# A New Look at the Two-Mode Problem

by

Marvin Kraus*

December 2002

*Department of Economics, Boston College, Chestnut Hill, MA 02467, USA. E-mail: kraus@bc.edu. I am grateful to Richard Arnott for helpful discussions of the problem. An earlier version of this paper was presented at the 2002 North American meetings of the Regional Science Association International in San Juan, Puerto Rico.

## Abstract

This paper considers the second-best policy problem that arises when auto travel is priced below its marginal cost and there is a substitute mass transit mode. We analyze the problem by combining a model of a rail line based on Kraus and Yoshida (*JUE* (2002)) with the highway bottleneck model. The model involves a transit authority which optimizes, in addition to the fare, two dimensions of transit capacity. These are (1) the number of train units serving the route and (2) the capacity of an individual train unit. Under a very weak condition, second-best optimality involves expanding both dimensions of transit capacity. The larger of the effects is on train size.

# A New Look at the Two-Mode Problem

## 1. Introduction

Most urban transportation economists would agree that, at least in the United States, the policy responsible for the greatest resource waste in urban transportation is the nearly universal pricing of auto travel below its marginal cost. As a result, there has been a great deal of interest in what Arnott and Yan [1] recently termed the "two-mode problem." The problem is how to price mass transit and how to set both highway and mass transit capacity to minimize the resource waste from underpricing auto travel.

Most of the relevant literature has focused on some particular aspect of the problem. In one set of papers, highway and transit capacity are taken as exogenous, and the focus is on second-best transit pricing. Including work on the corresponding two-road problem, papers in this set include the early contributions by Lévy-Lambert [7], Marchand [10] and Sherman [13], and the more recent contributions by Braid [2], de Palma and Lindsey [3], Liu and McDonald [8]-[9], Small and Yan [15], Verhoef, Nijkamp and Rietveld [16] and Verhoef and Small [17]. A second set of papers has employed models with no transit alternative to auto travel to analyze the relationship between first- and second-best levels of highway capacity. The most notable of these papers are Wheaton [18], Wilson [19] and d'Ouville and McDonald [4]. The only analyses of the two-mode problem in which transit capacity is optimized in the first- and second-best solutions are those by Henderson [5] and Arnott and Yan [1], and neither study looks at the relationship between first- and second-best levels of transit capacity.[1] The main focus of the present paper is on the relationship between the first- and second-best levels of transit capacity.

The model we construct combines a bottleneck model of a highway with a model of a rail line based on Kraus and Yoshida [6]. The latter is a counterpart for mass transit to the highway bottleneck model, and therefore couples nicely with a bottleneck specification for the highway. One reason for this choice is to have a time-of-use decision for each mode. The other is that the highway bottleneck model and its mass transit counterpart have fewer nonlinearities than traditional

models.

In many respects the model is similar to Braid's [2] two-road model. The similarities arise from Braid's use of the highway bottleneck model for both roads in analyzing second-best pricing of a road when a substitute road is unpriced. There are also important differences. In Braid's model, each bottleneck has an exogenous capacity which is available on an uninterrupted basis.[2] In the present paper, transit capacity is not only optimized, but is provided on an intermittent basis.

As a policy model, our model is quite rich, with the transit authority optimizing (in addition to the fare) both the number of train units serving the route and the capacity of an individual train unit. At the same time, we treat highway capacity as exogenous, since there is otherwise little that can be done analytically. We do not view this as particularly problematic, since in practical situations the level of highway capacity is not always a margin of adjustment – because of environmental concerns, expanding the level of highway capacity may not be an option.

Under a very weak condition, we show that second-best optimality is achieved with an expansion in both dimensions of transit capacity. Interestingly, the more pronounced effect is on train size. These are local results which do not necessarily hold away from the first-best optimum. We have encountered the same difficulty in establishing global results as others who have worked on the problem.

The next two sections present the model and the analysis, respectively. A final section concludes.

## 2. The Model

Consider two points, A and B, where individuals, assumed to be identical, live and work, respectively. A and B are connected by a highway which is used exclusively by auto commuters and by a rail line. Individuals have the same work start time $t^*$ and must be at work by $t^*$ (arriving late is prohibitively costly). An individual who arrives at work at time $t' \leq t^*$ incurs a time early cost of $\beta(t^* - t')$, where $\beta > 0$ is a given "schedule delay cost" parameter. $N_1$ and $N_2$ are the number of auto and mass transit commuters, respectively. $N_1$ and $N_2$ are endogenous, along

with their sum $N \equiv N_1 + N_2$. However, in the next two subsections, which present the highway and mass transit submodels, $N_1$ and $N_2$ are exogenous.

2.1 *Highway Submodel*

The highway is uncongested, except at a single bottleneck. The bottleneck's capacity – the maximum rate at which cars can pass through the bottleneck per hour – is fixed at *s*. If arrivals at the bottleneck ever occur at a rate exceeding *s*, then a queue forms.

For simplicity, an auto commuter is assumed to have no travel costs other than queuing time costs at the bottleneck. This means that an auto commuter's departure time from home is his arrival time at the bottleneck, and his arrival time at work is his departure time from the bottleneck.

Let $D(t)$ denote the number of cars in the queue at time *t*, and let $q(t)$ denote the queuing time experienced by an auto commuter who leaves home at time *t*. Queuing time is related to queue length by

$$q(t) = D(t)/s. \tag{1}$$

For an auto commuter whose departure time from home is *t*, queuing time cost is assumed to be $\alpha q(t)$, where $\alpha$ is a given queuing cost parameter. The individual's schedule delay cost is $\beta(t^* - (t + q(t)))$, since his arrival time at work is $t' = t + q(t)$. Denoting the total time cost by $c(t)$,

$$c(t) = \alpha q(t) + \beta(t^* - (t + q(t))). \tag{2}$$

Note that an increase in $q(t)$ increases $c(t)$ only if $\alpha > \beta$, which we assume to be the case.[3]

Given the number of auto commuters, optimal temporal utilization of the highway requires that the sum of their schedule delay and queuing costs is at a minimum. This is achieved when auto commuters depart (departures will always refer to home) at a uniform rate of *s* over a time interval from

$$t_o \equiv t^* - N_1/s \tag{3}$$

to $t^*$. To see that this pattern is optimal, simply note that a queue never forms, and schedule delay costs cannot be made lower.

We next consider decentralization of the optimal departure pattern. This requires an appropriate time-varying toll. Denote the toll at departure time $t$ by $\tau(t)$, giving a (full) trip price at $t$ of $c(t)+\tau(t)$. Taking $q(\cdot)$ as given, individuals choose departure times to minimize trip price. In equilibrium, trip prices are equal at all departure times that are used, and are no less at departure times that are not used. The common trip price at chosen departure times is the mode 1 *equilibrium trip price* and is denoted by $P_1$.

It follows that for any value of the parameter $\varphi_1$, the toll function defined by

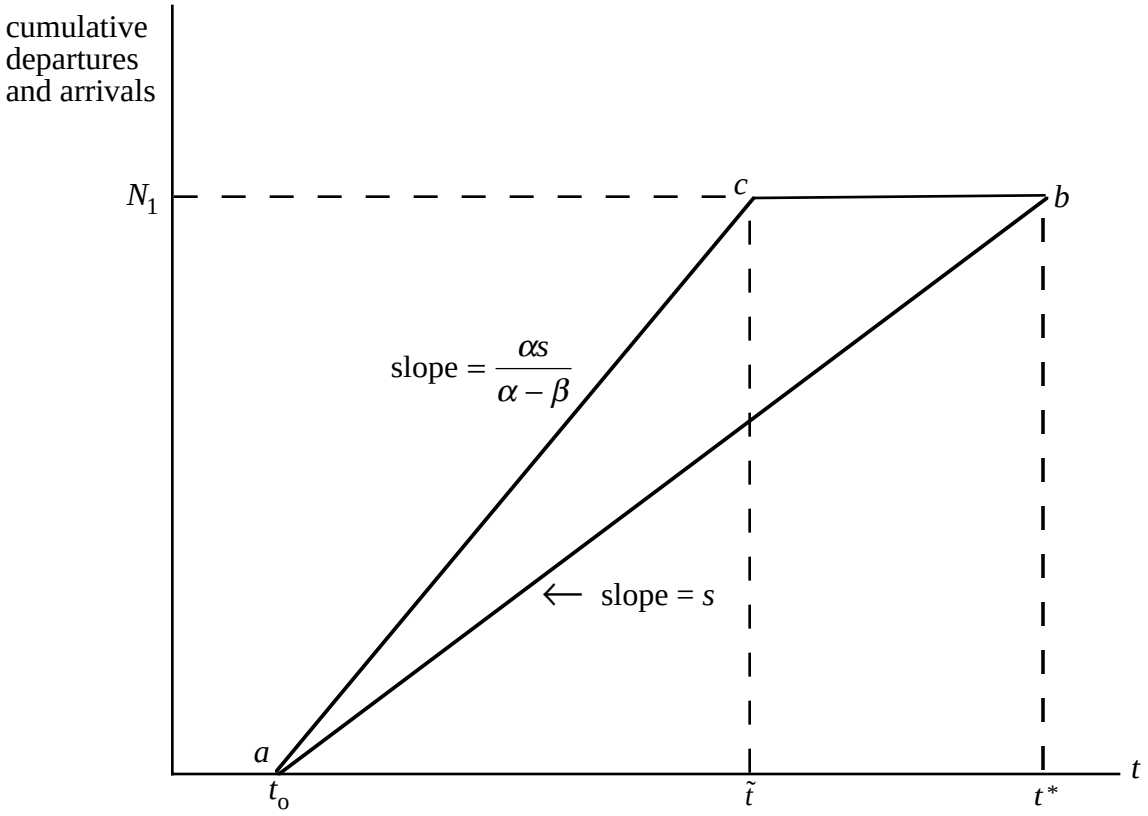
$$\begin{aligned} \tau(t) &= \varphi_1 \qquad t \le t_o \\ &= \varphi_1 + \beta(t - t_o) \qquad t \in [t_o, t^*] \end{aligned} \tag{4}$$

results in decentralization of the optimal departure pattern. With (4), the toll increases over the optimal departure interval at a rate of $\beta$, exactly offsetting the decrease in schedule delay cost. As for $\varphi_1$, there is nothing in the model yet to pin down its value. Once the specification of the model is complete, we will see that the first-best value of $\varphi_1$ is zero. Finally, we note that since auto commuters have an average schedule delay under (4) of $N_1/2s$, their aggregate schedule costs are $\beta N_1^2/2s$.

Next, let us consider the no-toll case of $\tau(t)=0$ for all $t \le t^*$. The fact that the toll is uniform in this case means that the equal-trip-price condition can be met only if queuing occurs in equilibrium. Since a commuter who departs later incurs a lower schedule delay cost, the queue he faces will have to be longer. The equilibrium is depicted in Figure 1, which shows cumulative departures from home, *acb*, and cumulative arrivals at work, *ab*. Departures begin at the same time $t_o$ as under the optimal departure pattern. The last departure occurs at $\tilde{t}$, where the height of the cumulative departures schedule reaches $N_1$. After $\tilde{t}$, there are arrivals, but no departures.

When the departure pattern is optimal, *ab* gives both cumulative arrivals and cumulative departures. Since a move to the present no-toll case does not change arrivals, it does not change aggregate schedule delay costs. However, queuing costs are introduced which in the aggregate are the same as schedule delay costs. As a result, the aggregate time costs of auto commuters in the

Figure 1


cumulative
departures
and arrivals
$N_1$
$c$
$b$
slope = $\frac{\alpha s}{\alpha - \beta}$
slope = $s$
$a$
$t_o$
$\tilde{t}$
$t^*$
$t$

no-toll case are $\beta N_1^2/s$.

We are now ready to specify what constitutes underpricing the highway. For this purpose, we introduce a parameter $\lambda \in [0, 1]$ and consider the following generalization of the toll function in (4):

$$\begin{aligned} \tau(t) &= \varphi_1 \qquad t \le t_o \\ &= \varphi_1 + (1-\lambda)\beta(t - t_o) \qquad t \in [t_o, t^*]. \end{aligned} \tag{5}$$

Because of its generality, (5) can be used to set up both the first- and second-best problems. In the case of the first-best problem, we set $\lambda = 0$ (in which case (5) reduces to (4)) and optimize $\varphi_1$. The optimal value of $\varphi_1$ in this case will be shown to be zero. In the case of the second-best problem, $\varphi_1$ is constrained to be zero, and $\lambda$ is assumed to satisfy $0 < \lambda \le 1$. For any such value of $\lambda$, the fact that $\varphi_1 = 0$ means that the highway is less than fully priced at all times in the first-best departure interval. The greater the value of $\lambda$, the greater the extent to which the highway is underpriced. Thus $\lambda$ serves as an ordinal index of the underpricing of the highway. Setting $\lambda = 1$ generates the no-toll case.

Any positive value of $\lambda$ results in a cumulative departures schedule steeper than *ab* in Figure 1 and therefore a certain amount of queuing. Appendix 1 gives the details of this and shows that in the equilibrium departure pattern under (5), the aggregate time costs of auto commuters are given by

$$C(N_1, s, \lambda) \equiv \Gamma(\lambda)\beta N_1^2/2s, \tag{6}$$

where

$$\Gamma(\lambda) \equiv \frac{\alpha - \beta + \lambda(\alpha + \beta)}{\alpha - \beta + \lambda\beta}. \tag{7}$$

It is easily checked that $\Gamma(\cdot)$ is a monotonically increasing function, and that $\Gamma(0) = 1$, while $\Gamma(1) = 2$. Thus, for $\lambda = 0$, (6) reduces to the cost expression $\beta N_1^2/2s$ that we derived for the toll function (4), while for $\lambda = 1$, it reduces to the cost expression we derived for the no-toll equilibrium.

The final piece of business for the highway submodel is to determine the mode 1 equilibrium trip price under (5). The easiest way to do so is as the price of a mode 1 trip at $t_o$. The toll at $t_o$ is $\varphi_1$, and since there is no queue at $t_o$, $c(t_o) = \beta N_1 / s$. Thus,

$$P_1 = \beta N_1 / s + \varphi_1. \tag{8}$$

2.2 *Mass Transit Submodel*

For simplicity, a transit commuter is assumed to have no travel costs except possibly for a waiting time cost at the origin train stop. We ignore time spent walking to and from train stops, as well as passenger transit and loading time.

The number of train departures from A will be referred to as the number of runs and denoted by *R*. Initially, *R* is taken to be integer-valued. Each run has a strict capacity of $\sigma$ passengers. *R* and $\sigma$ are related by

$$R \geq N_2 / \sigma. \tag{9}$$

Without loss of generality, *R* is taken to be the smallest integer satisfying (9).

The number of physically distinct trains used to make the runs, which we refer to as the number of train units, is denoted by *K*. Like *R*, *K* is initially integer-valued. If each train unit is used to make a single run, then $K = R$. If one or more train units makes multiple runs, then $K < R$.

The time it takes a train to make the roundtrip from A to B is *T* minutes. Thus, the successive runs of a train unit must be scheduled at least *T* minutes apart. The other constraint we impose on scheduling is that successive train departures from A are at least $\delta$ minutes apart. $\delta$ is a positive parameter known as the safe headway.

The cost of providing transit service is specified as in Kraus and Yoshida [6]:

$$(\nu_o + \nu_1 \sigma) TR + \nu_2 \sigma K + \nu_3 \sigma + \nu_4 T, \tag{10}$$

where $\nu_1, \ldots, \nu_4$ and $\nu_o$ are positively-valued parameters. The first term in (10) is the operating costs of runs. It assumes a cost per train-minute of operation which increases linearly with the capacity of a train. The second term in (10), which we will refer to as fleet costs, gives the

nonoperating capital costs for the transit authority's fleet of cars. It implicitly assumes that the cost of a car is proportional to its capacity. The third term in (10) represents capital costs for terminals (at A and B). It is based on two implicit assumptions. One is that a terminal's cost is proportional to its area. The other is that a terminal's area is proportional to the capacity of the trains that it serves. The final term in (10) represents right-of-way and construction costs for trackage. Trackage costs are assumed to be proportional to the distance between A and B and therefore to *T*.

Given $N_2$, we now consider the problem of minimizing total mode 2 costs. These are the schedule delay and waiting costs of transit commuters and the cost of providing transit service. $\sigma$ and *K* are policy variables, but for now we take them as given. *R* is implied by $N_2$ and $\sigma$, so this fixes a value for (10). With (10) fixed, the problem is to minimize the time costs of transit commuters, where the choice variables are the departure times of the runs and the pattern of commuter arrivals at the origin stop.

Given the departure times of the runs, a mass of passengers of size $\sigma$ should arrive at each departure time except the earliest (by arrivals, we mean those at the origin stop), with a possibly smaller mass of size $N_2 - (R-1)\sigma$ arriving at the earliest departure time. Under this arrival pattern, there is no waiting, and given the departure times of runs, schedule delay costs cannot be made lower.

The next step is to determine how the runs should be scheduled. Starting with an example will help, so with $\delta$ and *T* measured in minutes, suppose that $t^*$ = 9:00, $\delta$ = 2, *T* = 45, *K* = 3 and *R* = 7. Then, given that the successive runs of a train unit must be at least *T* minutes apart, the optimal train departure times, from latest to earliest, are: 9:00, 8:58, 8:56, 8:15, 8:13, 8:11, 7:30. The idea, then, is to run trains in clusters, with trains within a cluster separated by the safe headway. In the example, there is a partial cluster in addition to two full clusters. In general, the number of full clusters is given by $R/K$ rounded *down* to the nearest integer. We write this as

$$C = (R/K)_-, \tag{11}$$

where *C* denotes the number of full clusters.

Next, change the value of *K* in the example to 23, and assume that *R* is now 47. Since

departures in the latest cluster begin at 8:16, the safe headway restriction pushes the last departure in the middle cluster back to 8:14, $K\delta$ minutes before the 9:00 departure. In the previous example, this separation was $T$ minutes. In general, the number of minutes separating the latest runs from two neighboring clusters is given by $max\{T, K\delta\}$. In what follows, we will be concerned only with the case $K\delta < T$. This is highly inclusive, since any value of $K$ greater than $(T/\delta)_+$ is clearly inefficient.

Under the preceding assumption, the solution to the scheduling problem is for trains to leave A at the following times: $t^*, t^*-\delta, ..., t^*-(K-1)\delta$ (latest cluster), $t^*-T, t^*-T-\delta, ...,$ $t^*-T-(K-1)\delta$ (next-to-latest full cluster), …, $t^*-(C-1)T, t^*-(C-1)T-\delta, ...,$ $t^*-(C-1)T-(K-1)\delta$ (earliest full cluster), $t^*-CT, t^*-CT-\delta, \ldots, t^*-CT-(R-CK-1)\delta$ (partial cluster, if applicable).

Up until now, $\sigma$ and $K$ have been taken as given. It is not difficult to show that when $\sigma$ is optimal, all runs are at capacity. This means that there is no slack in (9), or that

$$R\sigma = N_2. \tag{12}$$

The proof is similar to one that can be found in Kraus and Yoshida [6] (Lemma 1 of that paper) and is therefore omitted.

Having identified the departure times that make up the optimal schedule, and knowing that each run will carry the same number of passengers, $\sigma$ (from (12)), we can easily make an accounting of aggregate schedule delay costs (*SDC*). Although the resulting expression looks complicated, it is straightforward to show that

$$SDC = \frac{\beta N_2}{2R}\{KC[\delta(K-1)+T(C-1)]+(R-CK)[2CT+\delta(R-CK-1)]\}. \tag{13}$$

For the remainder of the paper, $R$ and $K$ are taken to be continuously-valued, and we approximate (11) by

$$C = R/K. \tag{14}$$

Without this approximation, (13) would be a discontinuous function of $R$ and $K$, rendering the problem analytically intractable. The effect of (14) is to make all clusters full, while allowing them

to be fractional in number. Using (14), (13) simplifies to

$$SDC = \frac{\beta N_2}{2}[\delta(K-1) + T(\frac{R}{K} - 1)]. \tag{15}$$

which can be shown to understate (13).

Since optimizing *R* is equivalent to optimizing $\sigma$ (as a result of (12)), we can formulate the problem

$$\min_{R,K} \quad \Phi(N_2, R, K) \tag{16}$$

where

$$\Phi(N_2, R, K) \equiv \frac{\beta N_2}{2}[\delta(K-1) + T(\frac{R}{K} - 1)] + \nu_0 TR + \nu_1 TN_2 + \frac{(\nu_2 K + \nu_3)N_2}{R} + \nu_4 T \tag{17}$$

is total mode 2 costs. The first-order conditions are

$$\Phi_R = \frac{\beta N_2 T}{2K} + \nu_0 T - \frac{(\nu_2 K + \nu_3)N_2}{R^2} = 0 \tag{18}$$

$$\Phi_K = \frac{\beta N_2}{2}(\delta - \frac{TR}{K^2}) + \frac{\nu_2 N_2}{R} = 0. \tag{19}$$

The tradeoffs involved in the first-order conditions are as follows. First, consider the effect of an increase in *K*, with *R* held fixed. This shifts arrivals closer to $t^*$, resulting in a decrease in aggregate schedule delay costs. This has to be traded off against an increase in fleet costs, since the number of capacity units in the transit authority's fleet, $\sigma K$, is now greater.

To see the tradeoff involved in (18), we now consider the effect of a *decrease* in *R*, with *K* held fixed. Since decreasing *R* increases $\sigma$, this is an alternative way of decreasing aggregate schedule delay costs. There is also a decrease in $\nu_0 TR$, the portion of operating costs which is independent of train capacity. But since $\sigma$ increases, there are increases in both fleet costs and capital costs for terminals.

It will be important to know how the solution to (16) varies with $N_2$. It turns out there is quite a bit one can say about this. Our results are stated in the following lemma, in which $E_{K:N_2}$ denotes the elasticity of *K* with respect to $N_2$, and corresponding notation is used for other

elasticities.

*Lemma* 1. The following are properties of a solution to (16):

(i) $0 < E_{K:N_2} < E_{R:N_2} < 1/2.$

(ii) $E_{\sigma:N_2} > E_{K:N_2}.$

The proof of (i) is given in Appendix 2. Here, we demonstrate how (ii) follows from (i). From (12), $E_{R:N_2} + E_{\sigma:N_2} = 1$. Since $E_{R:N_2} < 1/2$ from (i), this implies $E_{\sigma:N_2} > 1/2$. Together with $E_{K:N_2} < 1/2$, this gives the stated property.

Before leaving the mass transit submodel, we briefly consider decentalization of the optimal arrival pattern of passengers. This requires an appropriate time-varying fare. In equilibrium, trip price must be the same at all train departure times. But since a later departure time means lower schedule delay cost, the fare at such a time must be correspondingly higher.

We will also need an expression for the equilibrium mode 2 trip price, $P_2$. Letting $\varphi_2$ and $L$ respectively denote the fare for the earliest run and the number of minutes before $t^*$ that the earliest run is scheduled, we can write $P_2$ as

$$P_2 = \varphi_2 + \beta L. \tag{20}$$

With optimal scheduling,

$$L = \delta(K-1) + T(C-1) = \delta(K-1) + T(\frac{R}{K} - 1), \tag{21}$$

and (20) becomes

$$P_2 = \varphi_2 + \beta[\delta(K-1) + T(\frac{R}{K} - 1)]. \tag{22}$$

2.3 *Demand and Overall Equilibrium*

We model demand in the simplest possible way, taking modal trip demands to be those of a representative consumer. We also assume that trip demands are independent of income. Under this assumption, ordinary demand functions are identical to compensated demand functions and can be written

$$N_1 = N_1(P_1, P_2) \tag{23}$$

$$N_2 = N_2(P_1, P_2). \tag{24}$$

Inverse demand functions take the form

$$P_1 = P_1(N_1, N_2) \tag{25}$$

$$P_2 = P_2(N_1, N_2). \tag{26}$$

The fact that (23)-(24) gives compensated demands means that its price derivatives give own- and cross-substitution effects. We therefore employ the notation $\partial N_i / \partial P_j = s_{ij}$ for all $i, j = 1, 2$. We assume that (23) and (24) are the demand functions of a utility-maximizing consumer, so that

$$s_{11} < 0, \; s_{22} < 0 \tag{27}$$

$$s_{11}s_{22} - s_{12}s_{21} > 0 \tag{28}$$

in addition to $s_{12} = s_{21}$. We also assume that mode 1 and mode 2 trips are substitutes, so that $s_{12} > 0$.

Given a particular policy (values for $\varphi_1, \varphi_2, R$ and $K$), a solution to the model for $N_1, N_2, P_1$ and $P_2$ can be obtained by solving the four-equation system consisting of (23)-(24) and the two supply relationships (8) and (22). A solution takes the form

$$N_i = N_i(\varphi_1, \varphi_2, R, K); \quad i = 1, 2 \tag{29}$$

$$P_i = P_i(\varphi_1, \varphi_2, R, K); \quad i = 1, 2. \tag{30}$$

Note that the system consisting of (8) and (22)-(24) does not involve $\lambda$. $\lambda$ does not have a direct effect on equilibrium trip prices and quantities. It affects them only indirectly by affecting the optimal values of $\varphi_2, R$ and $K$.

In Section 3, expressions will be needed for the various derivatives of (29). It is straightforward to derive these by varying policy variables one at a time in the four-equation system. The results are presented in Table 1.

## 3. First- and Second-Best Problems

The problem we consider is social surplus maximization. Benefits are given by the line

integral

$$B(N_1, N_2) \equiv \int_{(0,0)}^{(N_1,N_2)} P_1(n_1, n_2)dn_1 + P_2(n_1, n_2)dn_2, \tag{31}$$

while costs are given by (6) for auto trips, and by (17) for transit. It is well-known (see, e.g., Pressman [12]) that, under our assumption that trip demands are independent of income, (31) is not only path-independent, but also has the property that

$$\frac{\partial B(N_1, N_2)}{\partial N_i} = P_i(N_1, N_2); \quad i = 1, 2. \tag{32}$$

Denoting $N_i(\varphi_1, \varphi_2, R, K)$ in (29) by $N_i(\cdot)$, we write social surplus as

$$B(N_1(\cdot), N_2(\cdot)) - C(N_1(\cdot), s, \lambda) - \Phi(N_2(\cdot), R, K). \tag{33}$$

(33) is maximized in both the first- and second-best problems, albeit for different $\lambda$ values.

3.1 *First-Best Problem*

In the first-best problem, (33) is maximized with $\lambda = 0$. $\varphi_1$ is unconstrained and is optimized along with $\varphi_2, R$ and $K$. The first-order conditions, after making use of (32) and rearranging terms, are:

$\varphi_1$ and $\varphi_2$:
$$\begin{pmatrix} \frac{\partial N_1}{\partial \varphi_1} & \frac{\partial N_2}{\partial \varphi_1} \\ \frac{\partial N_1}{\partial \varphi_2} & \frac{\partial N_2}{\partial \varphi_2} \end{pmatrix} \begin{pmatrix} P_1 - \frac{\partial C}{\partial N_1} \\ P_2 - \frac{\partial \Phi}{\partial N_2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{34}$$

$R$:
$$\frac{\partial \Phi}{\partial R} = (P_1 - \frac{\partial C}{\partial N_1})\frac{\partial N_1}{\partial R} + (P_2 - \frac{\partial \Phi}{\partial N_2})\frac{\partial N_2}{\partial R} \tag{35}$$

$K$:
$$\frac{\partial \Phi}{\partial K} = (P_1 - \frac{\partial C}{\partial N_1})\frac{\partial N_1}{\partial K} + (P_2 - \frac{\partial \Phi}{\partial N_2})\frac{\partial N_2}{\partial K} \tag{36}$$

where derivatives of $C$ are evaluated at $\lambda = 0$. The $2 \times 2$ matrix in (34) has a determinant which (using the results in the first two rows of Table 1) simplifies to

$$\frac{s(s_{11}s_{22} - s_{12}s_{21})}{s - \beta s_{11}} > 0.$$

Table 1. Derivatives of Equation (29)[a]

| | $N_1$ | $N_2$ |
|---|---|---|
| $\varphi_1$ | $\frac{\partial N_1}{\partial \varphi_1} = \frac{s_{11}s}{s-\beta s_{11}} < 0$ | $\frac{\partial N_2}{\partial \varphi_1} = \frac{s_{21}s}{s-\beta s_{11}} > 0$ |
| $\varphi_2$ | $\frac{\partial N_1}{\partial \varphi_2} = \frac{s_{12}s}{s-\beta s_{11}} > 0$ | $\frac{\partial N_2}{\partial \varphi_2} = \frac{s_{22}s-\beta(s_{11}s_{22}-s_{12}s_{21})}{s-\beta s_{11}} < 0$ |
| $R$ | $\frac{\partial N_1}{\partial R} = \frac{\beta T s_{12}s}{K(s-\beta s_{11})} > 0$ | $\frac{\partial N_2}{\partial R} = \frac{\beta T s_{22}s-\beta^2 T(s_{11}s_{22}-s_{12}s_{21})}{K(s-\beta s_{11})} < 0$ |
| $K$ | $\frac{\partial N_1}{\partial K} = \frac{\beta s_{12}s(\delta K^2-TR)}{K^2(s-\beta s_{11})} < 0$ | $\frac{\partial N_2}{\partial K} = \frac{(\delta K^2-TR)(\beta s_{22}s-\beta^2(s_{11}s_{22}-s_{12}s_{21})}{K^2(s-\beta s_{11})} > 0$ |

[a] In the last row of the table, $\delta K^2 - TR < 0$, since $K\delta < T$ and $K \leq R$.

Since this is nonzero, (34) implies

$$P_1 = \frac{\partial C}{\partial N_1} \tag{37}$$

$$P_2 = \frac{\partial \Phi}{\partial N_2}, \tag{38}$$

which are the marginal cost pricing conditions. When these conditions are satisfied, (35) and (36) become

$$\frac{\partial \Phi}{\partial R} = \frac{\partial \Phi}{\partial K} = 0, \tag{39}$$

which are the first-order conditions for the transit cost minimization problem (16).

From (37), it is easy to derive $\varphi_1 = 0$. For $\lambda = 0$, $\Gamma(\lambda) = 1$. Using this in (6) gives

$$\frac{\partial C}{\partial N_1} = \beta N_1 / s. \tag{40}$$

Equating this to (8) gives $\varphi_1 = 0$.

The easiest way to see this result is to look at the congestion externality imposed by an auto commuter. If a marginal auto commuter is added at some time *t* in the optimal departure interval, some other auto commuter must be relocated from *t* to the beginning of the departure interval at $t_o$. The congestion externality imposed by the marginal auto commuter is the increase in schedule delay cost for the relocated individual, which is $\beta(t - t_o)$. $\varphi_1$ is the toll at the beginning of the departure interval. Since there is no congestion externality if the marginal auto commuter is added at $t_o$, equality between the toll and the congestion externality at time $t_o$ requires that $\varphi_1 = 0$.

In the case of mode 2, we have that $\varphi_2 > 0$. This can be demonstrated as follows. From (17),

$$\frac{\partial \Phi}{\partial N_2} = \frac{\beta}{2}[\delta(K-1) + T(\frac{R}{K} - 1)] + \nu_1 T + \frac{\nu_2 K + \nu_3}{R}. \tag{41}$$

Using this along with (22) in (38) and solving for $\varphi_2$ gives

$$\varphi_2 = -\frac{\beta}{2}[\delta(K-1) + T(\frac{R}{K} - 1)] + \nu_1 T + \frac{\nu_2 K + \nu_3}{R}. \tag{42}$$

From (18),

$$\frac{\nu_2 K + \nu_3}{R} = \frac{\beta TR}{2K} + \frac{\nu_o TR}{N_2}.$$

Using this in (42) gives

$$\varphi_2 = -\frac{\beta\delta(K-1)}{2} + \frac{\beta T}{2} + \nu_1 T + \frac{\nu_o TR}{N_2}$$

$$= \frac{\beta(T - K\delta)}{2} + \frac{\beta\delta}{2} + \nu_1 T + \frac{\nu_o TR}{N_2}.$$

Recalling that $K\delta < T,$ we have

$$\varphi_2 > \frac{\beta\delta}{2} + \nu_1 T + \frac{\nu_o TR}{N_2} > 0, \tag{43}$$

which is the result we wished to demonstrate.

Why is it that $\varphi_2 > 0,$ while $\varphi_1 = 0?$ There are two differences between mass transit and highway capacity that account for this. One is that highway capacity is available continuously with respect to time, while mass transit capacity is provided at a set of discretely spaced times. Thus, a transit passenger who is relocated from the earliest departure time to a new earlier time has her schedule delay cost discretely increased. The second is that there is unutilized highway capacity prior to $t_o,$ and its utilization entails no cost to the highway authority. Similarly, of the $\sigma K$ capacity units in the transit authority's fleet, none are utilized prior to the earliest departure. But, in order for passengers to utilize this capacity, the transit authority must incur higher operating costs for runs.

### 3.2 *Second-Best Problem*

As in the first-best problem, the objective is to maximize (33). The difference is that $\lambda$ is now positive and there is a constraint that $\varphi_1 = 0.$ As we explained in connection with (5), this results in the highway being less than fully priced.

For any value of the parameter $\lambda$ ($\lambda$ now satisfies $0 < \lambda \leq 1$), the first-order conditions are the second condition in (34), which we now write

$$P_2 - \frac{\partial \Phi}{\partial N_2} = -(P_1 - \frac{\partial C}{\partial N_1})\frac{\partial N_1}{\partial \varphi_2} \Big/ \frac{\partial N_2}{\partial \varphi_2}, \tag{44}$$

and (35)-(36). Derivatives of $C$ are evaluated for the specified value of $\lambda$, which determines the degree to which the highway is underpriced.

In (44),

$$P_1 - \frac{\partial C}{\partial N_1} < 0. \tag{45}$$

This follows from (8) and

$$\frac{\partial C}{\partial N_1} = \Gamma(\lambda)\beta N_1 / s, \tag{46}$$

since $\varphi_1 = 0$ and $\Gamma(\lambda) > 1$ for $\lambda > 0$. Also in (44), $\frac{\partial N_1}{\partial \varphi_2}$ and $\frac{\partial N_2}{\partial \varphi_2}$ are opposite in sign. From Table 1, one can see that this is a result of the assumption that mode 1 and mode 2 trips are substitutes $(s_{12} > 0)$. Using this in (44) along with (45) gives

$$P_2 - \frac{\partial \Phi}{\partial N_2} < 0 \tag{47}$$

– efficient pricing of mode 2 requires that it, too, be priced below marginal cost.

(47) is a well-known feature of the second-best solution. Its earliest derivation (Lévy-Lambert [7], Marchand [10]) assumed that mode 1 and mode 2 trips are perfect substitutes. Sherman [13] was the first of a number of authors to derive the result for the case of imperfect substitutes.

Despite the fact that pricing is no longer first-best, (35) and (36) again reduce to (39). First consider (35). Substituting (44) gives

$$\frac{\partial \Phi}{\partial R} = \left( P_1 - \frac{\partial C}{\partial N_1} \right) \left( \frac{\frac{\partial N_1}{\partial R}\frac{\partial N_2}{\partial \varphi_2} - \frac{\partial N_2}{\partial R}\frac{\partial N_1}{\partial \varphi_2}}{\frac{\partial N_2}{\partial \varphi_2}} \right).$$

The expressions in the middle two rows of the table imply that $\frac{\partial N_1}{\partial R}\frac{\partial N_2}{\partial \varphi_2} - \frac{\partial N_2}{\partial R}\frac{\partial N_1}{\partial \varphi_2} = 0$, giving $\frac{\partial \Phi}{\partial R} = 0$. Making similar use of (44) and (36) gives $\frac{\partial \Phi}{\partial K} = 0$.

The preceding result means that there is no distortion away from cost minimization in transit. The same result was obtained in previous analyses of the problem by Henderson [5] and Arnott and Yan [1] and depends crucially on the two modes having noninterdependent costs.

We are now ready to work towards our main result, which appears below as Corollary 1. The result has to do with the effect on transit policy of an infinitesimal increase in $\lambda$, coming off of an initial value for $\lambda$ of zero. Throughout what follows, it is to be understood that for any positive value of $\lambda$, there is a constraint $\varphi_1 = 0$.

We begin by establishing the following lemma:

*Lemma* 2. Given an infinitesimal increase in $\lambda$, the effect on equilibrium trip prices and quantities is either

$$N_1\downarrow,\ \ N_2\uparrow,\ \ P_1\downarrow,\ \ P_2\downarrow \tag{48}$$

or

$$N_1\uparrow,\ \ N_2\downarrow,\ \ P_1\uparrow,\ \ P_2\uparrow. \tag{49}$$

*Proof.* See Appendix 3.

*Remark.* The effects in (48) and (49) are the indirect effects we mentioned earlier, coming about through the effect of $\lambda$ on $\varphi_2, R$ and $K$. The reason why price-quantity movements can only combine as in (48) or (49) is that mode 1 and mode 2 trips are substitutes.

When the initial value of $\lambda$ is zero, a very weak assumption serves to eliminate (49). In the following proposition, $E_{N_2:P_2}$ denotes the own-price elasticity of demand for mode 2 trips.

*Proposition* 1. Suppose that $\lambda$ has an initial value of zero. Given an infinitesimal increase in $\lambda$, if $\left|E_{N_2:P_2}\right| \le 2$ in the first-best optimum, then the effect on equilibrium prices and quantities is given by (48).

*Proof.* For $\lambda = 0$,

$$\frac{d}{d\lambda}\left(P_2 - \frac{\partial \Phi}{\partial N_2}\right) < 0, \tag{50}$$

since as $\lambda$ increases, $P_2 - \dfrac{\partial \Phi}{\partial N_2}$ moves into negative values ((47)) from an initial value of zero ((38)). From (50),

$$0 > \frac{dP_2}{d\lambda} - \frac{d}{d\lambda}\frac{\partial \Phi}{\partial N_2}$$

$$= \frac{dP_2}{d\lambda} - \frac{\partial^2 \Phi}{\partial R \partial N_2}\cdot\frac{dR}{d\lambda} - \frac{\partial^2 \Phi}{\partial K \partial N_2}\cdot\frac{dK}{d\lambda} \qquad \left(\text{since } \frac{\partial \Phi}{\partial N_2} \text{ is independent of } N_2\right)$$

$$= \frac{dP_2}{d\lambda} + \frac{\nu_o T}{N_2}\cdot\frac{dR}{d\lambda} \qquad \left(\text{where } \frac{\partial^2 \Phi}{\partial K \partial N_2} = 0 \text{ from (19), and } \frac{\partial^2 \Phi}{\partial R \partial N_2} = -\frac{\nu_o T}{N_2} \text{ from (18)}\right)$$

$$= \frac{dP_2}{d\lambda} + \frac{\nu_o T}{N_2}\cdot\frac{dR}{dN_2}\cdot\frac{dN_2}{d\lambda}$$

$$= \frac{dP_2}{d\lambda}\left(1 + \frac{\nu_o T}{N_2}\cdot\frac{dR}{dN_2}\cdot\frac{\partial N_2}{\partial \varphi_2}\right) \qquad \text{(from (A14) in Appendix 3)} \qquad (51)$$

(49) is the case in which $\dfrac{dP_2}{d\lambda} > 0$. From (51), this can occur only if

$$-\frac{\nu_o T}{N_2}\cdot\frac{dR}{dN_2}\cdot\frac{\partial N_2}{\partial \varphi_2} > 1.$$

This is equivalent to

$$\left|E_{N_2:\varphi_2}\right| E_{R:N_2} > \frac{\varphi_2 N_2}{\nu_o T R}.$$

Since $\varphi_2$ must initially satisfy (43),

$$\left|E_{N_2:\varphi_2}\right| E_{R:N_2} > 1 + \frac{N_2}{\nu_o T R}\left(\frac{\beta\delta}{2} + \nu_1 T\right)$$

so, in particular,

$$\left|E_{N_2:\varphi_2}\right| E_{R:N_2} > 1.$$

From Lemma 1 of the mass transit submodel, $E_{R:N_2} < 1/2$. We therefore obtain

$$\left|E_{N_2:\varphi_2}\right| > 2.$$

Since $\varphi_2$ makes up just part of $P_2$,

$$\left|E_{N_2:P_2}\right| > \left|E_{N_2:\varphi_2}\right|.$$

Thus (49) could occur only if $\left|E_{N_2:P_2}\right|$ exceeds 2. Q.E.D.

We can now state the main result of the paper.

*Corollary* 1. Under the same elasticity condition as in Proposition 1, the following policy effects occur as $\lambda$ moves into positive values from an initial value of zero:

$$R\uparrow,\ K\uparrow,\ \sigma\uparrow,\ \varphi_2\downarrow.$$

Moreover, $\sigma$ should go through a larger percentage increase than *K*.

*Proof*. We know that $N_2\uparrow$ and that (16) is solved in both the first- and second-best problems. Together with Lemma 1 of the mass transit submodel, this gives the results for *R*, *K* and $\sigma$. There is also an increase in *L*, since in (21), *R* goes through a larger percentage increase than *K*. But $P_2\downarrow$, so from (20), there must be a decrease in $\varphi_2$. Q.E.D.

*Remark*. $\left|E_{N_2:P_2}\right|$ is generally believed to be well below one.

## 4. Conclusion

This paper has considered the second-best policy problem that arises when auto travel is underpriced and there is a substitute mass transit mode. We analyzed the problem by combining a model of a rail line based on Kraus and Yoshida [6] with the highway bottleneck model. The model involves a transit authority which optimizes, in addition to the fare, two dimensions of transit capacity. These are (1) the number of train units serving the route and (2) the capacity of an individual train unit. Under a very weak condition, second-best optimality involves expanding both dimensions of transit capacity. The larger of the effects is on train size.

## Appendix 1

Over the equilibrium departure interval,

$$\dot{c}(t) + \dot{\tau}(t) = 0.$$

Together with (2) and (5), this gives

$$\alpha\dot{q}(t) - \beta(1 + \dot{q}(t)) + (1 - \lambda)\beta = 0,$$

implying that

$$\dot{q}(t) = \frac{\lambda\beta}{\alpha - \beta}. \tag{A1}$$

Together with (1), (A1) gives

$$\dot{D}(t) = \frac{\lambda\beta s}{\alpha - \beta}, \tag{A2}$$

which indicates that the rate at which the queue builds up over the equilibrium departure interval is positively related to $\lambda$. Over the equilibrium departure interval,

$$\dot{D}(t) = r(t) - s, \tag{A3}$$

where $r(t)$ is the instantaneous departure rate at time $t$. Using (A2) in (A3) gives

$$r(t) = \frac{(\alpha - \beta + \lambda\beta)s}{\alpha - \beta}, \tag{A4}$$

which will be useful in deriving (6).

To derive (6), we use the fact that the aggregate time costs of auto commuters can be obtained by subtracting aggregate toll receipts from $P_1N_1$. Making use of (8),

$$P_1N_1 = \beta N_1^2/s + \varphi_1 N_1. \tag{A5}$$

The simplest way to derive aggregate toll receipts is to use the fact that $r(t)$ is the same throughout the equilibrium departure interval and that $\tau(\cdot)$ increases linearly. This means that aggregate toll receipts can be expressed as the product of $N_1$ and the toll at the midpoint of the equilibrium departure interval:

$$\textit{Toll Receipts} = N_1\tau(t_m), \tag{A6}$$

where $t_m$ denotes the midpoint of the equilibrium departure interval. The length of the equilibrium departure interval is implied by requiring that its product with the right-hand side of (A4) is equal to $N_1$. Thus the length of the equilibrium departure interval is

$$\frac{(\alpha - \beta)N_1}{(\alpha - \beta + \lambda\beta)s}. \tag{A7}$$

From (5) and the fact that $t_m - t_o$ is half of (A7),

$$\tau(t_m) = \varphi_1 + \frac{(1-\lambda)\beta(\alpha - \beta)N_1}{2(\alpha - \beta + \lambda\beta)s}. \tag{A8}$$

Equation (6) is obtained by substituting (A8) into (A6), subtracting the resulting expression from (A5), and simplifying.

## Appendix 2

*Proof of property* (i) *of Lemma* 1. We begin by showing that any solution to (18)-(19) must also solve

$$K^2 = \frac{\beta T R^2}{\beta\delta R + 2\nu_2} \tag{A9}$$

and

$$4(\nu_3 - \nu_o T(R^2/N_2))^2 = \frac{\beta^3\delta^2 T R^4}{\beta\delta R + 2\nu_2}. \tag{A10}$$

(A9) is obtained simply by solving (19) for $K^2$ in terms of *R*. To derive (A10), rewrite (19)

$$-\frac{\beta\delta N_2 K}{2R} + \frac{\beta N_2 T}{2K} - \frac{\nu_2 N_2 K}{R^2} = 0 \tag{A11}$$

and equate the left-hand side of this equation to the left-hand side of (18). This gives

$$K = \frac{2(\nu_3 N_2 - \nu_o T R^2)}{\beta\delta N_2 R}. \tag{A12}$$

Squaring (A12) and equating to (A9) gives (A10).

Note that (A9)-(A10) is a recursive system in which (A10) alone can be solved for *R*. From (A12), a necessary condition for $K > 0$ is

$$R^2 < \frac{\nu_3 N_2}{\nu_o T}. \tag{A13}$$

(A10) always has exactly one positive root satisfying (A13), and this corresponds to a local minimum of (16). It is straightforward to use (A10) to show that $dR/dN_2 > 0$. Thus, as $N_2$ increases, so does the right-hand side of (A10). The left-hand side of (A10) must therefore also increase. The only way this can happen is if $R^2/N_2$ decreases, which is equivalent to $E_{R:N_2} < 1/2$. From (A9), the only effect of $N_2$ on *K* comes about through its effect on *R*. The appearance of *R* in the denominator of (A9) implies that $E_{K:N_2} < E_{R:N_2}$. Q.E.D.

## Appendix 3

*Proof of Lemma* 2. To prove Lemma 2, we first establish the relationships

$$\frac{dN_i}{d\lambda} = \frac{\partial N_i}{\partial \varphi_2} \cdot \frac{dP_2}{d\lambda}; \quad i = 1, 2 \tag{A14}$$

and

$$\frac{dP_1}{d\lambda} = \frac{\partial P_1}{\partial \varphi_2} \cdot \frac{dP_2}{d\lambda}. \tag{A15}$$

The lemma is essentially a corollary to (A14) and (A15).

From (21), the transit authority has complete control over *L*, and here it will be simpler to work with the system consisting of (23)-(24), (8) and (20) for the determination of equilibrium trip prices and quantities. With the constraint on $\varphi_1$, a solution takes the form

$$N_i = N_i(\varphi_2, L); \quad i = 1, 2$$

$$P_i = P_i(\varphi_2, L); \quad i = 1, 2.$$

We have

$$\frac{dN_i}{d\lambda} = \frac{\partial N_i}{\partial \varphi_2} \cdot \frac{d\varphi_2}{d\lambda} + \frac{\partial N_i}{\partial L} \cdot \frac{dL}{d\lambda}$$

$$= \frac{\partial N_i}{\partial \varphi_2} \cdot \frac{d\varphi_2}{d\lambda} + \beta \frac{\partial N_i}{\partial \varphi_2} \cdot \frac{dL}{d\lambda}$$

(The previous step is evident from the way in which $\varphi_2$ and *L* enter the system; it can be established formally by deriving $\frac{\partial N_i}{\partial L}$ and comparing it to $\frac{\partial N_i}{\partial \varphi_2}$.)

$$= \frac{\partial N_i}{\partial \varphi_2}\left(\frac{d\varphi_2}{d\lambda} + \beta \frac{dL}{d\lambda}\right)$$

$$= \frac{\partial N_i}{\partial \varphi_2} \cdot \frac{dP_2}{d\lambda}. \qquad \text{(from (20))}$$

This is (A14). (A15) is derived in the same way.

Now, consider an infinitesimal increase in $\lambda$ and suppose that $P_2 \downarrow$. Then from (A14) and

(A15) and the signs in the second row of Table 1, along with the sign of

$$\frac{\partial P_1}{\partial \varphi_2} = \frac{\beta s_{12}}{s - \beta s_{11}} > 0,$$

all of the effects indicated in (48) must occur. If $P_2 \uparrow$, then everything is reversed. What remains to be established is that $P_2$ must in fact change. If it did not, then from (A14)-(A15) there could be no change in $P_1$, $N_1$, or $N_2$. With no change in $N_2$, there could be no change in $R$ or $K$. In (44), there would be no change other than in $\frac{\partial C}{\partial N_1}$, which would violate that condition. Q.E.D.

## References


1. R. Arnott and A. Yan, The two-mode problem: Second-best pricing and capacity, *Review of Urban and Regional Development Studies*, 12, 170-199 (2000).
2. R.M. Braid, Peak-load pricing of a transportation route with an unpriced substitute, *Journal of Urban Economics*, 40, 179-197 (1996).
3. A. de Palma and R. Lindsey, Private toll roads: Competition under various ownership regimes, *Annals of Regional Science*, 34, 13-35 (2000).
4. E.L. d'Ouville and J.F. McDonald, Optimal road capacity with a suboptimal congestion toll, *Journal of Urban Economics*, 28, 34-49 (1990).
5. J.V. Henderson, "Economic Theory and the Cities," 2nd ed., Academic Press, Orlando (1985).
6. M. Kraus and Y. Yoshida, The commuter's time-of-use decision and optimal pricing and service in urban mass transit, *Journal of Urban Economics*, 51, 170-195 (2002).
7. H. Lévy-Lambert, Tarification des services à qualité variable: Application aux péages de circulation, *Econometrica*, 36, 564-574 (1968).
8. L.N. Liu and J.F. McDonald, Efficient congestion tolls in the presence of unpriced congestion: A peak and off-peak simulation model, *Journal of Urban Economics*, 44, 352-366 (1998).
9. L.N. Liu and J.F. McDonald, Economic efficiency of second-best congestion pricing schemes in urban highway systems, *Transportation Research*, 33B, 157-188 (1999).
10. M. Marchand, A note on optimal tolls in an imperfect environment, *Econometrica*, 36, 575-581 (1968).
11. H. Mohring, 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.), Johns Hopkins University Press, Baltimore (1979).

12. I. Pressman, A mathematical formulation of the peak-load pricing problem, *Bell Journal of Economics and Management Science*, 1, 304-326 (1970).
13. R. Sherman, Congestion interdependence and urban transit fares, *Econometrica*, 39, 565-576 (1971).
14. K.A. Small, The scheduling of consumer activities: Work trips, *American Economic Review*, 72, 467-479 (1982).
15. K.A. Small and J. Yan, The value of "value pricing" of roads: Second-best pricing and product differentiation, *Journal of Urban Economics*, 49, 310-336 (2001).
16. E. Verhoef, P. Nijkamp and P. Rietveld, Second-best congestion pricing: The case of an untolled alternative, *Journal of Urban Economics*, 40, 279-302 (1996).
17. E. Verhoef and K.A. Small, "Product Differentiation on Roads: Second-Best Congestion Pricing with Heterogeneity under Public and Private Ownership," Irvine Economics Paper 99-00-01, University of California at Irvine (1999).
18. W.C. Wheaton, Price-induced distortions in urban highway investment, *Bell Journal of Economics*, 9, 622-632 (1978).
19. J.D. Wilson, Optimal road capacity in the presence of unpriced congestion, *Journal of Urban Economics*, 13, 337-357 (1983).

## Footnotes

1. Mohring [11] presents some simulation results, but does not treat the problem analytically.
2. The same is true of the two-road bottleneck specification in de Palma and Lindsey [3], which is mainly concerned with private toll roads.
3. For strong evidence supporting this assumption, see Small [14].