queue length over the rush hour evolve such that trip cost is the same for all departures. Those who depart earliest and latest face no queue; their trip cost is entirely schedule delay cost. Those who arrive at work on time and hence incur no schedule delay cost must incur the highest travel time cost — by joining the morning rush hour queue when it is at its longest. Schedule delay cost, as a function of arrival time at work, declines linearly over the early–morning rush hour and then increases linearly over the late–morning rush hour. To satisfy the equal–trip–price condition, travel time cost, as a function of arrival time at work, must increase linearly over the early–morning rush hour and then decrease linearly

over the late–morning rush hour. Because the bottleneck is used to capacity over the rush hour, the arrival rate is uniform at s over the rush hour, and the total travel time costs across all commuters (TTC) equals total schedule delay costs (SDC). Thus, where superscript n denotes the no–toll equilibrium

$$TTC^n = SDC^n. \tag{i}$$

Now, consider setting a toll, as a function of arrival time, which equals travel time cost in the no–toll equilibrium (hence, the toll rises linearly over the early–morning rush hour and decreases linearly over the late–morning rush hour). If there is no queuing, each commuter faces exactly the same trip price as in the no–toll equilibrium. Since trip price was equalized over the rush hour in the no–toll equilibrium, so is it equalized with the toll and no queuing. Thus, application of this toll results in an equilibrium with no queuing. The distribution of arrival times and hence total schedule delay costs are the same as in the no–toll equilibrium, but total travel time costs equal zero. It is straightforward to show (Arnott et al. (1989)) that the equilibrium with this toll coincides with the social optimum. Thus, where superscript o denotes the social optimum:

$$TTC^o = 0 \qquad SDC^o = SDC^n. \tag{ii}$$

Total travel costs, TC, equal total travel time costs plus total schedule delay costs. Hence

$$TC^n = TTC^n + SDC^n = 2SDC^n = 2SDC^o = 2TC^o. \tag{iii}$$

Total travel costs are twice as high in the no–toll equilibrium as in the equilibrium with the optimal toll. Put alternatively: Even though the number of rush–hour commuters is fixed, application of the optimal toll reduces total travel costs to one–half their level in the absence of the toll. These results are illustrated in Figure 3.

The essential point is that, even when the demand for commuting trips is highly inelastic, very substantial efficiency gains can be achieved by tolling, via the rescheduling of commuter trips over the rush hour that the toll induces. In fact, the lion's share of the benefits from tolling may come about from the rescheduling of trips that a toll induces, rather than from a reduction in the total volume of trips. Since trip scheduling is ignored in the standard urban economic, flow model of congestion, the literature which uses that model (Arnott and MacKinnon (1978), Segal and Steinmeier (1980), Sullivan (1980), and Kraus (1986)) probably substantially underestimates the benefit from tolling.

The purpose of the above digression was to show that there are substantial efficiency gains to be had from reallocating traffic over the rush hour. In other models, such as Henderson (1981), there are also substantial efficiency gains from spreading out and lengthening the rush hour. The question at issue is whether, in the absence of tolling, these gains can be achieved via regulation.

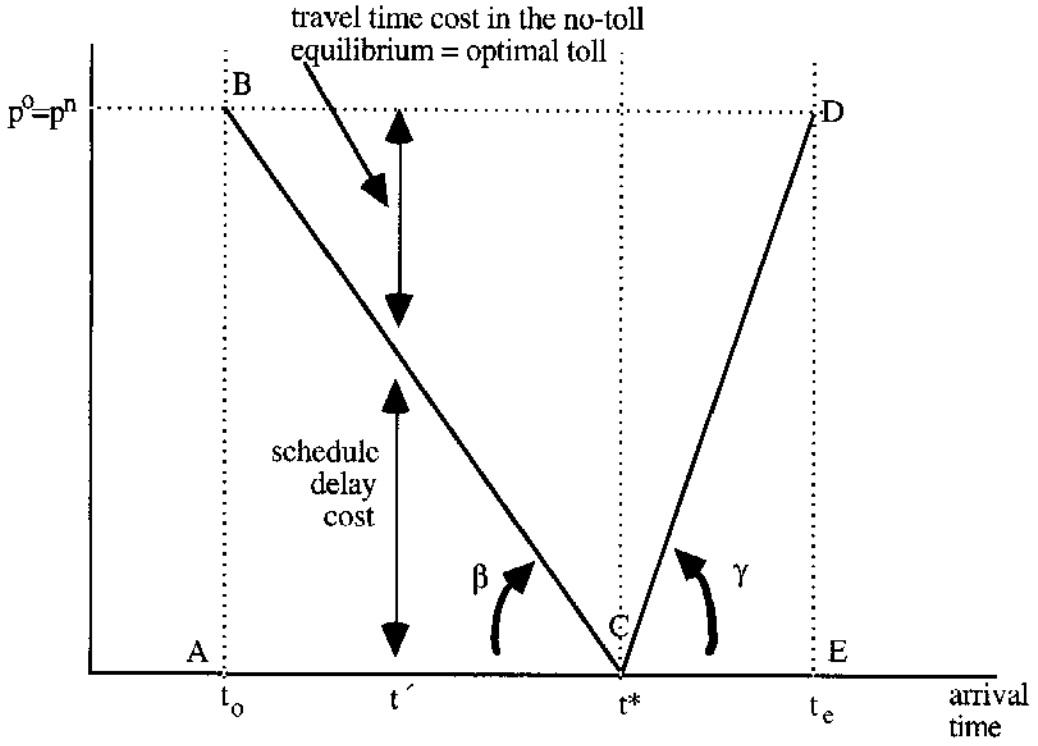

Figure 3: Equilibrium in the bottleneck with no toll and with the optimal toll
: t_o – start of rush hour, t_e – end of rush hour, p = trip price
: $TC^n = s(ABDE)$ $TC^o = s(ABC + CDE) = \frac{1}{2}TC^n$

In Washington, D.C. (Ref.?) and Ottawa (Bonsall (1978)), rush–hour congestion has been reduced considerably through the use of flex–time and staggered working hours. But this has been possible only because there is a single dominant employer — the federal government. Perhaps the regulation of the distribution of work start times is possible for other major employers. This would go part of the way towards solving the problem. But, according to the bottleneck model, it is the distribution of departure times and not arrival times that should be regulated, and this would appear infeasible. Thus, it would appear that there is a very important efficiency gain that can be achieved by time–varying tolls but not by any practical regulation.

2.4 Information

It was noted earlier that a substantial fraction of traffic congestion stems from traffic incidents. If drivers are informed of these incidents, they can change their routes to avoid them. Information on weather conditions and the general level of traffic can be used by drivers in deciding when to travel.

Improved information need not, however, reduce congestion. One example is where drivers are informed that travel time is lower along a more congestible route. Consider a sample driver who, upon receiving this information, decides to travel on the more congestible rather than the less congestible route. This will reduce his private cost, but will lower social welfare if the social cost on the more congestible route is higher. Another example occurs where information is provided but with a lag. Drivers are told, with a lag, that route A is less congested than route B. As a result, drivers switch from B to A, and route A becomes the more congested. Thus, it is important to design vehicle information systems to provide the kind of information that will lower congestion the most. It is also important to take into account the interaction between pricing and information. These issues are dealt with analytically by de Palma and Lindsey (1992) and El Sanhouri (1994), and via simulation by Mahmassani and Jayakrishna (1991). An important tentative finding by de Palma and Lindsey is that the benefits from information and congestion pricing are normally super–additive, i.e. the benefits from both together exceed the sum of the benefits from each without the other.

Despite a huge amount of research by traffic engineers on the design of IVHS (Intelligent Vehicle Highway Systems) systems, as well as a large number of pilot studies, the implementation of a practical and cost–effective system appears a long way off. Combined with congestion pricing, such systems hold the promise of providing substantial efficiency gains. But over the short and medium terms, traffic management policies should be designed without them.

A related policy is the provision of information on the availability of parking spaces at public parking lots and parking garages. This is common in European cities and works well. Electronic signs are posted on major access roads, indicating whether particular parking lots are full, or have a few or many parking spaces available. Since Marseille, which is probably the most chaotic city in France, can do it, at least some U.S. cities should be able to manage such a policy. Eventually, IVHS will allow drivers to reserve parking spots, paying from the time they make their reservation. A certain number of minutes from her destination, at the driver's choice, a driver will ask her in–car

computer for the nearest parking spaces in various price categories, choose the one she wants, and reserve it.

There is another potentially very important, though largely unrelated, interaction between urban traffic congestion and information -- electronic communication. There have been numerous articles in popular journals concerning the trend towards work at home, which is made possible by e-mail, the FAX machine, reduced long-distance telephone prices, and the electronic transmission of computer files. With the implementation of the much-heralded information superhighway and improvements in video communication, this trend will no doubt continue. Is this the wave of the future and will it lead to a substantial reduction in rush-hour congestion? The answer is far from obvious; recall that the telephone resulted in increased face-to-face interaction. Whatever the outcome, the issue is what appropriate government policy should be. First, because of adoption externalities, a case can be made that the government should subsidize the adoption of technological improvements in telecommunications. Second, with underpriced congestion, efficiency can be improved by subsidizing working at home. One way this can be achieved and which we would all like to see — I am not being entirely facetious -- is for the IRS to relax the criteria for the home office deduction. Another way would be for the government to permit firms to take deductions associated with their employees' work at home; such deductions could, however, be abused.

2.5 Transport-related pricing other than congestion pricing

Numerous forms of transport-related pricing, other than congestion pricing, are relevant to urban traffic congestion — mass transit and taxi pricing, parking pricing, the pricing of employee work start times, the pricing of accidents, the pricing of auto insurance, the pricing of cars and trucks, the personal income tax and corporate income tax treatment of travel-related expenses, and the pricing of gasoline.

i) mass transit and taxi pricing

Empirical studies suggest that the own-price elasticity of mass transit travel is low, as is the cross-price elasticity between mass transit and cars (Small (1992, p. 11)). Non-car owners are captive mass transit passengers. And in deciding on car versus mass transit, car owners are more sensitive to the quality aspects of mass transit, including service frequency, density, and reliability, than the mass transit fare. These observations

suggest that the direct effect of the mass transit fare on urban traffic congestion is of only secondary importance. However, to the extent that a higher fare translates into improved service, there may be an important indirect effect. As noted earlier, the determination of the optimal mass transit fare is a complex exercise in the theory of the second best.

It was also noted earlier that, in contrast to the current system, taxi travel should be subsidized, at least in the first best. Whether taxi subsidization is possible without serious abuse is open to question. For example, if a taxi's revenue were subsidized, taxi drivers would have an incentive to leave their meters running when idle.

ii) parking pricing

The distortions associated with existing parking practices were discussed in a previous section. With perfect congestion pricing, the optimal policies with respect to parking are relatively straightforward. First, employer–provided parking should be cashed out (Shoup (1982)), *viz.* the parking should be rationed by price. A complication, alluded to earlier, and to be discussed later, is the tax treatment of such parking. Second, on–street parking, too, should be cashed out. At parking meters, this entails preferably responsive pricing or, if that is not possible, time–dependent pricing. Competitive private parking would be priced efficiently, but account should be taken of the possibility that the owners of private parking facilities may have significant market power due to the friction of space. The amount of land allocated to on–street parking at various locations should be such that the shadow rent on land in on–street parking equals the shadow rent on land for street traffic. And the amount of land allocated to off–street parking should be such that the shadow rent on land in parking equals that in the best alternative use.

In the absence of congestion pricing on roads, or with only imperfect congestion pricing, the determination of (second–best) optimal parking fees is considerably more difficult since the parking fees should be designed not only to ration scarce parking spaces but also to restrict the amount of auto traffic. Likewise, the determination of the amount of public on– and off–street parking to provide, and of the amount of private parking to allow, will be that much more difficult.

Additional problems will arise if the government raises parking fees on only major streets in the downtown area. There may then be an increase in cross–traffic that traverses the downtown to park in adjacent areas where parking is significantly cheaper. There will also be an increase in parking in residential areas downtown. An ingenious

suggestion to deal with this problem is to install parking meters in higher–density residential neighborhoods, but give the resident(s) adjacent to a parking meter the revenue from the meter. That way residents are not made worse off by having free parking eliminated, but face a price for parking.

iii) pricing of employee work start times

Instead of regulating the distribution of work start–times of major firms, the government could impose a tax on employee work start times. Regulation in this situation seems preferable. For one thing, it would be difficult to determine the level of the tax. For another, firms would object more strongly if they had to pay a tax.

iv) pricing of accidents and breakdowns

It was argued earlier that insurance companies should pay for the delay caused by their clients' accidents and breakdowns. Passed on to clients, this should provide them with stronger incentives to avoid accidents, make sure that they have enough gas, and keep their cars in good running condition.

v) pricing of auto insurance

Most auto insurance rates (taxis excluded) are set independent of the miles driven. Why is a puzzle. Perhaps insurance companies figure that mileage would be difficult to monitor. But odometers could be made tamper–proof. And car–owners could be required to indicate current mileage at the annual vehicle registration. Perhaps drivers who drive more are safer, but one doubts that they are so much safer that the probabilities of accident and breakdown are independent of miles driven. The government could require that auto companies set their rates taking into account the number of miles a car is driven. In this way, drivers would come closer to facing the expected marginal accident and breakdown cost associated with their driving.

vi) the pricing of cars and trucks

Big cars cause more congestion than small cars. For this reason, it would seem appropriate to impose a tax on larger cars or, if revenue neutrality is a consideration, a tax on larger cars combined with a subsidy on smaller cars.

A tax on cars generally is probably inadvisable. First, such a tax is a crude way of dealing with congestion since it discriminates equally against city and non–city travel,

and for car owners does not change the price of travel. Second, such a tax would probably have adverse distributional consequences, since those who are on the margin between buying and not buying a car include many poor for whom a car is essential for access to a wider range of jobs.

Earlier it was argued that regulation should be imposed on the size of trucks permitted downtown and on the times trucks are permitted to make deliveries. Pricing would be preferable if full congestion pricing were in place because it would permit more flexibility. For example, trucks would then be able to make deliveries for which the benefit exceeded the social cost. As well, delivery van operators would have the appropriate incentives to choose the socially–optimal–sized truck. But in the absence of full congestion pricing, regulation seems appropriate.

vii) the tax treatment of travel expenses

Free employer–provided parking is treated as a non–taxable fringe benefit rather than taxable income–in–kind. If employer–provided parking were cashed out and employees were given a travel allowance equal to the rent on a parking space, the allowance would be taxable. Thus, the current tax system encourages subsidized employer–provided parking. This bias should be eliminated, by cashing out free parking, as I understand has been done in California. My suspicion is that employers would cash out free parking at substantially below the market rate if possible. To discourage this, large firms should be required to document the pricing of parking at the closest private parking garage. If a firm charges less than this, the difference should be treated as income in kind and employees be taxed on this income in kind.

The current corporate cum personal income tax systems are designed to be neutral with respect to business–related travel expenses. Perhaps they should not be. Perhaps the government should allow only a proportion of business–related travel miles to be deducted and adjust depreciation rules so as to encourage the use of smaller cars.

viii) the pricing of gasoline

Gas prices are much higher in Europe, and significantly higher in Canada, than in the United States. A gas tax is quite crude — though not as crude as a tax on car ownership — since it does not discriminate according to traffic conditions. Nevertheless, taking congestion and accidents into account, the shadow price of gasoline is probably significantly higher than the market price. Since recent attempts to raise significantly the

federal and Massachusetts gasoline taxes failed, it would appear, however, that a sharp rise in gasoline taxation is not politically feasible.

2.6 Changing driver behavior

Contrary to traffic engineers' models of traffic congestion that treat drivers as particles whose actions are governed by physical laws, drivers' actions are behavioral and can hence be influenced by carrots and sticks. Traffic flows much better in those cities where drivers are civil than in those where they are not.

Ideally, drivers should face the increased social cost associated with driving dangerously through increased insurance premia. Insurance companies should base their evaluation of a driver's risk on all the information available, including accident history (which they use) and traffic violations and miles driven (which they do not).

The optimal level of traffic fines is an interesting related issue. Fines seem much too low. A fine of $1000 for running a red light does not to me seem unreasonable. However, most people don't have $1000 on hand; perhaps people should be allowed to pay fines via installment and even via the tax system. Another problem is that traffic fines seem largely capricious. I walk by a stop sign trap every day. It is on the principal through street in a suburban neighborhood and is barely visible in one direction. The stop sign generates a considerable amount of revenue but running it says practically nothing about one's safety as a driver. Meanwhile, just up the road is a dangerous intersection that is never patrolled. Police should be given the incentive to go after genuinely dangerous driving rather than to give out their required number of tickets with the least bother. But how?

The young and elderly are, as groups, particularly bad drivers. They have the highest accident rates. And, on average, they also perceive themselves to be much better drivers than they are. Traffic laws should be stricter for both groups. Getting a license should be harder (the driving test is much harder in most European countries). Licenses should be suspended on the first violation for the young. And the aged should be required to retake the driving test every two years.

There is dangerous driving that breaks the law, dangerous driving that does not break the law (passing on the inside lane, rapid lane changes, barging in when the headway in minimal, tailgating, driving at the speed limit in the fast lane, etc.), and uncivil, antisocial driving (refusing to let a car merge or letting too many cars merge,

honking at the car in front before the light has turned green, not giving pedestrians the right of way when they have it, splashing pedestrians on rainy days) that not only impedes the flow of traffic but also makes driving in congested traffic (and being a pedestrian) so much more unpleasant. How can the latter two forms of driving behavior be discouraged? One scheme I have though about half–seriously, which will be feasible under IVHS, is ostracism. Each driver and pedestrian would have the right to report up to say ten drivers a month for bad behavior. If any driver is reported more than so many times during a one–month period, his license is temporarily suspended. I suspect that such a benign form of vigilante justice would be very popular. But how many politicians would not have their licenses suspended?

3. Overview

Let me start with an obvious point, but one that tends to get overlooked, even by economists and even by myself on occasion. The costs of congestion go hand–in–hand with the benefits of travel; zero congestion could be achieved if there were no travel. The optimal level of congestion has two characteristics. First, for a given level of benefits from travel, the costs of congestion are minimized. Second, when this efficiency condition is satisfied, the optimal level of congestion occurs when the benefits from increased travel are offset by the increase in travel costs induced by the increased travel. The optimal level of congestion could well be very high. Thus, while the paper title refers to "alleviating" traffic congestion, it should instead refer to achieving the optimal amount of travel at minimum congestion cost.

It is remarkable that economists have such a well–articulated theory of congestion, but such a poorly–integrated body of theory related to the benefits of travel. There is one branch of the latter theory that is quite well–developed –– that of firm–household interaction (Fujita and Ogawa (1982)). Households commute to work and locate so as to maximize utility, trading off wages, rents, and transport costs. Competitive firms operating under constant returns to scale locate so as to maximize profits, trading off wages, rents, and the unpriced productivity benefits from being close to other firms. The spatial dispersion of economic activity occurs because of the unpriced benefits of firm–firm interaction. The theory does provide a general equilibrium explanation of the location of economic activity but begs a couple of critical questions: What is the nature of the productivity benefits to a firm from proximity to other firms,

and why are they unpriced? Another branch of theory, stemming from the Hotelling model (Hotelling (1929), Beckmann and Thisse (1986)), takes household location as given and has price–setting firms trading off market size versus localized market power. Yet another branch of theory takes firm location as given, and examines an individual's cost–minimizing or utility–maximizing trip patterns — the traveling salesman problem, the knapsack problem, etc. No models that I know of address what seems to me an essential aspect of travel benefit — schedule coordination for workplace and household interaction. To enjoy the benefits from interaction, individuals must be at the same place at the same time.

The purpose of this digression is to make the point that, since we do not have a well–integrated theory of the benefits from travel, we should have little confidence that the marginal social benefit from a trip coincides with the marginal private benefit. Thus, in advocating congestion pricing, we are advocating first–best pricing rules when second–best pricing rules may be appropriate. Nevertheless, since we have little idea what the distortions are, it seems appropriate to retain first–best congestion pricing as the ideal.

In this paper, I have argued that optimal transport management entails a lot more than "congestion pricing." For political reasons it is unlikely that even crude forms of congestion pricing will be introduced in the United States for many years. Even when they are, they will fall so far short of the ideal that a wide range of complementary policies will be justified. Even sophisticated electronic tolling, of the form advocated by Vickrey (1963), which in my opinion will not be adopted in the United States for decades, falls a considerable way short of ideal congestion pricing. And even if perfect congestion pricing were applied, other traffic management policies would be needed because of other distortions. In short, we operate and will continue to operate squarely in the world of the second best. The world of the second best is messy enough even when there is only a single margin of choice that is distorted, since typically it is second–best optimal to introduce a battery of offsetting distortions. But in travel behavior, there are many margins of choice, which makes the intelligent design of traffic management policies exceedingly difficult and messy.

I have no panacea for traffic congestion; indeed, I think there is none. But there are so many distortions vis–à–vis travel behavior which are so large that there is considerable scope for improvement. Traffic congestion will get worse. But as it does, the political opposition to ameliorative changes will diminish. As well, policy

innovations will occur and those policy innovations that are successful will be widely adopted. Our politico–economic system is adaptive and we shall muddle through.

In concluding, rather than attempting to summarize my argument, I shall highlight three points.

The first is that there are considerable potential gains from expanding the scope of transport policy considered by economists. Traditionally, urban economic transport policy treats the pricing of car and mass transit travel and the choice of road and transit capacity. But there are many other margins on which public policy can be applied to reduce congestion — parking policy, policies to encourage smaller cars, trucks, and buses, policies to encourage civil driving behavior, policies with respect to car insurance, policies to reduce the disruption caused by traffic accidents and road work, tax policies, and policies to spread work start times, to name only a few. The application of basic economic principles to these new areas of transport policy holds considerable promise.

The second point is that we should not dismiss expansion of road capacity so quickly. The costs may be enormous, but so too may be the benefits. That auto congestion is underpriced reduces the benefits from expanding the capacity of more congestible roads but increases the benefits from expanding the capacity of less congestible roads.

The final point is that policies should be designed to exploit an important insight from the bottleneck model — that potentially very large efficiency gains are to be had from the rescheduling of departures over the rush hour.

Appendix 1

The marginal (net) benefit from capacity expansion, with linear demand and cost curves, and with congestion pricing.

The demand curve is $p = a - bq (a > 0, b \geq 0)$, and the private cost curve is $c = e + fq (e, f > 0)$, so that the marginal social cost curve is $MSC = e + 2fq$. A capacity expansion is modeled as a fall in f. Social benefit is

$$B = \int_0^q (a - bq')dq' = aq - \frac{bq^2}{2} \tag{i}$$

and total user cost is

$$C = eq + fq^2. \tag{ii}$$

With congestion tolling, $p = MSC$, so that

$$q^c = \frac{a-e}{b+2f}, \tag{iii}$$

and the marginal (net) benefit from capacity expansion is

$$\left.-\frac{d(B-C)}{df}\right|_{q^c} = \left.-\frac{d\left((a-e)q - \left(\frac{b}{2}+f\right)q^2\right)}{df}\right|_{q^c}$$

$$= -\left(-q^2 + (a-e)\frac{dq}{df} - 2\left(\frac{b}{2}+f\right)q\frac{dq}{df}\right)_{q^c}$$

$$= \left(\frac{a-e}{b+2f}\right)^2. \tag{iv}$$

Without congestion tolling, $p = c$, so that

$$q^{nc} = \frac{a-e}{b+f} \tag{v}$$

and the marginal (net) benefit from capacity expansion is

$$\left.\frac{d(B-C)}{df}\right|_{q^{nc}} = \left(\frac{a-e}{b+f}\right)^2\left(\frac{b}{b+f}\right). \qquad \text{(vi)}$$

Note that this equals zero if $b=0$.

Thus, the marginal (net) benefit from capacity expansion is greater with congestion pricing than without it if

$$-b^2 + bf + f^2 > 0, \qquad \text{(vii)}$$

and less with congestion pricing if the inequality is reversed.

BIBLIOGRAPHY

Arnott, R., and J. MacKinnon, 1978, "Market and Shadow Land Rents with Congestion," American Economic Review 68, 588–600.

Arnott, R., A. de Palma, and R. Lindsey, 1990, "Economics of a Bottleneck," Journal of Urban Economics 27, 111–30.

Arnott, R., A. de Palma, and R. Lindsey, 1991, "Does Providing Information to Drivers Reduce Traffic Congestion?" Transportation Research 25A, 309–18.

Bayliss, D., 1992, "British Views on Road Pricing," paper presented at the World Conference on Transportation Research, Lyon, France.

Beckmann, M. and J.–F. Thisse, 1986, "The Location of Production Activities," in P. Nijkamp, ed., Handbook of Regional and Urban Economics, vol. 1. Amsterdam: North–Holland, 21–95.

Ben–Akiva, M., A. de Palma, and J. Kaysi, 1991, "Dynamic Network Models and Driver Information Systems," Transportation Research 25A, 251–66.

Bonsall, T., 1978, "Flexible Work Hours and Public Transit in Ottawa," presented at the annual conference of the Roads and Transportation Association of Canada. Toronto, Ontario, September 23–4.

Borins, S.F., 1986, "The Political Economy of Road Pricing: The Case of Hong Kong," Proceedings of the World Conference on Transport Research, Vancouver, B.C., Vol. 2, 1367–78.

Boyer, M. and G. Dionne, 1987, "The Economics of Road Safety," Transportation Research 21 B, 413–431.

Busch, 1991, " ," in M. Papageorgiou, ed., Concise Encyclopedia of Traffic and Transportation Systems, , Pergamon.

Catling, I., and B. Harbord, 1985, "Electronic Road Pricing in Hong Kong: The Technology," Traffic Engineering and Control 26, 608–15.

d'Ouville, E. and J. MacDonald, 1990, "Optimal Road Capacity with a Suboptimal Congestion Toll," Journal of Urban Economics 28, 34–49.

de Palma, A. and R. Lindsey, 1992, "The Potential Benefits of a Combined Route Guidance and Road Pricing System: An Economic Analysis," CEM, Université Libre de Bruxelles, discussion paper 9217.

Downs, A., 1962, "The Law of Peak–Hour Expressway Congestion," Traffic Quarterly 16, 393–409.

Downs, A., 1992, Stuck in Traffic. Washington, D.C.: The Brookings Institution.

Dupuit, J., 1844, "On the Measurement of the Utility of Public Works." Reprinted in Transport, ed. Dennis Munby. London: Penguin, 1968, 19–57.

El Sanhouri, I., 1994, Evaluating the Joint Implementation of Congestion Pricing and Driver Information Systems, Ph.D. thesis, Dept. of Civil Engineering, M.I.T.

Frankena, M., 1987, "Capital–Biased Subsidies, Bureaucratic Monitoring, and Bus Scrapping," Journal of Urban Economics 21, 180–93.

Fujita, M., and H, Ogawa, 1982, "Multiple Equilibria and Structural Transition of Non–Monocentric Urban Configurations," Regional Science and Urban Economics 18, 161–96.

Gillen, D., 1977, "Estimation and Specification of the Effects of Parking Costs on Urban Transport Mode Choice," Journal of Urban Economics 4, 186–199.

Giuliano, G. and K. Small, 1994, "Alternative Strategies for Coping with Traffic Congestion," mimeo.

Goolsby, M., 1971, "Influence of Incidents on Freeway Quality of Service," Highway Research Board 349, 41–46.

Grenzeback, L. and C. Woodle, 1992, "The True Costs of Highway Congestion," ATE Journal, March, 16–20.

Gronau, R., 1994, "Optimal Capacity with a Suboptimal Congestion Toll," Journal of Urban Economics 36, 1–7.

Hau, T., 1992, Congestion Charging Mechanisms for Roads. World Bank Working Paper No. WPS–1071, Washington, D.C.

Hauser ?

Henderson, J.V., 1981, "The Economics of Staggered Work Hours," Journal of Urban Economics 9, 349–64.

Hotelling, H., 1929, "Stability in Competition," Economic Journal 39, 41–57.

Jones, P., 1991, "Gaining Public Support for Road Pricing through a Package Approach," Traffic Engineering and Control 32, 194–6.

Knight, F., 1924, "Some Fallacies in the Interpretation of Social Cost," Quarterly Journal of Economics, 38, 582–606.

Kraus, M., 1989, "The Welfare Gains from Pricing Road Congestion Using Automatic Vehicle Identification and On–Vehicle Meters," Journal of Urban Economics 25, 261–81.

Laffont, J.–J., and J. Tirole, 1993, A Theory of Incentives in Procurement and Regulation. Cambridge, Ma: M.I.T. Press.

Laih, C.–H., 1994, "Queuing at a Bottleneck with Single and Multi–step Tolls," Transportation Research 28A, 197–208.

Mahmassani, H. and R. Jayakrishnan, 1991, "System Performance and User Response under Real–Time Information in a Congested Traffic Corridor," Transportation Research 25A, 293–308.

May, A., 1993, "Potential of Next–Generation Technology," Transportation Research Board Study of Urban Transportation Congestion Pricing.

Mohring, H., 1970, "The Peak Load Problem with Increasing Returns and Pricing Constraints," American Economic Review 60, 693–705.

Mohring, H., 1972, "Optimization and Scale Economies in Urban Bus Transportation," American Economic Review 62, 591–604.

Newbery, D.M.G., 1988, "Road Damage Externalities and Road User Charges," Econometrica, 56, 295–316.

Pigou, A.C., 1912, Wealth and Welfare. London: MacMillan.

Ramjerdi, F., 1994, "The Norwegian Experience with Electronic Toll Rings," in Proceedings of the International Conference on Advanced Technologies in Transportation and Traffic Management," Centre for Transportation Studies, Nanyang Technological University, Singapore, 135–42.

Segal, D., and T. Steinmeier, 1980, "The Incidence of Congestion and Congestion Tolls," Journal of Urban Economics 7, 42–62.

Shoup, D., 1982, "Cashing Out Free Parking," Transportation Quarterly 36, 351–64.

Small, K., 1991, Urban Transportation Economics, vol. 51 in J. Lesourne and H. Sonnenschein, eds., Fundamentals of Pure and Applied Economics. Chur, Switzerland: Harwood.

Small, K., 1992, "Using the Revenues from Congestion Pricing," Transportation 19, 359–81.

Small, K. and J. Gomez–Ibanez, 1994, "Road Pricing for Congestion Management: The Transition from Theory to Policy," mimeo.

Sullivan, A., 1983, "Second–best Policies for Congestion Externalities," Journal of Urban Economics 14, 105–23.

Vickrey, W.S., 1954, "The Economizing of Curb Parking Space," Traffic Engineering Magazine, Nov. Reprinted in Journal of Urban Economics 36, (1994), 56–65.

Vickrey, W.S., 1959, "Statement on the Pricing of Urban Street Use." Hearings: U.S. Congress, Joint Committee on Metropolitan Washington Problems, 11 Nov., 466–77 plus attachments.

Vickrey, W.S., 1963, "Pricing in Urban and Suburban Transport," American Economic Review 53, 452–65.

Vickrey, W.S., 1967, "Optimization of Traffic and Facilities," Journal of Transport Economics and Policy 1, 123–136.

Vickrey, W.S., 1968, "Automobile Accidents, Tort Law, Externalities, and Insurance: An Economist's Critique," Safety: Law and Contemporary Problems 33(3), 464–87.

Vickrey, W.S., 1969, "Congestion Theory and Transport Investment," American Economic Review 59, 251–61.

Vickrey, W.S., 1971, "Responsive Pricing of Public Utility Services," Bell Journal of Economics and Management Science, 2, 337–46.

Vickrey, W.S., 1979, "Spacing Out Those Gregarious Buses," New York Times, June

Wheaton, W., 1978, "Price–Induced Distortions in Urban Highway Investment," Bell Journal of Economics 9, 622–32.

Wilson, J., 1983, "Optimal Road Capacity in the Presence of Unpriced Congestion," Journal of Urban Economics 13, 337–57.