Road Pricing with Optimal Mass Transit

by

Marvin Kraus*

April 2012

*Department of Economics, Boston College, Chestnut Hill, MA 02467, USA. E-mail: kraus@bc.edu. I am grateful to Richard Arnott, Clifford Winston and two anonymous referees for helpful comments.

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. Using analytical methods, a global comparison is made between the second-best levels of transit service and the fare and their first-best levels. The fact that the results are global permits an application to road pricing not possible with the local results of Kraus (2003).

Road Pricing with Optimal Mass Transit

1. Introduction

Kraus (2003) considered the second-best policy problem that arises when auto travel is priced below its marginal cost and there is a substitute mass transit mode. The present paper revisits the problem, obtaining much stronger results in a richer model. The new results have a direct application to road pricing.

As in Kraus (2003), the focus is on the relationship between first- and second-best levels of transit service. The earlier paper established that the second-best level is higher, but only as a local result that does not necessarily hold away from the first-best optimum. The present paper extends this to a global result that applies to discretely underpriced auto travel. The fact that the result is global is what makes the road pricing application possible.

The models of the two papers differ only slightly, but the slight change in specification makes for a much richer model. In the 2003 paper, the only externalities imposed by auto commuters are congestion externalities, and the bottleneck model used to model them gives rise to a first-best auto toll that rises linearly over the morning commuting period, with the toll being zero at the beginning of the period. If there is a shift from first-best tolling to a no-tolling regime, then inefficient queuing is introduced, driving up the marginal cost of an auto trip for the same number of auto commuters. While having this *direct marginal cost effect*, there is *no direct price effect*, since equilibrium trip price for the same number of auto commuters is unchanged. In the present paper, I assume that an auto commuter not only imposes a congestion externality, but also an exogenously-given environmental externality. The first-best auto toll now increases off of an initial level determined by the environmental externality, and when there is a shift from first-best tolling to a no-tolling regime, there is both a direct marginal cost effect and a direct price effect, which is much more realistic than having only a direct marginal cost effect.

The application of the analysis to road pricing is one of road pricing accompanied by efficient pricing and service provision in transit. It is assumed that there is initially no road pricing,

resulting in a below-marginal-cost price for the auto mode. There is a substitute mass transit mode for which pricing and service provision are second best. This calls for below-marginal-cost pricing in transit. Then road pricing is introduced, establishing a marginal-cost price for the auto mode. This is accompanied by reoptimization of both transit pricing and service, and a marginal-cost price is established for the transit mode. For analytical purposes, I consider this regime shift as a two-step process, the first of which is the pricing of the environmental externality through the introduction of a uniform toll. The second step is a shift from this uniform toll to the optimal time-varying toll. The first step involves a direct price increase for the auto mode, making a fare increase optimal to resolve the trading off of deadweight losses for the two modes. Not only does auto traffic decrease, but also transit ridership. As a result, transit service decreases. The second step is the peak-load pricing effect and involves a direct marginal cost reduction for the auto mode. This amounts to an increase in the efficiency of the auto mode relative to transit, making it optimal for auto use to increase and transit use to decrease. The transit policy that is called for is a further fare increase along with a further reduction in service.

A considerable literature has grown up on the second-best policy implications of having auto travel priced at less than marginal cost. Within this literature, there are two sets of papers closely related to the present one. The larger set of papers focuses on second-best pricing of a substitute mode or road and includes early contributions by Lévy-Lambert (1968), Marchand (1968) and Sherman (1971), and more recent ones by Braid (1996), Liu and McDonald (1998), Small and Yan (2001), Verhoef, Nijkamp and Rietveld (1996) and Verhoef and Small (2004). Papers in this set take the level of transit service as exogonous. In the second set of papers (Ahn (2009), Pels and Verhoef (2007)), transit service is endogenous and is optimized along with the fare. In that respect, these papers are like the present one. They differ from the present paper in that the results they present are obtained through simulation (the results presented here are obtained analytically) and the models used are not peak-load pricing models (which the model of the present paper is). A simulation study that does take into account peak-load pricing is Winston and Shirley (1998). It

differs from the other papers in that the base equilibrium that it uses as a reference point is not the second-best optimum.

The next section presents the model, while the analysis is presented in Sections 3 and 4. Section 5 provides the application to road pricing, while Section 6 concludes.

2. The Model

Consider two points, A and B, where individuals, who are assumed to be identical, respectively live and work. A and B are connected by a highway as well as by a rail line. Individuals have a common work start time t^*, by which time they must be at work (arriving late is prohibitively costly). An individual who arrives at work at time $t' \leq t^*$ incurs a schedule delay cost of $\beta(t^* - t')$, where $\beta > 0$ is a given schedule delay cost parameter.

2.1 *Highway Submodel*[1]

The highway submodel is a bottleneck model of the type studied by Arnott, de Palma and Lindsey (1993). The highway is assumed to be uncongested, except at a single bottleneck. The bottleneck's capacity – the maximum rate at which cars can pass through the bottleneck per hour – is exogenous and denoted by s. An arrival rate at the bottleneck exceeding s causes a queue to form. Queuing time costs at the bottleneck are an auto commuter's only travel costs.

We will see shortly how the bottleneck's limited capacity gives rise to the imposition of congestion externalities on the part of auto commuters. In addition, each auto commuter is assumed to impose an exogenous environmental externality of θ.

For any given number of auto commuters, N_1, optimal temporal use of the highway requires that the sum of their schedule delay and queuing costs is at a minimum (regardless of the time-of-use pattern, aggregate environmental cost is θN_1). This is achieved when auto commuters depart from the origin point A at a uniform rate of s over a time interval from $t_o \equiv t^* - N_1/s$ to t^*. Under this departure pattern, there is no queuing.

Turning to decentralization, the toll at departure time t is $\tau(t)$, with trip price given by the toll plus the time cost to the user. Individuals choose departure times to minimize trip price, implying

a common trip price at all chosen departure times in equilibrium. It follows that for any value of the parameter φ_1, the toll function defined by

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

results in decentralization of the optimal departure pattern. Once the specification of the model is complete, we will see that the optimal value of φ_1 is θ.

With a starting point of (1) with φ_1 optimized, we wish to consider a shift to the no-toll case of $\tau(t) = 0$ for all $t \le t^*$. We assume that an individual has a constant cost per unit of time spent queuing of α and that $\alpha > \beta$. We also introduce a pair of parameters γ and λ, each taking on a value in $[0,1]$, and consider the following generalization of (1):

$$\begin{aligned} \tau(t) &= (1-\gamma)\varphi_1 \qquad t \le t_o \\ &= (1-\gamma)\varphi_1 + (1-\lambda)\beta(t - t_o) \qquad t \in [t_o, t^*]. \end{aligned} \tag{2}$$

When γ and λ are both zero, (2) reduces to (1). The optimal value of φ_1 in this case is θ. When either γ or λ is positive, we constrain φ_1 to equal θ, creating a second-best problem in which either the toll level is suboptimal $(\gamma > 0)$ or there is a suboptimal degree of peak-load pricing $(\lambda > 0)$. $(\gamma, \lambda) = (1,1)$ generates the no-toll case. Our focus will therefore be on a shift in (γ, λ) from $(0,0)$ to $(1,1)$. We will identify the separate effects of the toll level and peak-load pricing by considering a two-part shift in which $(\gamma, \lambda) = (0,1)$ is used as an intermediate point.

For any positive value of λ, equilibrium in departures requires a certain amount of queuing, and as λ increases, queuing becomes more pronounced. A fuller treatment, along with a derivation of the aggregate time cost of auto commuters, is given in Kraus (2003). Adding the environmental cost, the total resource cost of auto trips is given by

$$C(N_1, s, \lambda) \equiv \Gamma(\lambda)\beta N_1^2 / 2s + \theta N_1, \tag{3}$$

where

$$\Gamma(\lambda) \equiv \frac{\alpha - \beta + \lambda(\alpha + \beta)}{\alpha - \beta + \lambda\beta}. \tag{4}$$

It is easily checked that $\Gamma(\cdot)$ is a monotonically increasing function, and that $\Gamma(0) = 1$. Also, the equilibrium price of an auto trip is given by

$$P_1 = \beta N_1 / s + (1-\gamma)\varphi_1. \tag{5}$$

2.2 *Mass Transit Submodel*

The mass transit submodel is identical to that in Kraus (2003), where it is presented in detail. The following is a summary.

A transit commuter has no travel costs except possibly for a waiting time cost at the origin train stop. In addition:

N_2 is the number of transit passengers, while

R is the number of train departures (runs) from A.

σ passengers is the capacity of a run. σ and R must satisfy $R\sigma \geq N_2$. At the optimum, $R\sigma = N_2$.

K is the number of physically distinct trains used to make the runs.

T is the time it takes for a train to make a roundtrip from A to B. The successive runs of a train unit must therefore be scheduled at least T minutes apart. There is also a safe headway constraint: successive train departures from A must be at least δ minutes apart.

The cost of providing transit service is given by

$$(\nu_o + \nu_1\sigma)TR + \nu_2\sigma K + \nu_3\sigma + \nu_4 T, \tag{6}$$

where $\nu_o, \nu_1, \ldots, \nu_4$ are given parameters. The first term in (6) is the operating costs of runs. Since a train requires a driver regardless of its capacity, operating costs have a component that is independent of σ. The second term in (6), which we refer to as fleet costs, gives the nonoperating capital costs for the transit authority's fleet of cars. The final two terms are respectively capital costs for terminals (at A and B) and right-of-way and construction costs for trackage. Trackage costs are proportional to the distance between A and B and therefore to T.

Given N_2, cost minimization involves a pattern of commuter arrivals at the origin stop in which a mass of passengers of size σ arrives at each of the train departure times. That way, there

is no waiting. Optimal scheduling, in turn, involves running trains in clusters, with runs within a cluster separated by δ, and clusters separated by T. Letting *SDC* denote the aggregate schedule delay costs of transit commuters under the optimal schedule, *SDC* is given by

$$SDC = \frac{\beta N_2}{2}[\delta(K-1) + T(\frac{R}{K} - 1)]. \tag{7}$$

The remaining problem is to optimize R and K. Using $R\sigma = N_2$ to express (6) as a function of N_2, R and K, we formulate the problem

$$\min_{R,K} \quad \Phi(N_2, R, K) \tag{8}$$

where

$$\Phi(N_2, R, K) \equiv \frac{\beta N_2}{2}[\delta(K-1) + T(\frac{R}{K} - 1)] + \nu_o TR + \nu_1 TN_2 + \frac{(\nu_2 K + \nu_3)N_2}{R} + \nu_4 T \tag{9}$$

is total cost for the transit mode.

For later use, we will need to know how the solution to (8) varies with N_2. Our results are stated in the following lemma from Kraus (2003), 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 (8):

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

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

Another property we will need is that (8) gives rise to a declining long run marginal cost curve for transit trips. The reason is having a fixed cost of runs in (6) ($\nu_o > 0$). The long run cost function for transit trips is given by the value function for (8) or $LRTC_2(N_2) \equiv \Phi(N_2, R(N_2), K(N_2))$, where $R(N_2)$ and $K(N_2)$ give the optimal values of R and K in terms of N_2. It is straightforward to use the first-order conditions for (8) to show that the slope of the long run marginal cost curve, $LRMC_2'(N_2)$, is given by

$$LRMC_2'(N_2) = -\nu_o T \frac{R}{N_2^2} E_{R:N_2} < 0. \tag{10}$$

We will also use the property that

$$\lim_{N_2 \to \infty} LRMC_2'(N_2) = 0, \tag{11}$$

which follows from the presence of N_2^2 in the denominator of (10) and that as N_2 increases, R in the numerator decreases relative to $\sqrt{N_2}$ (Thus, (10) decreases at least as rapidly as $N_2^{-3/2}$).

Remark. It is easily shown that the model gives rise to scale economies in providing transit trips (declining long run average cost curve for transit trips). This is partly accounted for by the fixed cost of runs, and partly by trackage costs. To see this, suppose there is a doubling in N_2, and that the transit authority responds by doubling σ, leaving R and K unchanged. From (7), aggregate schedule delay costs would double. In (6), the transit authority's fleet and terminals costs would double (second and third terms, respectively), as would its variable costs of runs (costs of runs dependent on σ. But its fixed costs of runs and tracking costs would remain unchanged, resulting in a less than doubling in the transit mode's total cost.[2]

For decentralization of the optimum, the higher schedule delay costs associated with earlier runs must be offset by lower fares. The equilibrium price of a transit trip can be written

$$P_2 = \varphi_2 + \beta L, \tag{12}$$

where φ_2 and L are respectively the fare for the earliest run and the number of minutes before t^* that the earliest run is scheduled. With optimal scheduling,

$$L = \delta(K-1) + T\left(\frac{R}{K} - 1\right), \tag{13}$$

and (12) becomes

$$P_2 = \varphi_2 + \beta\left[\delta(K-1) + T\left(\frac{R}{K} - 1\right)\right]. \tag{14}$$

2.3 *Demand and Overall Equilibrium*

We employ the simplest possible demand specification, taking modal trip demands to be those of a representative utility-maximizing 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

$$N_1 = N_1(P_1, P_2) \tag{15}$$

$$N_2 = N_2(P_1, P_2). \tag{16}$$

The fact that (15)-(16) 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 $i, j = 1, 2$. In addition to $s_{ii} < 0$ for $i = 1, 2$, we have that $s_{11}s_{22} - s_{12}s_{21} > 0$ and $s_{12} = s_{21}$. We also assume that auto and transit trips are substitutes, so that $s_{12} > 0$.

Given values for $\varphi_1, \varphi_2, R, K$ and γ, 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 (15)-(16) and the two supply relationships (5) and (14). A solution takes the form

$$N_i = N_i(\varphi_1, \varphi_2, R, K, \gamma); \quad i = 1, 2 \tag{17}$$

$$P_i = P_i(\varphi_1, \varphi_2, R, K, \gamma); \quad i = 1, 2. \tag{18}$$

Note that the system consisting of (5) and (14)-(16) does not involve λ. λ does not have a direct effect on equilibrium trip prices and quantities. It affects them only indirectly by affecting the optimal values of φ_2, R and K. γ affects equilibrium trip prices and quantities both directly and indirectly through φ_2, R and K.

3. First- and Second-Best Problems

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

$$B(N_1, N_2) \equiv \int_{(0,0)}^{(N_1, N_2)} P_1(n_1, n_2)dn_1 + P_2(n_1, n_2)dn_2, \tag{19}$$

while costs are given by (3) for auto trips, and by (9) for transit. Under our assumption that trip demands are independent of income, (19) is not only path-independent, but also has the property that the marginal benefit of a mode i trip is equal to its demand price.

Denoting $N_i = N_i(\varphi_1, \varphi_2, R, K, \gamma)$ in (17) by $N_i(\cdot)$, we write social surplus as

$$B(N_1(\cdot), N_2(\cdot)) - C(N_1(\cdot), s, \lambda) - \Phi(N_2(\cdot), R, K), \tag{20}$$

where the cost functions C and Φ come from (3) and (9). (20) is maximized in both first- and

second-best problems, albeit for different (γ, λ) combinations.

3.1 *First-Best Problem*

In the first-best problem, (20) is maximized with $(\gamma, \lambda) = (0,0)$. φ_1 is unconstrained and is optimized along with φ_2, R and K. The first-order conditions reduce to the marginal cost pricing conditions $P_1 = \partial C/\partial N_1$ and $P_2 = \partial \Phi/\partial N_2$ and the first-order conditions,

$$\frac{\partial \Phi}{\partial R} = \frac{\partial \Phi}{\partial K} = 0, \tag{21}$$

for the transit cost minimization problem (8). For a demonstration, see Kraus (2003).

From the first of the marginal cost pricing conditions, it is easy to derive $\varphi_1 = \theta$. For $\lambda = 0$, $\Gamma(\lambda) = 1$. Using this in (3) gives

$$\frac{\partial C}{\partial N_1} = \beta N_1/s + \theta. \tag{22}$$

The result $\varphi_1 = \theta$ then follows from using $\gamma = 0$ in (5) and equating to (22).

The first-best highway toll is thus given by (2) with $\varphi_1 = \theta$. Its level (at t_o) of θ internalizes the environmental externality, while its slope of β internalizes congestion externalities. To see the latter, note that if a marginal auto commuter is added at some time t in the first-best departure interval, then 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 $\beta(t - t_o)$.

φ_2, the fare for the earliest transit run, is also positive. The explanation for this result, which is derived in Kraus (2003), is twofold. First, a transit passenger who is relocated from the earliest departure time to a new earlier time (to make room for a marginal passenger at her initial departure time) has her schedule delay cost increased *discretely*. Second, a new run would have to be added, resulting in higher operating costs for the transit authority.

3.2 *Second-Best Problem*

We now turn to what we will refer to as the general second-best (γ, λ) problem: Given a (γ, λ) combination such that either γ or λ (or both) is positive (neither parameter can take on a value greater than unity), and with φ_1 constrained to equal θ, (20) is maximized with respect to φ_2, R and K. After rearranging terms, the first-order condition for φ_2 can be written

$$P_2 - \frac{\partial \Phi}{\partial N_2} = -(P_1 - \frac{\partial C}{\partial N_1})\frac{\partial N_1}{\partial \varphi_2} \Big/ \frac{\partial N_2}{\partial \varphi_2}. \tag{23}$$

In (23), $P_1 - \partial C/\partial N_1 < 0$. This follows from (5) and $\partial C/\partial N_1 = \Gamma(\lambda)\beta N_1/s + \theta$, since either γ or λ is now positive, $\varphi_1 = \theta$, and $\Gamma(\lambda) > 1$ whenever $\lambda > 0$. Also in (23), $\partial N_2/\partial \varphi_2 < 0$, while $\partial N_1/\partial \varphi_2 > 0$, the latter from the assumption that auto and transit trips are substitutes $(s_{12} > 0)$.[3] Combining these results gives $P_2 - \partial \Phi/\partial N_2 < 0$ or that second-best pricing of transit requires that it, too, be priced below marginal cost.[4]

Despite the fact that pricing is no longer first-best, there is no distortion away from cost minimization in transit, and the first-order conditions for R and K again reduce to (21). The demonstration follows Kraus (2003). The same result was obtained in previous analyses of the problem by Henderson (1985) and Arnott and Yan (2000) and depends crucially on the two modes having noninterdependent costs.

4. First-Best versus the No-Toll Case

The best way to understand road pricing (a regime shift from $(\gamma, \lambda) = (1,1)$ to $(0,0)$) is to first consider a shift in (γ, λ) from $(0,0)$ to $(1,1)$ (this section) and then the reverse shift from $(1,1)$ to $(0,0)$ (next section). In each case, we consider a two-part shift with $(0,1)$ as an intermediate point.

4.1 $(0,0) \rightarrow (0,1)$

We start by considering a shift from $(0,0)$ to $(0,1)$. This amounts to eliminating peak-load pricing, while maintaining the first-best toll level of θ. This has the effect stated in the following proposition, where $E_{N_2:P_2}$ denotes the own-price elasticity of demand for mode 2 trips.

Proposition 1. Suppose that $\left|E_{N_2:P_2}\right| \leq 2$ at the first-best optimum and that there is a shift in (γ, λ) from $(0,0)$ to $(0,1)$. Then the following effects take place on equilibrium trip prices and quantities and transit authority policy variables:

$$N_1 \downarrow,\ N_2 \uparrow,\ P_1 \downarrow,\ P_2 \downarrow \tag{24}$$

and

$$\varphi_2 \downarrow,\ R \uparrow,\ K \uparrow,\ \sigma \uparrow. \tag{25}$$

Proof. See the Appendix.

In Kraus (2003), it was shown that under the same elasticity condition that appears in Proposition 1, the effects indicated in (24) and (25) hold locally at the first-best optimum – that is, for an infinitesimal increase in λ coming off of $\lambda = 0$.[5] Proposition 1 has far greater practical importance, holding for a discrete increase in λ from 0 to 1.[6]

The fact that Proposition 1 rests on an elasticity condition can be seen as follows. A shift in (γ, λ) from $(0,0)$ to $(0,1)$ has the direct effect of driving the marginal cost of a mode 1 trip above its price, creating a deadweight loss for mode 1. A reduction in P_2 has two opposing effects on the mode 1 deadweight loss. By the assumption that mode 1 and mode 2 trips are substitutes, it leads to a decrease in N_1, the effect of which is to make the mode 1 deadweight loss smaller. But it also leads to a decrease in P_1 – the reduction in N_1 makes the mode 1 departure interval shorter – and this has the effect of making the mode 1 deadweight loss larger. As long as the own-price elasticity of transit demand is not too great, the first effect dominates the second, and a reduction in P_2 is desirable.[7]

The policy effects in (25) are easily understood in light of the price-quantity effects in (24). The key is that the second best involves no distortion away from cost minimization in transit. This means that transit service always accords to (8) and that Lemma 1 applies. Hence the results for R, K and σ. Lemma 1 also implies that R goes through a larger percentage increase than K, which makes the mode 2 departure interval longer (equation (13)). Thus the only way to effect a reduction in P_2 is to lower φ_2 (equation (12)).

4.2 $(0,1) \rightarrow (1,1)$

We next consider the shift from (0,1) to (1,1). This amounts to a shift to a no-tolling regime from one that sets a uniform toll of θ. Equilibrium trip prices and quantities are now affected both directly and indirectly through φ_2, R and K. What follows takes into account all of these effects.

Before proceeding, we note that what makes this second part of the shift part of the analysis is the present paper's extension of the 2003 model through the inclusion of the environmental externality θ. It is in this second part of the shift that the direct price effect of the overall shift referred to in the introduction is captured. As a result, the analysis of this and the following subsection has no counterpart in the 2003 paper.

We proceed by introducing a perturbation in γ into the seven-equation system of supply and demand functions and optimal policy conditions given by (5), (14)-(16), (23) and (21) and then deriving expressions for the comparative statics derivatives $\frac{dN_1}{d\gamma}$ and $\frac{dN_2}{d\gamma}$. To focus on the most important effects, for each (γ, λ) combination we employ a linear approximation to the demand functions (15) and (16), with the linearization carried out about the equilibrium price combination (P_1, P_2) for the (γ, λ) combination of interest. The expressions we obtain for $\frac{dN_1}{d\gamma}$ and $\frac{dN_2}{d\gamma}$ are therefore approximations. They are:

$$\frac{dN_1}{d\gamma} = -\frac{\theta s}{H}(s_{11}s_{22} - s_{12}s_{21})(1 - LRMC_2'(N_2)\cdot\frac{\partial N_2}{\partial \varphi_2}) \tag{26}$$

$$\frac{dN_2}{d\gamma} = \frac{\beta\theta}{H}\frac{\partial N_1}{\partial \varphi_2}(s_{11}s_{22} - s_{12}s_{21})(1 - \Gamma(\lambda)), \tag{27}$$

where H is the expression

$$H \equiv (s - \beta s_{11})\frac{\partial N_2}{\partial \varphi_2}(1 - LRMC_2'(N_2)\cdot\frac{\partial N_2}{\partial \varphi_2}) + \beta s_{12}\frac{\partial N_1}{\partial \varphi_2}(1 - \Gamma(\lambda)). \tag{28}$$

It is at this point that we use the asymptotic property of $LRMC_2$. From (11), the limiting value of H for large N_2 is negative (recall that $\Gamma(\lambda) > 1$). Thus, in the limit we have that (26) and (27) are both positive. This is the basis of:

Proposition 2. Consider a shift in (γ, λ) from $(0,1)$ to $(1,1)$. Then up to an approximation that ignores nonlinearities in demand, the following effects hold asymptotically for large N_2:

$$N_1\uparrow,\ N_2\uparrow,\ P_1\downarrow,\ P_2\downarrow \tag{29}$$

and

$$\varphi_2\downarrow,\ R\uparrow,\ K\uparrow,\ \sigma\uparrow. \tag{30}$$

Proof. Having established that N_1 and N_2 increase, the price effects indicated in (29) follow from the properties of substitution effects (Section 2.3), in particular that mode 1 and mode 2 trips are substitutes. The policy effects in (30) are qualitatively identical to those in Proposition 1 and follow from the mode 2 price-quantity effects in (29) in the same way as with Proposition 1.

Remark. Proposition 2 can be thought of as a result that applies to large cities. Because they offer the potential for large welfare gains from road pricing, it is large cities that are of greatest interest. Nothing in the proposition requires that N_2 is large relative to N_1, only that N_2 is large in an absolute sense. Under this scale condition, we can exploit the asymptotic behavior of $LRMC_2$ to unambiguously sign (26) and (27). And as noted in Section 2, (10) decreases at least as rapidly as $N_2^{-3/2}$, enhancing the applicability of the proposition.

4.3 $(0,0)\rightarrow(1,1)$

Combining Propositions 1 and 2 gives:

Proposition 3. Suppose that $\left|E_{N_2:P_2}\right| \le 2$ at the first-best optimum and that there is a shift in (γ, λ) from $(0,0)$ to $(1,1)$. Then up to an approximation that ignores nonlinearities in demand, the following effects hold asymptotically for large N_2:

$$N_1?,\ N_2\uparrow,\ P_1\downarrow,\ P_2\downarrow \tag{31}$$

and

$$\varphi_2\downarrow,\ R\uparrow,\ K\uparrow,\ \sigma\uparrow. \tag{32}$$

Proposition 3 is best understood as follows. Starting from the first-best, first consider an increase in λ from 0 to 1. This eliminates peak-load pricing of the highway, causing queuing to

occur at the bottleneck. The highway becomes a less efficient facility, and the marginal cost of an auto trip now exceeds its price. This is the result of a direct marginal cost effect. There is no direct price effect, since queuing substitutes for peak-load pricing. The result is that a deadweight loss is introduced for the highway. In order to reduce it, the transit fare is decreased.

Holding the value of λ fixed at 1, now consider an increase in γ from 0 to 1. This amounts to a shift to a no-toll regime from one that tolls uniformly at the first-best level θ. The gap between the price and marginal cost of an auto trip becomes larger, this time as a result of a direct price effect. This increases the deadweight loss for the highway, and in order to reduce it, the transit fare is decreased further.

5. Road Pricing

Road pricing is a regime shift from $(\gamma, \lambda) = (1,1)$ to $(0,0)$ (the reverse of the shift considered in the preceding section). A regime that employs no pricing of the externalities imposed by auto commuters is replaced by one that employs first-best pricing. Each regime optimizes mass transit pricing and service subject to its policy for pricing the highway. We analyze the regime shift as a two-part shift with $(0,1)$ as an intermediate point.

First consider the shift from $(1,1)$ to $(0,1)$. This introduces a uniform toll at the first-best level θ, narrowing the gap between price and marginal cost for the auto mode. This reduction in the degree of underpricing of the auto mode makes a fare increase optimal, so that there is also a reduction in the degree of underpricing of the transit mode (which is priced below marginal cost at $(1,1)$). The result is that both N_1 and N_2 decrease, the latter leading to a reduction in transit service.

Now consider the shift from $(0,1)$ to $(0,0)$. This introduces peak-load pricing of the highway, while maintaining the first-best toll level θ. Queuing is eliminated at the bottleneck, making the highway a more efficient facility. There is a reduction in the marginal cost of an auto trip, resulting in marginal cost pricing of the highway. This makes a further increase in the fare optimal, so that the transit mode is also priced at marginal cost. There is a further decrease in N_2,

but an increase in N_1 relative to (0,1). Transit service decreases along with N_2.

Overall, N_2 and transit service decrease; the effect on N_1 is ambiguous.

6. 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. By obtaining global results (relative to the first best) in a model in which both transit pricing and service provision are endogenous, the paper has been able to go considerably farther with this problem than have previous attempts to treat it analytically. The main policy application of the paper is to the introduction of an optimal road pricing scheme. Initially, it is assumed that there is no road pricing and that transit pricing and service provision are second best. Then road pricing is introduced and, along with it, transit pricing and service provision are reoptimized. The transit policy adjustments that are called for are an increase in the fare and a reduction in service. What drives these results is that the initial second best is characterized by below-marginal-cost pricing in transit, and once road pricing is introduced, the efficiency rationale for this is eliminated.

I conclude with a couple of comments about the road pricing application. First, there is nothing inconsistent about this paper's finding that a move from second to first best entails a reduction in transit service and the policy that London followed of increasing transit service when it introduced road pricing in 2003 (Leape (2006)). The reason is that in London there was no change in the level of transit fares. In contrast, this paper's results apply when the transit fare is unconstrained and is reoptimized with the introduction of road pricing so that the maximum efficiency gain from road pricing is realized. Second, there is evidence that the below-marginal-cost price of transit in the base equilibrium off of which road pricing is introduced in this paper and which is so crucial to the analysis is also a characteristic of real-world urban areas. The evidence comes from a recent empirical analysis by Parry and Small (2009), who did price/marginal cost comparisons for transit operations in Washington, DC, Los Angeles and London. Price was found to be below marginal cost in all cases.

Appendix

Proof of Proposition 1. It will be assumed throughout that $\gamma = 0$.

Lemma 2 of Kraus (2003) proves that none of the equilibrium trip prices and quantities can be stationary with respect to λ (no sideways arrows for any of the variables in (24) for an infinitesimal increase in λ). The important thing is that this holds not just for $\lambda = 0$, but for any value of λ in the unit interval. Proposition 1 of Kraus (2003) proves that under the same elasticity restriction that appears in this paper's Proposition 1, an infinitesimal increase in λ coming off of an initial value for λ of zero has the effects on equilibrium trip prices and quantities indicated in (24). Combining the two and using continuity, we have that the effects indicated in (24) hold for an infinitesimal increase in λ coming off of any initial value of λ in the unit interval. Thus, the effects indicated in (24) must hold when λ is increased discretely from 0 to 1.

The proof of the policy effects in (25) is insightful and is therefore included in the text (the final paragraph of Section 4.1).

References

Ahn, Kijung. 2009. "Road Pricing and Bus Service Policies," *Journal of Transport Economics and Policy*, 43(1): 25-53.

Arnott, Richard, de Palma, Andre and Robin Lindsey. 1993. "A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand," *American Economic Review*, 83(1): 161-79.

Arnott, Richard and An Yan. 2000. "The Two-Mode Problem: Second-Best Pricing and Capacity," *Review of Urban and Regional Development Studies*, 12(3): 170-99.

Braid, Ralph M. 1996. "Peak-Load Pricing of a Transportation Route with an Unpriced Substitute," *Journal of Urban Economics*, 40(2): 179-97.

Henderson, J. Vernon. 1985. *Economic Theory and the Cities*, 2nd ed., New York: Academic Press.

Kraus, Marvin. 2003. "A New Look at the Two-Mode Problem," *Journal of Urban Economics*, 54(3): 511-30.

Leape, Jonathan. 2006. "The London Congestion Charge," *Journal of Economic Perspectives*, 20(4): 157-76.

Lévy-Lambert, H. 1968. "Tarification des Services à Qualité Variable: Application aux Péages de Circulation," *Econometrica*, 36(3): 564-74.

Liu, Louie N. and John F. McDonald. 1998. "Efficient Congestion Tolls in the Presence of Unpriced Congestion: A Peak and Off-Peak Simulation Model," *Journal of Urban Economics*, 44(3): 352-66.

Marchand, Maurice. 1968. "A Note on Optimal Tolls in an Imperfect Environment," *Econometrica*, 36(3): 575-81.

Mohring, Herbert. 1972. "Optimization and Scale Economies in Urban Bus Transportation," *American Economic Review*, 62(4): 591-604.

Parry, Ian W. H. and Kenneth A. Small. 2009. "Should Urban Transit Subsidies Be Reduced?,"

American Economic Review, 99(3): 700-24.

Pels, Eric and Erik T. Verhoef. 2007. "Infrastructure Pricing and Competition Between Modes in Urban Transport," *Environment and Planning A*, 39: 2119-38.

Sherman, Roger. 1971. "Congestion Interdependence and Urban Transit Fares," *Econometrica*, 39(3): 565-76.

Small, Kenneth A. and Jia Yan. 2001. "The Value of 'Value Pricing' of Roads: Second-Best Pricing and Product Differentiation," *Journal of Urban Economics*, 49(2): 310-36.

Verhoef, Erik, Nijkamp, Peter and Piet Rietveld. 1996. "Second-Best Congestion Pricing: The Case of an Untolled Alternative," *Journal of Urban Economics*, 40(3): 279-302.

Verhoef, Erik and Kenneth A. Small. 2004. "Product Differentiation on Roads: Constrained Congestion Pricing with Heterogeneous Users," *Journal of Transport Economics and Policy*, 38(1): 127-56.

Winston, Clifford and Chad Shirley. 1998. *Alternate Route*, Washington, D.C.: The Brookings Institution.

Footnotes

1. The discussion here is deliberately kept brief. A more detailed discussion can be found in Kraus (2003, Subsection 2.1).
2. Since the long run cost structure involves no passenger waiting costs, a well-known source of scale economies related to waiting costs, the so-called "Mohring effect," (Mohring (1972)) does not operate in the model.
3. Expressions for these comparative statics derivatives can be found in Kraus (2003, Table 1).
4. For references for this well-known result, see Kraus (2003).
5. The local analysis in Kraus (2003) that corresponds to the present subsection's global analysis runs from the middle of p. 524 to the end of Section 3. The present paper's Proposition 1 is a global counterpart to the proposition of the same number and its corollary in the 2003 paper.
6. It is evident from the proof in the Appendix that Proposition 1 holds for any discrete increase in λ such that λ is confined to the unit interval and $\gamma = 0$.
7. The elasticity condition in Proposition 1 is quite unrestrictive. The specific upper bound of 2 is related to our earlier result (Section 2, Lemma 1) that $E_{R:N_2} < 1/2$.