THE TWO-MODE PROBLEM:
SECOND-BEST PRICING AND CAPACITY

Richard Arnott*†
and
An Yan*

August 2000

*Department of Economics
Boston College
Chestnut Hill, MA 02467, USA
e-mail: richard.arnott@bc.edu

Tel: 617-552-3674
Fax: 617-552-2308

†Corresponding author

Abstract[*]

Suppose that there are two congestible modes of travel from A to B — road and rail for concreteness – which are imperfect substitutes in demand. Road congestion from A to B is underpriced; this is an unalterable distortion. Compared to the first best, should the transportation planner choose a wider or narrower road, raise or lower the rail fare, and expand or contract rail capacity? This paper provides a synthetic review of the literature on the problem, presents some new results, and discusses directions for future research on this and related second-best problems.

Key words: second-best, pricing, investment, road, rail.

JEL Codes: R41, R42, D61

[*] We would like to thank the following for helpful comments: seminar participants at Berkeley, Cornell, McMaster, Tsukuba, Washington, Lausanne, and Lyon Universities; session participants at the 1999 AREUEA/ASRES Conference and the spring 2000 NBER environmental economics workshop, especially the discussant Ken Small; and the editor and two referees. We are also grateful to Yoshitsugu Kanemoto and Masayuki Doi for giving us copies of their work on components of the problem. Arnott wrote part of the first draft of the paper during a visit to the Department of Economics at the University of Tokyo, and part of the second while visiting the Fisher Center at the Haas School of Business, Berkeley. He would like to thank those institutions for their hospitality.

The Two-mode Problem: Second-best Pricing and Capacity

I. Introduction

This paper returns to a classic problem in urban transportation economics: Suppose there are two congestible modes of travel from A to B — road and rail for concreteness — which are imperfect substitutes in demand. There is a single distortion in the economy: congestion on the road is unalterably underpriced. How does the underpricing of road travel affect the optimal road width, the optimal rail capacity, and the optimal rail fare?

While considerable progress has been made in understanding the economics of the problem, we are still a long way from being able to provide precise quantitative answers to this question. This paper provides a synthetic review of the literature, explains why the problem has been such a tough nut to crack, presents a couple of new perspectives on the problem, and discusses promising directions for future research on this and related second-best problems.

Though it has not provided a complete analysis, the existing literature has examined a number of facets of the problem. First, Wheaton (1978), Wilson (1983) and d'Ouville and McDonald (1990) have addressed the question: If there is a single road connecting A and B on which congestion is unalterably underpriced, should the road be wider or narrower than when congestion is efficiently priced? Put alternatively: How does the second-best road width compare to the first-best? Wheaton found that reducing the toll infinitesimally below the first-best level results in an increase in optimal road width. Wilson provided a general local analysis, indicating how an infinitesimal change in the toll from any suboptimal level alters optimal width. He also derived global conditions, which he argued are likely to be satisfied empirically, under which second-best road width exceeds the first-best level. d'Ouville and McDonald (1990) provided an explanation of Wilson's main result and generalized it somewhat. Second, a number of researchers, including Lévy-Lambert (1968), Marchand (1968), Mohring (1979), Verhoef, Nijkamp and Reitfeld (1996) and Liu and McDonald (1998, 1999), have considered the short-run, two-mode problem for the case in which the two modes are perfect substitutes in demand[1]: Holding capacities fixed, is the second-best rail fare higher or lower than the first-best? The basic result illustrates a general principle in the theory of the second best, that an offsetting distortion is desirable or that two small Harberger triangles are better than one large one: The rail fare should be lowered from its first-best level to the point where the marginal decrease in deadweight loss from an inefficient modal split equals the marginal increase in deadweight loss from excessive travel. Sherman (1971), Bertrand (1977), and Doi (1986) extended the analysis to the situation where the two modes are imperfect substitutes or even complements in demand. The same general principle applies, but now with the optimal degree of underpricing of rail being positively related to the substitutability in demand between road and rail. Finally, Kanemoto (1996, 1999) has written a pedagogical paper outlining the principles underlying the determination of second-best optimal capacity for two special cases of the two-mode problem. In the first, (using our terminology) road travel is congestible and underpriced, while rail travel is uncongestible with trip cost depending only on capacity; how does the underpricing of road travel affect optimal rail capacity? In the second, under the same assumptions on congestion, the first-best toll is charged on the road, but rail travel is underpriced; how does the underpricing of rail travel affect optimal rail capacity?

While we shall continue to refer to the two modes from A to B as road and rail, it should be clear that the pair of modes has various interpretations. In the context of inter-city travel, the pair of modes can be any two of freeway, highway, air, and rail travel; and in the context of intra-city travel, the pair of modes can be any two of city street, urban freeway, LRT, subway, and commuter train. The analysis does not, however, apply to situations in which there is congestion interaction between the two modes, such as bus and car travel on the same city streets, though modification of the analysis to treat this class of situations would not be conceptually difficult (see Sherman (1971) for an analysis of the corresponding second-best pricing problem, capacities

[1] Some papers adopted a partial equilibrium approach, others a general equilibrium approach. The results are similar for the two approaches, apart from the presence of income effects in the general equilibrium model, and the insights gained are the same. Braid (1996) also examined these issues but in the context of Vickrey's (1969) bottleneck model, and hence focused on the time variation of the second-best fare over the rush hour. Our paper focuses on a

fixed). And as we shall comment on subsequently, our interpretation of the second mode as rail is inexact since the description of the rail technology is over-simplified. In some contexts, the relevant policy issues concern the interaction between more than two modes; for example, Pickerell (1992) has undertaken retrospective cost-benefit analysis of several light rail transit (LRT) systems in U.S. cities, accounting for both bus and car travel as alternative modes. The extension of our analysis to such situations is conceptually straightforward but algebraically non-trivial (see Bertrand (1977) for the corresponding second-best pricing problem, capacities fixed).

Our analysis follows most of the previous literature in employing partial equilibrium analysis. Doing so simplifies the analysis and aids intuition by permitting the application of consumer surplus analysis. Also, in common with all the previous literature, we ignore other distortions in the rest of the economy.[2]

We shall proceed in steps. In section II, we shall describe the general model, introduce notations, and present the first-best solution as a benchmark. In section III, we analyze first the Wheaton-Wilson problem and then the Doi problem. Then in section IV, we indicate the nature of the general problem and discuss directions for future research. Section V concludes.

II. The Model

INSERT FIGURE 1

Figure 1: Diagrammatic depiction of the problem

There are two modes of travel between a single origin A and a single destination B. Mode 1 is interpreted as road or car travel, mode 2 as rail (a generic term for any form of mass transit that

static problem, which can be interpreted either as ignoring time variation in congestion or as treating a reduced-form representation of the dynamic problem (see Arnott, de Palma, and Lindsey (1993)).

[2] It should not be surprising that interaction between distortions are potentially important. This has recently been highlighted in the literature on the double-dividend hypothesis. The earlier partial equilibrium literature on the Pigouvian taxation of pollution argued that such taxation has a double dividend; not only does it internalize the pollution externality but it also raises revenue which permits a reduction in income tax rates, leading to a fall in the labor-leisure distortion. Subsequent general equilibrium analysis (e.g. Bovenberg and Goulder (1998), and Parry and

has an exclusive right of way), though as noted earlier the pair of modes has other interpretations as well.

The demand for travel on mode 1 is given by $m(p_1, p_2)$ and on mode 2 by $n(p_1, p_2)$, where p_i is the full price of travel on mode i. This specification allows the two modes to be imperfect substitutes or even complements, places no restriction on the aggregate demand for travel, and abstracts from income effects. Individuals are assumed identical, so that the analysis can be cast in terms of a representative traveler.[3] The full price of travel on mode i equals the user cost on that mode, c_i, which includes both the money and time costs of travel, plus the toll on that mode, τ_i. Thus,

$$p_i = c_i + \tau_i, \qquad \mathrm{i} = 1,2. \tag{1}$$

Both modes are subject to congestion, and there is no congestion interaction between modes. To simplify somewhat, it is assumed, as is normally done in this literature, that the user cost for mode i depends on the volume-capacity ratio ("width" and "capacity" will be used interchangeably) for that mode:[4]

Bento (2000)) has shown the argument to be fallacious since it neglects that the Pigouvian tax raises the price of consumer goods, lowering the real wage.

[3] The representative traveler's utility maximization problem can be written as

$$\max_{c,m,n} \; U(m,n) + c \qquad s.t. \quad y = c + p_1 m + p_2 n,$$

where $U(\cdot)$ is the strictly concave utility function, c other goods and y income. On the realistic assumption that the individual never chooses to spend all his income on m and n, the demand functions $m(p_1, p_2)$ and $n(p_1, p_2)$ are independent of income, so that the consumer surplus function is $B(p_1, p_2) \equiv U(m(p_1, p_2), n(p_1, p_2)) + y - p_1 m(p_1, p_2) - p_2 n(p_1, p_2)$. Comparative static analysis yields: $\partial m/\partial p_1 < 0$; $\partial n/\partial p_2 < 0$; $\mathrm{sign}(\partial m/\partial p_2) = -\mathrm{sign}(\partial^2 U/\partial n \partial m)$ is positive if m and n are substitutes and negative if they are complements; and $(\partial m/\partial p_1)(\partial n/\partial p_2) - (\partial m/\partial p_2)(\partial n/\partial p_1) > 0$ by concavity of the utility function.

There are two rationales for ignoring income effects. The first is to isolate substitution effects. The second derives from the Slutsky relation in elasticity form:

$$E_{X:p} = E^c_{X:p} - \left(\frac{pX}{Y}\right) E_{X:Y},$$

where $E_{a:b}$ denotes the elasticity of a with respect to b, X a good, p the price of that good, E^c a compensated elasticity and Y income. The income effect is of second-order importance if the product of the expenditure share on the good (pX/Y) and the income elasticity of demand for the good is small, which is a reasonable assumption for urban transportation.

[4] This assumption rules out the situation considered in Kanemoto (1999) where travel on one mode is uncongestible and expansion of its capacity lowers its user costs, i.e., $c_2 = c_2(w_2)$. We shall consider this situation in section IV.

$$c_1 = c_1\left(\frac{m}{w_1}\right) \qquad c_2 = c_2\left(\frac{n}{w_2}\right) \tag{2}$$

where w_i is the capacity of mode i, with $c_i(0) \geq 0$, $c_i' > 0$ and $c_i'' \geq 0$ for $i = 1,2$.

Capacity construction costs are given by $K_i = K_i(w_i)$, with $K_i(0) = 0$ and $K_i' > 0$ but no restrictions on K_i'', $i = 1,2$.

The first-best planning problem, including the choice of decentralizing tolls, can be written as

$$\begin{aligned} \max_{\substack{p_1, p_2, w_1, w_2 \\ \tau_1, \tau_2}} \quad & B(p_1, p_2) + m(p_1, p_2)\tau_1 + n(p_1, p_2)\tau_2 - K_1(w_1) - K_2(w_2) \\ \text{s.t.} \quad & p_1 - c_1\left(\frac{m(p_1, p_2)}{w_1}\right) - \tau_1 = 0 \\ & p_2 - c_2\left(\frac{n(p_1, p_2)}{w_2}\right) - \tau_2 = 0. \end{aligned} \tag{3}$$

The maximand equals consumer surplus, $B(p_1, p_2)$ (see footnote 3), plus government surplus which equals toll revenue minus capacity construction costs. This is maximized w.r.t. p_i, w_i and τ_i, $i = 1,2$, and subject to the constraints (1). The planner has direct control over w_1, w_2, τ_1, and τ_2, but only indirect control over p_1 and p_2, as characterized by the constraints (1). It is assumed, as is empirically reasonable, that this problem has a unique maximum which is interior;[5] we write the solution as $\{p_i^*, w_i^*, \tau_i^*, m^*, n^*\}$ where * indicates the first-best optimum. The characteristics of the first-best optimum, including decentralizing tolls, are well-known. For each mode, optimal capacity should be such that the marginal social benefit of capacity, $-\frac{\partial}{\partial w_1}\left(mc_1\left(\frac{m}{w_1}\right)\right)$ and $-\frac{\partial}{\partial w_2}\left(nc_2\left(\frac{n}{w_2}\right)\right)$ respectively, equal the corresponding marginal cost; thus

[5] Or in other words, we assume that the second-best conditions are everywhere satisfied.

$$\left[\left(\frac{m}{w_1}\right)^2 c_1' - K_1'\right]^* = 0 \qquad \text{and} \qquad \left[\left(\frac{n}{w_2}\right)^2 c_2' - K_2'\right]^* = 0. \tag{4a}$$

Also, for each mode, the toll should be set equal to the corresponding congestion externality, $e_1 = \frac{m}{w_1} c_1'$ and $e_2 = \frac{n}{w_2} c_2'$, evaluated at the optimum — the standard Pigouvian prescription; thus,

$$\tau_i^* - e_i^* = 0, \qquad i = 1,2. \tag{4b}$$

The second-best problem can now be posed precisely. Append the constraint that $\tau_1 = \bar{\tau}_1\left(< \tau_1^*\right)$ to (3), and denote variables evaluated at the second-best optimum by **. Many papers simply write down and interpret the corresponding first-order conditions. Those papers that dig deeper adopt one of two approaches. The "local" approach is to determine $d\chi^{**}/d\bar{\tau}_1$ via comparative static analysis, where $\chi^{**} = \left(p_1^{**}, p_2^{**}, w_1^{**}, w_2^{**},\ \tau_2^{**}, m^{**}, n^{**}\right)$. The "global" approach is to determine $\chi^{**} - \chi^*$, using non-local arguments.

III. Components of the Problem

In the section, we analyze components of the problem. In the first subsection we analyze the Wheaton-Wilson problem, and in the second subsection the Doi problem.

III.1 The Wheaton-Wilson problem

Wheaton (1978) and Wilson (1983) considered the situation where there is only a single mode of transport from A to B — which they interpreted as car travel — and asked the question: If congestion on the road is underpriced $\left(\bar{\tau} < \tau^*\right)$, how does this affect optimal road width? Locally[6], what is $dw^{**}/d\bar{\tau}$? And globally, what determines $w^{**} - w^*$? We shall concentrate on the local approach, though we shall also comment on the global approach.

[6] Since there is only one mode, we shall drop the modal subscript for this subsection only.

The planner's problem analogous to (3) with the constraint that travel is underpriced, $\tau = \bar{\tau}\left(< \tau^*\right)$, is written as

$$\max_{p,w} \quad \mathrm{L} = B(p) + m(p)\bar{\tau} - K(w) + \lambda\left(p - c\left(\frac{m(p)}{w}\right) - \bar{\tau}\right), \tag{5}$$

where λ is the Lagrange multiplier on the constraint that $p - c\left(\frac{m(p)}{w}\right) - \bar{\tau} = 0$.

Intuition suggests that ${dw^{**}}/{d\bar{\tau}}$ is ambiguous in sign. Start at the first-best optimum and then lower the toll a discrete amount. This reduces the price of a trip, which in turn causes the number of trips to increase. Too many trips are taken and associated with this is a deadweight loss. The aim is to adjust road width, relative to the first-best level, to maximize social surplus. Expanding the road reduces the severity of congestion and hence the deadweight loss per traveler $(e - \bar{\tau})$ but causes the excessive number of travelers to increase; narrowing the road has the opposite effects.

There are two alternative ways of conceptualizing the problem. The first is a standard one in the theory of the second best. Altering road width from its first-best level can be viewed as another distortion. Then road width should be chosen to minimize the sum of the deadweight losses associated with the two distortions: $\bar{\tau} < \tau^*$ and $w \neq w^*$. From this point of view, adjusting road width entails an offsetting distortion. The second way of conceptualizing the problem is perhaps more intuitive in this context, and entails decomposing the marginal social benefit (MSB) from expanding the road. The direct effect is the immediate benefit, *holding the number of drivers fixed*. The indirect effect is the change caused by the road expansion in the deadweight loss from the underpricing of congestion, deriving from the induced increase in the number of drivers on the road.

Since the pricing of car travel does not affect the (construction) cost function of road expansion, raising the toll infinitesimally causes optimal road width to rise if and only if it causes the marginal social benefit (MSB) of road width to increase. Accordingly, the way we shall proceed is to examine the effect of raising the toll on the marginal social benefit of road width.[7]

- First-order conditions: a new decomposition

Let us now turn to the algebra. The first-order conditions are[8]

$$\text{p:}\quad B'(p)+m'(p)\bar{\tau}+\lambda\left(1-\frac{m'(p)}{w}c'\left(\frac{m(p)}{w}\right)\right)=0 \tag{6a}$$

$$\text{w:}\quad -K'(w)+\lambda\frac{m(p)}{w^2}c'\left(\frac{m(p)}{w}\right)=0\,. \tag{6b}$$

Noting that $B'(p)=-m(p)$ (by an argument analogous to that in fn. 3) and substituting out λ yields

$$-K'+\left(\frac{m-m'\bar{\tau}}{1-\frac{m'}{w}c'}\right)\frac{mc'}{w^2}=0, \tag{7}$$

where the arguments of functions are suppressed to simplify notation. Now, K' is the marginal social cost of road expansion, and so the other term in (7) is the marginal social benefit. To obtain a more illuminating expression, note that differentiation of the constraint with respect to w yields $dp/dw=-\left(c'm/w^2\right)\div\left(1-c'm'/w\right)$ so that $dm/dw=-\left(m'c'm/w^2\right)\div\left(1-c'm'/w\right)$, and recall that $e=\frac{c'm}{w}$ is the congestion externality. Then the marginal social benefit from the road expansion may be rewritten as

[7] We cast the analysis in terms of how a change in the toll affects *MSB*, holding *w* fixed. This entails viewing *MSB* as a function of *w* and $\bar{\tau}$, $MSB(w,\bar{\tau})$. Defining elasticities relative to this function, from the optimal road width condition $MSB(w,\bar{\tau})=K'(w)$:

$$E_{w:\bar{\tau}}=\frac{E_{MSB:\bar{\tau}}}{-E_{MSB:w}+\frac{K''w}{K'}},$$

Since the denominator is positive from the second-order conditions, the sign of $E_{w:\bar{\tau}}$ is determined by the sign of $\frac{dMSB}{d\bar{\tau}}$.

[8] We shall not record second-order conditions since we are examining local maxima. It should be noted, however that if the second-order conditions are not everywhere satisfied, there may be multiple local maxima so that a small

$$MSB = \left(\frac{m\left(1 - \frac{c'm'}{w}\right) + m'(e - \bar{\tau})}{1 - \frac{c'm'}{w}} \right) \frac{mc'}{w^2}$$

$$= m\left(-\frac{\partial c}{\partial w}\right) - \frac{dm}{dw}(e - \bar{\tau}), \qquad (8)$$

using $\partial c / \partial w = -c'm / w^2$. The first term on the RHS of (8) is the direct marginal benefit from the road expansion; the second is the indirect marginal cost, which equals the induced increase in the number of travelers times the deadweight loss associated with each traveler.[9] Since the literature refers to the induced increase in demand from a road expansion as *latent demand*, we may say that the indirect cost equals the increase in the deadweight loss due to the latent demand generated by the road expansion.

INSERT FIGURE 2

Figure 2: The marginal social benefit of a road expansion from w' to w'':
First-best and no-toll equilibria

The decomposition in (8) is particularly useful because it has a neat geometric interpretation. Figure 2 portrays the effects of a slight road expansion from w' to w'' for both the first-best equilibrium and the no-toll equilibrium. msc denotes marginal social cost $\frac{\partial}{\partial m}\left(mc\left(\frac{m}{w}\right)\right)$ and uc user cost $c\left(\frac{m}{w}\right)$, with $e = msc - uc$. Consider initially the first-best equilibrium (labeled E^0 in Figure 2) at which the full price equals marginal social cost. From (8), the benefit from the

change in τ can cause a jump from one local maximum to another. We also ignore corner maxima since they are of only secondary interest in this context.

[9] What is the interpretation of λ? Comparing (6b) and (8), we have that λ / m equals the ratio of the marginal social benefit to the direct marginal benefit. Thus, $(m - \lambda) / m$ gives the proportion of the direct marginal benefit from the road expansion that is dissipated through the combination of distorted pricing and latent demand.

road expansion is approximately $m\left(-\partial c/\partial w\right)(w''-w')$, evaluated at the first-best equilibrium, which equals the area *abcd*. Now consider the no-toll equilibrium (labeled *g* in Figure 2) at which the full price equals the user cost. From (8), the benefit from the road expansion has two components. The direct benefit is approximately $m\left(-\partial c/\partial w\right)(w''-w')$, evaluated at the no-toll equilibrium, which equals the area *efgh*. The indirect cost is approximately $e\left(dm/dw\right)(w''-w')=(msc-uc)\left(dm/dw\right)(w''-w')$ since the toll equals zero, which is given by the area *gijk*. Thus, the benefit from the road expansion is greater in the no-toll equilibrium than in the first-best equilibrium if and only if area *efgh* – area *gijk* exceeds area *abcd*. On one hand, the direct benefit from the road expansion is larger in the no-toll equilibrium than in the first-best equilibrium (area *efgh* exceeds area *abcd*), for two reasons: first, because congestion is unpriced, more people travel on the road in the no-toll equilibrium; and second, because the road is more congested in the no-toll equilibrium, each traveler derives greater benefit from a given road expansion. On the other hand, expansion of the road generates indirect costs due to latent demand in the no-toll equilibrium but not in the first-best equilibrium. The figure suggests that which of these two effects dominates is *a priori* ambiguous, so that the marginal social benefit from a road expansion may be either higher or lower in the no-toll equilibrium than in the first-best equilibrium. And such is indeed the case, as is shown below.

- Comparative statics and informational requirements

Write the expression derived earlier for dm/dw in terms of the congestion externality and the price elasticity of demand for travel, $\varepsilon \equiv -m'p/m$:

$$\frac{dm}{dw}=\frac{m}{w}\left(\frac{e\varepsilon}{p+e\varepsilon}\right). \tag{9a}$$

Then (8) can be rewritten as

$$MSB=\frac{me}{w}-\frac{m}{w}\left(\frac{e\varepsilon}{p+e\varepsilon}\right)(e-\overline{\tau}) \tag{9b}$$

so that (holding w fixed)

$$\frac{dMSB}{d\bar{\tau}} = -\frac{me\varepsilon p(1+\gamma)}{w(p+e\varepsilon)^2} - (e-\bar{\tau})\frac{d}{d\bar{\tau}}\left(\frac{dm}{dw}\right) \qquad (9c)$$

where $\gamma \equiv c'' \frac{m}{w} \Big/ c'$ is the elasticity[10] of c' with respect to $m\!/\!_w$. With an increase in the toll, the road should be expanded if $dMSB\!/\!_{d\bar{\tau}} > 0$ and narrowed when the inequality is reversed. Three points are worthy of note.

1. $\left.\frac{dMSB}{d\bar{\tau}}\right|_{\bar{\tau}=e} < 0$ for $\varepsilon > 0$ and finite (from (9c)). Thus, lowering the toll incrementally from its optimal level causes optimal capacity to increase. This is the result obtained by Wheaton. It can be explained with reference to Figure 2. Because deadweight loss is a second-order term when the toll is set incrementally below the first-best level, the analog to area *gijk* disappears, and only the analog to area *abcd* exists.

2. When demand is perfectly elastic, so that road expansion results in an equiproportional increase in drivers $\left(\frac{dm}{dw} = \frac{m}{w}\right)$, (9c) reduces to

$$\left.\frac{dMSB}{d\bar{\tau}}\right|_{\varepsilon=\infty} = \frac{m}{w}\left(1-\frac{\bar{\tau}}{e}\right)$$

which is positive for $\bar{\tau} < e$. In this case, when the road is underpriced, lowering the toll marginally lowers second-best capacity.

3. Eqs. (9a) and (9c) together indicate that the sign of $\frac{dMSB}{d\bar{\tau}}$ depends not only on the magnitude of the demand elasticity and of the excess burden from an additional traveler, but also on the rate at which these change with price, and hence on *second* derivatives of the demand and user cost functions. The demand for trips is not a primitive of tastes, and the demand elasticity varies over time and place, depending among other things on the flexibility of work hours and the

configuration of the road and rail networks. Therefore in a specific situation the demand elasticity is known only imprecisely. The rate at which the demand elasticity varies with price is known much less precisely, so imprecisely in fact that we doubt whether it is feasible to determine even the sign of $\frac{dMSB}{d\bar{\tau}}$ with any confidence. We have therefore come to the view that attempting to determine second-best road width through strictly local analysis is futile; the approach is simply too informationally demanding.

- Global analysis

The "broad" or global shape of the demand curve may be known reasonably well, even when the shapes of small segments of the demand curve are not. In these circumstances, the empirical application of global analysis may be informationally feasible even when the informational requirements for the empirical application of local analysis are excessive. Let us explore this idea.

INSERT FIGURE 3 HERE

Figure 3: w^{**} as a function of $\bar{\tau}$

Figure 3 plots two possible ways in which second-best road width may vary with the toll – indicated by locus (a) and locus (b). Wheaton's result that $\left.\frac{dw^{**}}{d\bar{\tau}}\right|_{\bar{\tau}=\tau^*} < 0$, indicates that both loci are negatively-sloped in the neighborhood of the first-best optimum.[11] Locus (a) illustrates that, because of this, it is quite possible for second-best road width to exceed first-best road width for all suboptimal levels of the toll ($\tau \in [0, \tau^*)$) even when second-best road width is increasing in the level of the toll over part of this range. This demonstrates the obvious but important point that

[10] The most commonly-used congestion cost function is the so-called Vickrey congestion cost function: $c\left(\frac{m}{w}\right) = h_0 + h_1\left(\frac{m}{w}\right)^{\beta}$. In terms of this function, $\gamma = \beta - 1$.

[11] Wheaton's result also implies that it is not possible for $w^{**}(\bar{\tau}) < w^*$ for all $\bar{\tau} \in [0, \tau^*)$.

$w^{**}(\bar{\tau}) > w^*$ for all $\bar{\tau} \in [0, \tau^*)$ is a weaker condition than $\frac{dw^{**}}{d\bar{\tau}} < 0$ for all $\bar{\tau} \in [0, \tau^*)$. Locus (b) illustrates that, despite Wheaton's result, second-best road width may be less than first-best road width over an interval of suboptimal tolls. A final observation is that since $w^{**}(\bar{\tau}) - w^*$ is (minus) the integral of $\frac{dw^{**}}{d\tau}$ from $\bar{\tau}$ to τ^*, and since $\frac{dw^{**}}{d\bar{\tau}}$ depends on derivatives of demand elasticities, the properties of the demand function which affect $w^{**}(\bar{\tau}) - w^*$ are arc elasticities or average slopes.

These observations prompted Wilson to enquire into sufficient *global* conditions on the demand function and the form of the congestion cost function such that the underpricing of congestion causes second-best road width to exceed first-best road width for all $\bar{\tau} \in \left[0, \tau^*\right)$. He established the following result (rewritten according to our terminology and notation):

If: i) $c\left(\frac{m}{w}\right) = h_0 + h_1\left(\frac{m}{w}\right)^{\beta}$, h_0, h_1, β constants, $h_0 \geq 0$, $h_1 > 0$, $\beta \geq 1$

ii) $\frac{d\varepsilon}{dp} \geq 0$

iii) $\varepsilon < \frac{p}{c - h_0}$

then $w^{**}(\bar{\tau}) > w^*$.

These are all primitive or structural conditions on tastes and technology, and are much simpler than the conditions which would be required to establish that $\frac{dw^{**}}{d\bar{\tau}} < 0$ for all $\bar{\tau} \in \left[0, \tau^*\right)$.

Using global analysis, it may be possible to establish upper and lower bounds relating w^{**} to w^* and $\bar{\tau}$. This line of investigation holds more promise for practical policy application than the alternative local, comparative-static approach with its unreasonably demanding empirical requirements for implementation. Appendix 1 describes one such approach. Given the congestion cost function, the demand function for which $w^{**} = w^*$ is solved for; we term this the "w^*-demand function". Then a necessary condition for $w^{**} > w^{*}$ for *all* suboptimal levels of the toll is derived based on a global comparison of the w^{*}-demand function and the actual demand function. d'Ouville and McDonald (1990) have undertaken an analysis along these lines. They model congestion using a production function rather than a cost function; in particular, they assume that trips are produced under constant returns to scale with aggregate travel time and capacity as inputs. They establish the following sufficient conditions (using our notation unless indicated otherwise) for $w^{**} > w^{*}$ with a suboptimal toll: i) $\varepsilon < 1 + \frac{Tm}{L}$, where L is aggregate travel time and T the toll expressed in units of time; and ii) $\sigma < 1$, where σ is the elasticity of substitution between aggregate travel time and capacity in the trip production function. They also provide a nice geometric interpretation of this result.[12]

- An alternative perspective: volume-capacity ratio as policy instrument

Often additional insight into a problem may be achieved by looking at it from a different perspective. Consider the transformation of variables $\theta \equiv \frac{m}{w}$; θ is the volume-capacity ratio, a standard measure of the degree of congestion. We suppose that the planner chooses θ rather than w; she adjusts road width to achieve a desired level of congestion. Then the planning problem may be written as

$$\max_{p,\theta} \quad B(p) + m(p)\bar{\tau} - K\left(\frac{m(p)}{\theta}\right) \qquad s.t. \quad p = c(\theta) + \bar{\tau}.$$

Substituting out p yields the unconstrained maximization problem

$$\max_{\theta} \quad B\big(c(\theta) + \bar{\tau}\big) + m\big(c(\theta) + \bar{\tau}\big)\bar{\tau} - K\left(\frac{m\big(c(\theta) + \bar{\tau}\big)}{\theta}\right). \tag{10}$$

The first-order condition is

[12] Unfortunately, this result is not as general as it at first appears to be. The standard approach to modelling congestion assumes that user cost is a function of the *ratio* of volume to capacity, which accords very well with empirical observation. d'Ouville and McDonald make the quite different assumption that number of trips is a function of the ratio of aggregate travel time to capacity. The two assumptions are equivelent when the money cost

$$B'c' + m'c'\bar{\tau} - K'\frac{m'c'}{\theta} + K'\frac{m}{\theta^2} = 0,$$

which may be rewritten as

$$-m\left(c' - \frac{K'}{\theta^2}\right) - c'm'\left(\frac{K'}{\theta} - \bar{\tau}\right) = 0. \qquad (11)$$

It is straightforward to demonstrate that at the optimum with underpriced congestion and $\varepsilon > 0$, $c' - \frac{K'}{\theta^2} > 0$ and $\frac{K'}{\theta} - \bar{\tau} > 0$.

Consider adding an extra traveler to the road, *holding* θ *fixed*. Doing so generates toll revenue of $\bar{\tau}$. Also, to hold θ fixed, the road must be expanded by $\left.\frac{dw}{dm}\right|_{\bar{\theta}} = \left.\frac{d\left(\frac{m}{\theta}\right)}{dm}\right|_{\bar{\theta}} = \frac{1}{\theta}$, which has a cost of $\frac{K'}{\theta}$. Thus, $\frac{K'}{\theta} - \bar{\tau}$ may be interpreted as the *long-run marginal deadweight loss* — the increase in deadweight loss from adding an extra traveler to the road, holding constant the volume-capacity ratio rather than, as in the standard measure, road width.

Now consider marginally increasing the road's volume-capacity ratio. Decompose the resulting expression into two effects: the direct effect or the effect holding m constant, and the indirect effect due to latent demand. The first term on the LHS of (11) is the direct effect and the second term the indirect effect. Observe that the indirect effect is the marginal (as θ increases) deadweight loss from the underpricing of congestion, while the direct effect is the marginal deadweight loss from setting the volume-capacity ratio above its first-best level. Thus, this formulation nicely relates the two alternative ways of looking at the second-best problem — direct/indirect effects and offsetting distortions.

This formulation generates the additional insight that, with congestion underpriced, the second-best volume-capacity ratio should exceed the first-best volume-capacity ratio, which we

of travel is ignored and the fixed component of trip cost is zero. Thus, d'Ouville and McDonald's specification essentially ignores the fixed component of trip cost.

term “overcrowding”. Thus, even though it is not possible to say in general whether second-best road width is less than or greater than first-best road width, it is always the case with underpriced congestion that the road should be more congested than in the first-best optimum. This result was first intimated by Mohring (1970) but seems to have been overlooked in most of the subsequent literature.

INSERT FIGURE 4 HERE

Figure 4: The second-best optimum: Constant marginal construction costs and no toll

Figure 4 provides a diagrammatic depiction of (11) for the case of constant marginal construction costs ($K' = k$ so that $-m\left(c' - \frac{k}{\theta^2}\right) - c'm'\left(\frac{k}{\theta} - \bar{\tau}\right) = 0$) and no toll. At the second-best optimum, the marginal deadweight loss from underpricing congestion, given by area A, equals the marginal deadweight loss from overcrowding, given by area B.

- Modern respecifications

The “Wheaton-Wilson problem” is a particular specification of the problem of determining the second-best capacity of a congestible facility given that it is underpriced. Its specification reflects the state of the second-best/optimal-tax literature in the early seventies. A modern formulation of the problem would take into account that the government’s deficit related to the road must be financed by distortionary taxation for reasons related to asymmetric information. Following Vickrey (1954), this can be done by assuming an exogenous “marginal deadweight loss of public funds”: ϕ; for example, $\phi = .8$ indicates that the marginal dollar of revenue raised by the government generates a deadweight loss of eighty cents. Taking this into account, the first-best problem we have considered would become a second-best problem, and the second-best problem a third-best problem. Let us consider the latter. Since the magnitude of the deficit is $K(w) - m(p)\tau$, the analog to (5) is

$$\max_{p,w} \quad \mathsf{L} = B(p) + (1+\phi)\left(m(p)\bar{\tau} - K(w)\right) + \lambda\left(p - c\left(\frac{m(p)}{w}\right) - \bar{\tau}\right). \qquad (5')$$

The marginal social cost of capacity rises from $K'(w)$ to $(1+\phi)K'(w)$ and the marginal social benefit is given by (8) plus an extra term $\frac{dm}{dw}\phi\bar{\tau}$, which incorporates the reduction in the deadweight loss due to distortionary taxation from the toll revenue raised from the "latent demanders". The modern formulation might also: account for the adverse incentive effects from automatically covering the road authority's deficit, à la Laffont-Tirole (1993); replace the underpricing constraint by a political economy constraint incorporating those political considerations that preclude efficient tolling; and include distributional considerations[13] by maximizing social welfare instead of social surplus and perhaps by incorporating constraints on the redistribution possible through the tax system.

- Naïve cost-benefit analysis

Thus far, we have followed Wheaton and Wilson in comparing the first- and second-best road widths. Another comparison, which is perhaps more relevant, is between second-best road width and the road width chosen by transportation planners using "naïve" cost-benefit analysis, under which the benefits of a road expansion are measured as the direct benefits — the planners ignore the indirect costs associated with the combination of underpricing and latent demand. It seems that naïve cost-benefit analysis is still widely employed by transportation planners.

Let MSB and $MSB^\dagger$ be the marginal social benefit of road capacity computed correctly and computed incorrectly according to naïve cost-benefit analysis. Adapting (8) and applying (9b):

$$\frac{MSB^\dagger}{MSB} = \frac{m\left(-\frac{\partial c}{\partial w}\right)}{m\left(-\frac{\partial c}{\partial w}\right) - \frac{dm}{dw}(e-\bar{\tau})} = \frac{1}{1 - \frac{\varepsilon}{p+e\varepsilon}(e-\bar{\tau})}.$$

[13] The argument against including distributional considerations is that they are best dealt with through the income tax. It is noteworthy that Vickrey, a champion of the poor and underprivileged, did not incorporate distributional considerations into most of his discussions of urban transport policy.

Thus, with underpricing of congestion, the procedure employed in naïve cost-benefit analysis consistently overestimates the marginal social benefit of road capacity. In terms of Figure 3, naïve cost-benefit analysis measures marginal social benefit as area *efgh* whereas the true marginal social benefit is area *efgh* – area *gijk*.

Let us consider a numerical example. Suppose that the elasticity of demand for road travel is $2/3$, that the congestion cost function has the Vickrey form $c\left(\frac{m}{w}\right) = h_0 + h_1\left(\frac{m}{w}\right)^{\beta}$, with $h_0 = 1$, $h_1 = 1$, and $\beta = 3$. Suppose furthermore that the toll is zero, that average construction costs are constant, and that the level of demand and construction costs are such that road capacity chosen (incorrectly) according to naïve cost-benefit analysis is 1.0, and that the volume-capacity ratio at this level of capacity is 1.0 (so that in the corresponding equilibrium user cost is double free-flow user cost). Simple calculation yields that the corresponding second-best optimal capacity is .64, implying a volume-capacity ratio of 1.23, and an equilibrium user cost 2.87 times free-flow user cost. Table 1 below presents these and other variables of interest for the first-best optimum, the second-best optimum, and the naïve optimum.

Table 1 : Numerical example of the naïve cost-benefit analysis

	p	m	w	θ	c	e	τ	DWL
First-best	5	.54	.54	1	2	3	3	0
Second-best	2.87	.79	.64	1.23	2.87	5.61	0	.54
Naive	2	1	1	1	2	3	0	.86

Note: $c\left(\frac{m}{w}\right) = 1 + \left(\frac{m}{w}\right)^3$, $K(w) = 3w$, $m = \sqrt[3]{4}\,p^{-2/3}$

The deadweight loss is calculated as the loss in surplus relative to the first-best optimum. Three results are of particular interest. First, in the naïve optimum the deadweight loss per traveler is 43% of the full price, which is substantial; second, the deadweight loss in the naïve optimum is

1.6 times that in the second-best optimum; and third, the naïve *w* is 56% larger than the second-best *w*, which suggests that the inappropriate use of the naïve cost-benefit rule may be more important in the context of policy analysis than the earlier question of second-best vs. first-best capacity.

III.2 The Doi problem

- Statement of problem, first-order conditions, and interpretation

Consider two modes in parallel. The capacities of both modes are fixed. One mode — car travel — is unalterably underpriced $\left(\tau_1 = \bar{\tau}_1 < \tau_1^*\right)$. How should the other mode be priced? As noted earlier, numerous papers have studied this problem. Most of these papers make the empirically unrealistic assumption that the two modes are perfect substitutes. Doi (1986) gave a very comprehensive analysis, providing a general treatment of demand and considering a variety of transit authority objective functions. His algebraic analysis is accompanied by an insightful geometric analysis. We shall follow Doi (1986) in providing a general treatment of demand, so that the two modes may be imperfect substitutes or even complements.

The planner's problem analogous to (3) but with the constraints that capacities are fixed and car travel underpriced is

$$\begin{aligned}\max_{p_1, p_2} \quad & B(p_1, p_2) + m(p_1, p_2)\left(p_1 - c_1\left(\frac{m(p_1, p_2)}{w_1}\right)\right) \\ & + n(p_1, p_2)\left(p_2 - c_2\left(\frac{n(p_1, p_2)}{w_2}\right)\right) + \lambda\left(p_1 - c_1\left(\frac{m(p_1, p_2)}{w_1}\right) - \bar{\tau}_1\right). \qquad (12)\end{aligned}$$

The corresponding first-order conditions are

$$p_1: \quad -m + m_1(p_1 - c_1) + m\left(1 - c_1' \frac{m_1}{w_1}\right) + n_1(p_2 - c_2) - nc_2' \frac{n_1}{w_2} + \lambda\left(1 - c_1' \frac{m_1}{w_1}\right) = 0 \qquad (13a)$$

$$p_2: \quad -n + m_2(p_1 - c_1) - mc_1' \frac{m_2}{w_1} + n_2(p_2 - c_2) + n\left(1 - c_2' \frac{n_2}{w_2}\right) - \lambda c_1' \frac{m_2}{w_1} = 0 \qquad (13b)$$

where the subscripts on the *m*'s and *n*'s denote partial derivatives, e.g., $m_2 = \frac{\partial m}{\partial p_2}$. Substituting out λ from this pair of equations, employing $p_i - c_i - \tau_i = 0,\ i = 1,2$, and then simplifying the resulting equations using the expressions for ${}^{dm}\!/\!{}_{d\tau_2}$ and ${}^{dn}\!/\!{}_{d\tau_2}$ obtained from total differentiation of $p_i - c_i - \tau_i = 0,\ i = 1,2$, yields

$$\frac{dm}{d\tau_2}\left(e_1 - \bar{\tau}_1\right) + \frac{dn}{d\tau_2}\left(e_2 - \tau_2\right) = 0 \qquad (14a)$$

or

$$\frac{e_2 - \tau_2}{e_1 - \bar{\tau}_1} = \frac{m_2}{\frac{e_1}{m}\left(n_2 m_1 - n_1 m_2\right) - n_2}. \qquad (14b)$$

Eq. (14a) demonstrates clearly that the choice of τ_2 entails offsetting distortions. τ_2 should be above or below the rail congestion externality so as to offset the distortion associated with underpriced car travel. But this causes distortion in the market for rail travel. τ_2 should be set at the level which minimizes the sum of the two deadweight losses. Put alternatively, τ_2 should be set such that the increase in deadweight loss for one mode from raising or lowering τ_2 an incremental amount is just offset by the decrease in deadweight loss for the other mode.

Eq. (14a), while insightful, does not indicate whether τ_2 should be set above or below the level of the congestion externality in rail travel. This is indicated by (14b). Note that $m_1 = n_2 < 0$ (substitution effects) and $n_2 m_1 - n_1 m_2 > 0$ (strict concavity of the utility function). Thus, if road and rail are substitutes in demand $\left(n_1 = m_2 > 0\right)$, with underpriced road congestion the rail fare should be set *below* the congestion externality for rail; in the less likely event that they are complements, the rail fare should be set above the corresponding congestion externality. The intuition is as follows. Start at the situation where the rail fare is set equal to the rail congestion externality. Adjust the fare infinitesimally so as to reduce the level of car travel; if car and subway are substitutes, the rail fare should be lowered, and if they are complements the fare should be raised. This reduces the deadweight loss associated with car travel but by the Envelope Theorem causes only a second-order welfare loss for rail travel. Then continue adjusting the rail fare in the same direction until condition (14a) is satisfied.

INSERT FIGURE 5

Figure 5: Diagrammatic depiction of (14a) when road and rail are substitutes in demand

Eq. (14a) has an easy geometric interpretation. Lower τ_2 from τ_2' to $\tau_2' - \Delta_2$. The immediate effect is to lower the price of rail travel and increase both the number of rail travelers and the deadweight loss from underpriced rail travel. Assuming that road and rail are substitutes in demand, the decrease in the price of rail travel causes the demand curve for car travel to shift down, which results in a fall in the number of car travelers and hence in the deadweight loss from underpriced car travel. The price of car travel falls as well, which causes the demand curve for rail travel to shift down slightly, and so on. The final result is as shown in the Figure 5, for the situation where (14a) is satisfied, so that the marginal reduction in the rail fare causes a fall in the deadweight loss for car travel (the heavily shaded area in the left-hand diagram) just equal to the rise in the deadweight loss for rail travel (the heavily shaded area in the right-hand diagram).

Eq.(14a) can be expressed in other ways. From $p_1 = c_1\left(\frac{m(p_1, p_2)}{w_1}\right) + \bar{\tau}_1$ and $p_2 = c_2\left(\frac{n(p_1, p_2)}{w_2}\right) + \tau_2$, the p_i's , and hence all other variables may be expressed as functions of $\bar{\tau}_1$ and τ_2, e.g., $m = \tilde{m}(\bar{\tau}_1, \tau_2)$ and $n = \tilde{n}(\bar{\tau}_1, \tau_2)$. Not only does this change of variables simplify the algebra, but also empirical work measures demand elasticities with respect to measurable components of trip prices, including fares and tolls (Oum et al. (1992)), and not with respect to the full prices which are unobservable. Thus,

$$\frac{\frac{msc_2 - p_2}{p_2}}{\frac{msc_1 - p_1}{p_1}} = -\left(\frac{\frac{d\tilde{m}}{d\tau_2}}{\frac{d\tilde{n}}{d\tau_2}}\right)\left(\frac{p_1}{p_2}\right) \qquad \text{(using } e_i - \tau_i = msc_i - p_i,\ i = 1,2 \text{ and (14a))}$$

$$= \left(\frac{E_{\tilde{m}:\tau_2}}{-E_{\tilde{n}:\tau_2}}\right)\left(\frac{p_1 m}{p_2 n}\right), \tag{14c}$$

which indicates that, when rail and road are substitutes in demand, the proportional underpricing of rail $\left(\frac{msc_2 - p_2}{p_2}\right)$ is positively related to the proportional underpricing of road $\left(\frac{msc_1 - p_1}{p_1}\right)$, inversely related to the ratio of consumer "expenditure" on rail relative to road $\left(\frac{p_2 n}{p_1 m}\right)$, inversely related to the rail elasticity of demand with respect to the rail fare, and positively related to the cross-price elasticity of demand for road travel with respect to the rail fare.[14]

- Respecification of the rail technology

In discussing an earlier draft of this paper, Small argued that the two-mode problem as specified by us and in the previous literature is more appropriately interpreted as "streets and freeways" than as "road and rail" since it provides a poor description of the technology of rail travel. We now consider this criticism in the context of the Doi problem, with capacities fixed, in which case our treatment is deficient in ignoring the service frequency of rail.

We suppose that the trackage and the number of rail carriages are fixed, but that *service frequency* is variable, more frequent service therefore entailing fewer carriages per train.[15] Increasing service frequency increases costs because more operators have to be employed, etc. Let f denote the service frequency of rail, a policy variable, $s(f)$ the variable costs associated with service frequency $(s' > 0)$, and $\hat{c}(n, f; \Omega_2)$ the user cost associated with rail travel (where Ω_2

[14] The amount by which the rail fare should be raised in response to a unit rise in the road toll can be calculated. We do not present the result since it is a rather unenlightening mess of demand elasticities, rates of change of demand elasticities, etc.

[15] This specification draws on Kraus and Yoshida (2000). We have in mind an LRT or subway system where each carriage has an engine. For train travel, increasing service frequency requires increasing the number of engines.

incorporates the rail capacity variables) which captures both crowding and the schedule delay associated with infrequency of service

$$\left({\partial \hat{c}}/{\partial n} > 0,\ {\partial^2 \hat{c}}/{\partial n^2} > 0,\ {\partial \hat{c}}/{\partial f} < 0,\ {\partial^2 \hat{c}}/{\partial f^2} > 0,\ {\partial^2 \hat{c}}/{\partial n \partial f}\ ? \right).$$

The planner's problem analogous to (12) is

$$\max_{p_1, p_2, f} \quad B(p_1, p_2) + m(p_1, p_2)\left(p_1 - c_1\left(\frac{m(p_1, p_2)}{w_1} \right) \right)$$

$$+ n(p_1, p_2)\left(p_2 - \hat{c}\left(n(p_1, p_2), f; \Omega_2 \right) \right) - s(f) + \lambda_1\left(p_1 - c_1\left(\frac{m(p_1, p_2)}{w_1} \right) - \bar{\tau}_1 \right). \qquad (12')$$

It is straightforward to demonstrate that (14a) continues to hold. Indeed, this can be seen by inspection. Set service frequency at its second-best optimal level, conditional on $\bar{\tau}_1$. Then the maximization problem (12´) reduces to the maximization problem (12). Thus, consideration of the service frequency of rail, while practically important, does not substantially alter the economics of rail *pricing*.

IV. The Full Problem

The stage is now set to consider the full second-best problem in which the rail fare, rail capacity, and road capacity are all adjusted in response to the underpricing of car travel.

IV.1 Paradoxes[16]

An interesting place to start is the Pigou-Knight-Downs (PKD) Paradox restated in the context of the paper. Suppose that road and rail are perfect substitutes in demand, that congestion in rail travel is insignificant, that rail operating costs are independent of the number of rail passengers, and that both road and rail are untolled. The PKD Paradox states that with total demand for travel fixed ($m + n =$ constant) an expansion of the road generates zero *gross* benefits, as long as some travelers use the rail after the expansion. Travelers divide themselves across the two modes so as to equalize the full price of travel. Since an expansion of the road does not alter the full price of rail travel, it does not alter the equilibrium full price of travel and therefore confers no benefit, nor does it generate any reduction in rail operating costs. All it does is to divert traffic from rail to road. The Paradox disappears when congestion pricing is applied to the road. Thus, Vickrey (1963) captures the Paradox aptly in stating: The road is worthless precisely because it is free.

It is not difficult to modify the model so that the road is worse than worthless.[17] First, if there is a rail fare, the diversion of traffic to the road causes a reduction in fare revenue collected, and this loss of revenue is pure deadweight loss. Second, suppose that, in response to the reduction in rail traffic, the rail authority reduces train frequency while holding the fare constant. This raises average waiting time and hence the full price of a train trip, which causes the equilibrium full price of travel to rise. If, in deciding on train frequency, the rail authority fails to take the effects of its decision on road travel fully into account, as seems realistic, the increase in total travel costs will exceed the cost savings from reducing train frequency. In this paradoxical world, a reduction in the rail fare is very effective, so effective in fact that set at the right level it can completely neutralize the deadweight loss associated with the underpricing of auto travel.

These paradoxes are much alluded to by environmentalists who use them to support their opposition to new highway projects. But paradoxes are often paradoxical because they derive from unrealistic assumptions. The above paradoxes are due, at least in part, to the unrealistic assumption that road and rail are perfect substitutes in demand. Let us examine how robust the PKD Paradox, extended to incorporate tolls, is to a more general specification of demand.

The marginal social benefit from a road expansion is given by

$$MSB_1(w_1) = \frac{d}{dw_1}\left(B(p_1, p_2) + m(p_1, p_2)\bar{\tau}_1 + n(p_1, p_2)\bar{\tau}_2\right),$$

[16] These paradoxes are discussed at greater length in Arnott and Small (1994).

[17] The paradox is then labeled the Downs-Thomson Paradox.

where $\frac{dp_1}{dw_1}$ and $\frac{dp_2}{dw_1}$ are obtained from total differentiation of the pricing constraints

$p_1 - c_1\left(\frac{m(p_1, p_2)}{w_1}\right) - \bar{\tau}_1 = 0$ and $p_2 - \bar{c}_2 - \bar{\tau}_2 = 0$, where $\bar{c}_2$ is the constant user cost for rail. Thus,

$$\begin{aligned} MSB_1(w_1) &= (-m + m_1\bar{\tau}_1 + n_1\bar{\tau}_2)\frac{dp_1}{dw_1} + (-n + m_2\bar{\tau}_1 + n_2\bar{\tau}_2)\frac{dp_2}{dw_1} \\ &= (-m + m_1\bar{\tau}_1 + n_1\bar{\tau}_2)\left(\frac{-c_1'\frac{m}{w_1^2}}{1 - c_1'\frac{m_1}{w_1}}\right) \qquad \left(\text{since } \frac{dp_2}{dw_1} = 0\right) \qquad (15) \\ &= \left(\frac{me_1}{w_1}\right)\left(\frac{p_1 + \varepsilon_1\bar{\tau}_1 - \eta_{21}\frac{n}{m}\bar{\tau}_2}{p_1 + e_1\varepsilon_1}\right). \end{aligned}$$

where $\varepsilon_1 \equiv -\frac{m_1 p_1}{m}$, $\eta_{21} \equiv \frac{n_1 p_1}{n}$, and $e_1 = c_1'\frac{m}{w_1}$. This is identical to (9b), except for the addition of the cross-price elasticity term, which captures the fare revenue effect noted above. Since $\frac{me_1}{w_1}$ is the direct marginal social benefit of the road expansion,

$$D = \frac{(e_1 - \bar{\tau}_1)\varepsilon_1 + \eta_{21}\frac{n}{m}\bar{\tau}_2}{p_1 + e_1\varepsilon_1} \qquad (16)$$

is the proportion of the direct benefit dissipated through distortion. Suppose, for example, that $\bar{\tau}_1 = 0$ so that $p_1 = c_1$, $\frac{e_1}{p_1} = 3$, $\varepsilon_1 = .5$, $\eta_{21} = 1$, $\frac{n}{m} = .2$, and $\bar{\tau}_2 = p_1$; then $D = .68$. With this set of parameters, the latent demand generated by the road widening, which derives from both *traffic creation* and *traffic diversion* from rail to road, does not completely neutralize the direct benefits, but does eliminate a considerable proportion of them. With other sets of parameter values, it is possible that *D*>1, implying that the marginal social benefit of the road expansion is negative.

Now, let us consider the optimal second-best rail fare. Since there is no congestion on the railway, the first-best fare is zero. With $\bar{\tau}_1 = 0$ the optimal second-best fare is

$\tau_2 / e_1 = \frac{dm}{d\tau_2} \bigg/ \frac{dn}{d\tau_2} < 0$, which is simply the particularization of (14a) with $e_2 = \bar{\tau}_1 = 0$. Therefore, the second-best fare τ_2 should be set such that the marginal gain in social surplus due to traffic diversion from the underpriced and congestible road to the uncongestible rail equals the increase in the budgetary cost of subsidizing rail travel deriving from increased rail traffic.

Having sampled some of its economics, let us now turn to analysis of the full problem.

IV.2 Second-best optimality conditions

The problem is the same as the first-best planning problem described in (3) except that the road toll is fixed:

$$\max_{\substack{p_1, p_2, w_1, w_2 \\ \tau_2}} \quad B(p_1, p_2) + m(p_1, p_2)\bar{\tau}_1 + n(p_1, p_2)\tau_2 - K_1(w_1) - K_2(w_2)$$

$$s.t. \qquad p_1 - c_1\left(\frac{m(p_1, p_2)}{w_1}\right) - \bar{\tau}_1 = 0 \qquad \lambda_1 \tag{17}$$

$$p_2 - c_2\left(\frac{n(p_1, p_2)}{w_2}\right) - \tau_2 = 0 \qquad \lambda_2$$

The first-order conditions are

$$p_1: \qquad -m + m_1\bar{\tau}_1 + n_1\tau_2 + \lambda_1\left(1 - c_1' \frac{m_1}{w_1}\right) - \lambda_2 c_2' \frac{n_1}{w_2} = 0 \tag{18a}$$

$$p_2: \qquad -n + m_2\bar{\tau}_1 + n_2\tau_2 - \lambda_1 c_1' \frac{m_2}{w_1} + \lambda_2\left(1 - c_2' \frac{n_2}{w_2}\right) = 0 \tag{18b}$$

$$\tau_2: \qquad n - \lambda_2 = 0 \tag{18c}$$

$$w_1: \qquad -K_1' + \lambda_1 c_1' \frac{m}{w_1^2} = 0 \tag{18d}$$

$$w_2: \qquad -K_2' + \lambda_2 c_2' \frac{n}{w_2^2} = 0. \tag{18e}$$

The interpretation of the first-order conditions is straightforward. Their only remarkable feature is that $\lambda_2 = n$; thus, at the second-best optimum, none of the direct benefits from rail capacity expansion are dissipated through the combination of distorted pricing and latent demand. We shall explain this result shortly.

After some manipulation, the first-order conditions can be rewritten as

$$\tau_2 : \qquad \frac{dm}{d\tau_2}(e_1 - \overline{\tau}_1) + \frac{dn}{d\tau_2}(e_2 - \tau_2) = 0 \tag{19a}$$

$$w_1 : \qquad m\left(-\frac{\partial c_1}{\partial w_1}\right) - \frac{dm}{dw_1}(e_1 - \overline{\tau}_1) - \frac{dn}{dw_1}(e_2 - \tau_2) - K_1' = 0 \tag{19b}$$

$$w_2 : \qquad n\left(-\frac{\partial c_2}{\partial w_2}\right) - \frac{dm}{dw_2}(e_1 - \overline{\tau}_1) - \frac{dn}{dw_2}(e_2 - \tau_2) - K_2' = 0. \tag{19c}$$

Eq. (19a) has the same explanation as (14a) in section III.2. The rail fare should be set so as to minimize the sum of the deadweight losses associated with, first, the road toll not equaling the road congestion externality and, second, the rail fare not equaling the rail congestion externality. Eqs. (19b) and (19c) are the generalizations one would expect on the basis of the intuition given for (8) — the analogous result for the Wheaton-Wilson problem.

Comparing (19c) and (18e) with $\lambda_2 = n$, it follows that at the second-best optimum

$$\frac{dm}{dw_2}(e_1 - \overline{\tau}_1) + \frac{dn}{dw_2}(e_2 - \tau_2) = 0; \tag{20}$$

the indirect cost of expanding rail capacity, resulting from distortion and latent demand, equals zero. Why? Since a change in both τ_2 and w_2 operate through the same "channel"[18] — the full price of rail travel — their effects on m and n are identical up to a constant of proportionality, so that $\left.\frac{dm}{dw_2}\middle/\frac{dn}{dw_2}\right. = \left.\frac{dm}{d\tau_2}\middle/\frac{dn}{d\tau_2}\right.$. Since the level of the rail fare is chosen to neutralize indirect costs (eq. (19a)), the indirect cost of an expansion of rail capacity is also neutralized. The same is not true of an expansion of road capacity, *viz.* $\frac{dm}{dw_1}\Big/\frac{dn}{dw_1} \neq \frac{dm}{d\tau_2}\Big/\frac{dn}{d\tau_2}$, and hence the indirect cost of road expansion is not neutralized by the rail fare and indeed looms large in the analysis.

Eqs. (19) are simple and intuitive. However, since they implicitly incorporate the dependence of the full prices on the policy variables and on $\bar{\tau}_1$, their total differentiation results in considerable algebraic complexity. Furthermore, since there are now three policy variables rather than one, the problem is no longer amenable to simple geometric analysis.

IV.3 Discussion

- Analytical issues

How then to proceed? In unpublished notes, we have derived expressions for $\frac{dw_1}{d\bar{\tau}_1}$, $\frac{d\tau_2}{d\bar{\tau}_1}$, $\frac{dw_2}{d\bar{\tau}_1}$ by applying comparative static analysis to (17) and (18). Even though we wrote the results in terms of dimensionless shares and elasticities to improve comprehensibility, the results are still too complex to be helpful. Possible approaches include transformation of variables, analysis of special cases, non-local analysis, and combinations thereof.

An earlier draft of this paper (Arnott and Yan (1999)) contained a section which made the unrealistic simplifying assumption that rail capacity construction is characterized by constant returns to scale. With this assumption, (19a) reduces to $-n\frac{\partial c_2}{\partial w_2} = \theta_2^2 c_2'(\theta_2) = k_2$, where k_2 is the unit rail capacity construction cost, which implies that the second-best volume-capacity ratio of rail equals its first-best level, independent of demand conditions, w_1, $\bar{\tau}_1$ and τ_2, and depends *only* on the rail technology, specifically k_2 and the form of $c_2'(\cdot)$. Thus, under the assumption of constant unit cost of rail construction, the second-best rail volume-capacity ratio can be treated as a

[18] From the two full price equations, $\frac{dp_i}{dw_2} = \frac{-c_2'n}{w_2^2}\frac{dp_i}{d\tau_2}$; thus $\frac{dm}{dw_2} = \frac{-c_2'n}{w_2^2}\frac{dm}{d\tau_2}$ and $\frac{dn}{dw_2} = \frac{-c_2'n}{w_2^2}\frac{dn}{d\tau_2}$, so that

parameter. We showed earlier that the Wheaton-Wilson problem can be analyzed replacing road capacity by the road volume-capacity ratio. The full problem too can be rewritten with the volume-capacity ratios of road and rail replacing road and rail capacities. With this transformation of variables, the constancy of the second-best rail volume-capacity ratio results in a reduction in the dimensionality of the problem, which facilitates the analysis considerably. Even so, the general case is discouragingly complex. We did however succeed in getting interpretable results for three special cases: constant demand elasticities[19], linear demand, and small deviations in the road toll away from the first-best optimum.

Other simplifications are possible too. One obvious line of attack is to analyze the problem, holding one of the policy variables fixed. For example, with the secular decline in transit ridership in the U.S., it may not be unrealistic to assume that existing rail capacity is excessive and, since it is largely irreversible, essentially fixed. Alternatively, one might assume, for the same reason, that congestion on the rail is unimportant, returning us to the world of the paradoxes discussed earlier. These simplifications ignore service frequency, however, which is potentially important.

All such analyses will be insightful, permitting the identification of operative effects and how they interact. When it comes to quantification for policy purposes, however, no local analysis can avoid the "elasticities-of-demand-elasticities" problem. Econometricians have enough difficulty estimating traffic demand elasticities with any precision, particularly since those elasticities vary from place to place, depending on the income distribution, the configuration of road and public transit networks in relation to residences, etc. It is unrealistic to suppose that in the near future we will be able to obtain useful estimates of elasticities of demand elasticities. But ascertaining whether an *incremental* change in road capacity, rail fare, or rail capacity is beneficial requires knowing these elasticities.

It seems therefore that policy prescription will have to rest on non-local analysis and on econometric estimation of the parameters of such analysis. Suppose, for example, that it can be

$$\frac{dm}{dw_2}\bigg/\frac{dn}{dw_2}=\frac{dm}{d\tau_2}\bigg/\frac{dn}{d\tau_2}.$$

established with acceptable confidence that over the range of policy interest the *arc* elasticity of road demand lies between .4 and .6. Using non-local analysis, this information and other available parameters (for instance, related to the construction technology) may be enough to establish that a major expansion of transit capacity is desirable. As evidenced by Wilson (1983) and d'Ouville and McDonald (1990), non-local analysis will not be easy, but for theory that serves urban transportation policy analysis we judge it necessary.

Since our aim in this paper has been to elucidate the basic economics of capacity and pricing decisions in urban transportation when auto congestion is underpriced, we have been content to follow the literature in its specification of the congestion and construction technologies. We should point out, however, that for policy analysis purposes, those specifications are over-simplified in a number of respects.

First, some capacity variables do not enter the congestion cost functions only via the volume-capacity ratios. For example, improving the surface of a road or the rail-bed increases travel speed for any volume-capacity ratio. This has potentially important implications for transportation policy. Consider, for instance, the Pigou-Knight-Downs Paradox discussed in section IV.1, with congested auto travel and uncongested rail travel. The standard policy implication taken from the Paradox is that, with underpriced auto congestion, the benefits of road expansion may be almost completely neutralized by latent demand, in which case the traffic congestion problem seems well nigh intractable. But the standard specification of the Paradox ignores the possibility of reducing rail travel costs, since it assumes that increasing capacity affects only the cost of *congested* travel. However, even if rail travel is uncongested, rail travel costs can be reduced by improving the quality of the rail-bed and of the signal system, since these improvements permit an increase in speed. Such improvements are doubly beneficial. Not only do they confer direct benefits on rail travelers, but the traffic diversion from road to rail induced by the rail improvements reduces the deadweight loss from underpriced auto congestion as well.

[19] Account must of course be taken of relationships between elasticities.

Second, as was discussed in section III.1, the technology of rail is not adequately captured by a user cost function which depends only on the volume-capacity ratio and a capacity construction cost function. An alternative specification can be constructed which includes the quality of the rail-bed (which incorporates other elements of capacity affecting free-flow travel speed) and separates construction costs into capital costs and operating costs.[20]

- Policy issues

Following the literature, our analysis focused on comparing the first- and second-best allocations, where the second-best allocation is the best allocation subject to the road toll being given exogenously. Analysis of this problem is certainly useful in developing the economics of the second best in the context of transportation. But its policy relevance is open to question. If indeed the road toll is given exogenously and if the planner implements the second-best allocation, what does it matter how the second-best allocation compares to the first-best allocation?[21] What is of policy interest is solving for the optimal allocation subject to the full set of constraints on policy, and investigating how that constrained optimal allocation should be implemented.

Our analysis considered only one constraint – that the road toll is specified exogenously. But there are many other possibly relevant constraints, and a correspondingly rich set of second-best policy problems that fit the two-mode model. We mention only a few.

1. Each mode may be operated by a different transportation agency, with either or both agencies subject to a deficit constraint.

[20] Let q be the quality of the rail-bed, r the rolling stock (measured in passenger capacity), $K(\cdot)$ capital costs, and $V(\cdot)$ operating costs. Then the user cost function can be specified as $c_2(\theta_2, s, q)$, and the train authority's long-run cost as $K(q,r)+V(s,r)$. Service frequency and the train's volume-capacity ratio are in turn functions of q and r. The duration of a train run is $t(q,\theta_2)$. Then assuming that the rolling stock is fully utilized, we have that flow capacity is $\frac{r}{t(q,\theta_2)}$, so that the volume-capacity ratio is $\theta_2 = \frac{nt(q,\theta_2)}{r}$ which can be solved to give $\theta_2 = \theta_2(n,q,r)$.

[21] Perhaps this is too negative an assessment. Throughout the paper we have employed static analysis. In fact, of course, because travel demand changes over time and because transportation infrastructure is durable, capacity decisions are intertemporal. If, therefore, the transportation planner has reason to believe that road pricing will be less distorted in the future, he should choose optimal capacity taking into account this movement towards first-best pricing.

2. Each mode may be operated by a different transportation agency. Suppose that one agency employs naïve cost-benefit analysis, while the other is sophisticated. How should the sophisticated agency respond to the overcapacity in the other mode caused by that mode's agency using naïve cost-benefit analysis?
3. Extend the model to treat peak and off-peak periods, and suppose that even though congestion tolls cannot be applied to auto travel, gasoline taxes can. At what level should the gasoline tax be set?
4. Recent experience[22] suggests that congestion pricing of roads may be politically acceptable if the toll revenue raised is dedicated to improvements in mass transit. What is the optimal policy subject to this form of political constraint?
5. Advocates of urban mass transit in the U.S. argue that transit ridership has fallen due to the infrequency (as well as unreliability and poor quality) of service, and that a large expansion of capacity would remedy the problem. Quantitatively, does this argument have merit?
6. Pickrell (1992) has undertaken retrospective cost-benefit analysis of the LRT (light rail transit) systems constructed recently in several U.S. cities, and found that most have negative *gross* benefits through their effect of diverting traffic from buses thereby reducing their quality of service. Does this discouraging result still hold when account is taken of the full set of policy instruments?

V. Conclusion

In this paper we returned to a classic problem in urban transportation. How does the underpricing of car travel affect the (second-best) optimal choices of road capacity, "rail" capacity and rail pricing? This problem received considerable attention in the academic literature in the 1970's, but has since been relatively neglected. We started with a synthetic exposition of the

[22] This policy was prominent in the platform of the recently-elected mayor of the City of London. It also featured in the debate over the conversion of the underutilized high-occupancy vehicle (HOV) lanes on Interstate 15 in the northern suburbs of San Diego into HOT lanes (lanes admitting both untolled high-occupancy and tolled, low-occupancy vehicles). The opposition to these "Lexus-lanes" (so-called because only wealthy, low-occupancy vehicle

earlier literature which tended to provide separate treatments of the pricing and capacity problems, and extended it in several directions. We then investigated some aspects of an integrated treatment and discussed directions for future research taking into account modern developments in second-best theory and current policy issues. The "two-mode problem" remains the centerpiece of second-best urban transport economic theory. While its study over more than thirty years has yielded many important insights, its lode is still far from exhausted.

drivers can afford the tolls) was reduced by dedicating the toll revenues collected to improving the suburban bus service.

Appendix 1

An Illustrative Global Result

This Appendix illustrates the style of reasoning that will be needed to establish global results. Part A derives the "w^*-demand function" — the demand function for which second-best road width equals first-best road width for *all* levels of the toll. Part B demonstrates that if the actual demand function intersects the w^*-demand function from below at some $p < p^*$, then at the point of intersection $w^{**} < w^*$. Thus, a necessary condition for $w^{**} > w^*$ for *all* suboptimal levels of the toll is that the actual demand function not intersect the w^*-demand function from below for $p < p^*$.

A. *Derivation of the demand function for which $w^{**} = w^*$ for all $\bar{\tau}$ given the congestion cost and capacity cost functions.*

If $w^{**} = w^*$ for all $\bar{\tau}$, then MSB is constant for all m; that is, from (9b):

$$\frac{m(p+\varepsilon\bar{\tau})e}{p+\varepsilon e} = m^* e^* \quad \text{for all } m. \tag{A.1}$$

By assumption, we know the forms of $c\left(\frac{m}{w^*}\right)$ and $e\left(\frac{m}{w^*}\right)$. Furthermore, $\bar{\tau} = p(m) - c\left(\frac{m}{w^*}\right)$.

Thus, suppressing the dependence of $c(\cdot)$ and $e(\cdot)$ on w^* we may rewrite (A.1) as

$$m\big(p(m)+\varepsilon(m)\big(p(m)-c(m)\big)\big)e(m) = \big(p(m)+\varepsilon(m)e(m)\big)m^* e^*$$

or

$$p(m)\left(me(m)-m^*e^*\right) - \varepsilon(m)\left(c\left(\frac{m}{w^*}\right)e(m)m + e(m)m^*e^*\right) + \varepsilon(m)p(m)me(m) = 0. \tag{A.2}$$

Now $\varepsilon(m) = -\frac{p(m)}{p'(m)m}$. Hence, suppressing the arguments of the functions, (A.2) may be rewritten as

$$-p'm\left(me - m^*e^*\right) - e\left(cm + m^*e^*\right) + pme = 0. \tag{A.3}$$

This is a first-order, non-linear differential equation in m. Imposing the boundary condition $m\left(p^*\right)=m^*$, the solution gives the unique demand function for which $w^{**}=w^*$ for all $\bar{\tau}$. We term this the w^*-demand function.

B. A necessary condition for $w^{**}>w^*$ whenever congestion is underpriced.

Any demand function which cuts the w^*-demand function from below for some $\bar{\tau}\in\left[0,\tau^*\right)$ has $w^{**}<w^*$ at the point of intersection. Let I denote the intersection point, ϕ the actual demand function, and w^* the w^*-demand function. At the intersection point, $(p,m,e)_I^\phi=(p,m,e)_I^{w^*}$. If the actual demand function cuts the w^*-demand function from below, then at the intersection point $(-p')_I^\phi<(-p')_I^{w^*}\Rightarrow(\varepsilon)_I^\phi>(\varepsilon)_I^{w^*}$. Since congestion is underpriced at I:

$$(\varepsilon)_I^\phi>(\varepsilon)_I^{w^*}\Rightarrow\left(\frac{p+\varepsilon\bar{\tau}}{p+\varepsilon e}\right)_I^\phi<\left(\frac{p+\varepsilon\bar{\tau}}{p+\varepsilon e}\right)_I^{w^*}$$

$$\Rightarrow MSB_I^\phi\text{ lies below }MSB_I^{w^*}\Rightarrow\left(w^{**}\right)_I^\phi<\left(w^{**}\right)_I^{w^*}=w^*.$$

Note that only two points of the actual demand function may be enough to establish that the function cuts the w^*-demand function from below. Thus, this global condition may be applied with very little information about the actual demand function.

Bibliography

Arnott, R., A. de Palma, and R. Lindsey, 1993,“A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand”, American Economic Review 83, 161-179.

Arnott, R., and K. Small, 1994, “The Economics of Traffic Congestion”, American Scientist, 82, 446-455.

Arnott, R., and A. Yan, 1999, “Road and Rail: Second-best Pricing and Capacity”, mimeo.

Bertrand, T., 1977, “Second-Best Congestion Taxes in Transportation Systems”, Econometrica, 45, 1703-1715.

Bovenberg, A.L. and L. Goulder, 1998, “Environmental Taxation”, in A. Auerbach and M. Feldstein, eds. Handbook of Public Economics, Second edition. New York, North-Holland

Braid, R., 1996, “Peak-load Pricing of a Transportation Route with an Unpriced Substitute”, Journal of Urban Economics, 40, 179-197.

Doi, M., 1986, Multimodal Urban Transportation Pricing Theory, Ph.D. thesis, University of Pennsylvania. University Microfilms International (UMI), Ann Arbor, MI, USA.

d’Ouville, E.L., and J.F. McDonald, 1990, “Optimal Road Capacity with a Suboptimal Congestion Toll”, Journal of Urban Economics, 28, 34-49.

Kanemoto, Y., 1996, “Kotu Toshi no Ben-eki Hyoka: Shohisha Yojo Aurochi (Benefit Evaluation of Transportation Investment – The Consumer Surplus Approach)”, A-201, Nihon Kotsu Seisaku Kenkyukai.

Kanemoto, Y., 1999, “Benefit-Cost Evaluation of Transportation Investment with Underpriced Congestion”, University of Tokyo, mimeo.

Kraus, M., and Y. Yoshida, 2000, “The Commuter's Time-of-Use Decision and Optimal Pricing and Service in Urban Mass Transit”, Boston College, mimeo.

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

Lévy-Lambert, H., 1968, "Tarification des Services à Qualité Variable -- Application aux Péages de Circulation", Econometrica, 36, 564-574.

Liu, L., and J.F. McDonald, 1998, "Efficient Congestion Tolls in the Presence of Unpriced Congestion: A Peak and Off-Peak Simulation Model", Journal of Urban Economics, 44, 352-366.

McDonald, J.F., 1995, "Urban Highway Congestion: An Analysis of Second-best Tolls", Transportation, 22, 353-369.

Marchand, M., 1968, "A Note on Optimal Tolls in an Imperfect Environment", Econometrica, 36, 575-581.

Mayeres, I., and S. Proost, 1997, "Optimal Tax and Public Investment Rules for Congestion Type of Externalities", Scandinavian Journal of Economics, 99, 261-279.

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

Mohring, H., 1979, "The Benefits of Reserved Bus Lanes, Mass Transit Subsidies, and Marginal Cost Pricing in Alleviating Traffic Congestion" in Current Issues in Urban Economics, P. Mieszkowski and M. Straszheim, eds., Maryland, MD, Johns Hopkins University Press, 165-195.

Oum, T.H., W.G. Waters II and J. S. Yong, 1992, "Concepts of Price Elasticities of Transport Demand and Recent Empirical Estimates", Journal of Transport Economics and Policy 26,139-154.

Parry, I., and A. Bento, 2000, " Tax Deductions, Environmental Policy, and the 'Double Dividend' Hypothesis", forthcoming, Scandinavian Journal of Economics

Pickrell, D., 1992, "A Desired Named Street Car", Journal of American Planning Association, 58 (2), 158-173.

Verhoef, E., P. Nijkamp and P. Rietveld, 1996, "Second-best Congestion Pricing, the Case of an Unpriced Alternative", Journal of Urban Economics, 40, 279-302.

Sherman, R. 1971, "Congestion Interdependence and Urban Transit Fares", Econometrica, 39, 565-576.

Vickrey, W.S., 1955, “A Proposal for Revising New York’s Subway Fare Structure”, Journal of the Operations Research Society of America, 3, 38-69.

Vickrey, W.S., 1969, "Congestion Theory and Transport Investment" American Economic Review, 59, 251-261.

Wheaton, W.C., 1978, “Price-induced Distortions in Urban Highway Investment”, Bell Journal of Economics, 9, 622-632.

Wilson, J.D., 1983, “Optimal Road Capacity in the Presence of Unpriced Congestion”, Journal of Urban Economics, 13, 337-357.