JOURNAL OF REGIONAL SCIENCE, VOL. 44, NO. 3, 2004, pp. 405–435

PRICING, CAPACITY CHOICE, AND FINANCING IN TRANSPORTATION NETWORKS*

Erik T. Verhoef
Department of Spatial Economics, Free University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. E-mail: everhoef@econ.vu.nl

Jan Rouwendal
Department of Spatial Economics, Free University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. E-mail: jrouwendal@feweb.vu.nl

ABSTRACT. This paper explores interrelations between pricing, capacity choice, and financing in transportation networks. We build on the Mohring-Harwitz result on self-financing of optimally designed and priced roads and investigate it in a network environment under various types of second-best regulation. A small network model with endogenous car ownership demonstrates that optimal congestion pricing and capacity choice over an entire network may cause user prices to increase more in initially mildly congested areas compared to heavily congested areas. Furthermore, a flat kilometer charge under optimal capacity choice may result in first-best efficiency gains.

1. INTRODUCTION

Traffic congestion is one of the daily recurring problems resulting from the high car-dependence in most modern societies. The economic approach to analyzing traffic congestion and congestion policies can be summarized as viewing a congested road as a distorted market on which travelers demand a service (the use of the road network), and supply is defined by the capacity of the road(s) in the network. A distortion exists because travel time losses from congestion constitute an externality; individual users consider their own travel times when deciding whether or when to use the road, but typically ignore the implied travel time losses for others (e.g., Small, 1992). Travel speeds are simultaneously determined—among other factors—by the intensity of use and the capacity on offer (e.g., the number of lanes). Two archetype policies for coping with congestion can therefore be distinguished immediately in such economic frameworks. The first involves measures that aim to affect the

*The research of Erik Verhoef has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. Jan Rouwendal is also at the Department of Social Science of Wageningen University. We thank three anonymous referees for helpful comments on an earlier version of this paper. The usual disclaimer applies.

Received: March 2003; Revised: September 2003; Accepted: November 2003.

demand for road use or its distribution over time, place, or links given the network's capacity. The second involves adjustments—usually increases—in capacity levels. Capacity expansion has been the dominant practice of policymaking for most societies. In contrast, much of the economic literature has focused on demand management in general and road pricing in particular (e.g., Button and Verhoef, 1998), motivated by the inefficiency of unregulated congested road use and by the economic insight that corrective pricing could lead to an improvement in efficiency when externalities are present.

Nevertheless, one of the most famous results in transport economics, due to Mohring and Harwitz (1962), establishes an important relationship between demand management and capacity policies; under certain technical conditions, the revenues from optimal congestion pricing will be just sufficient for financing the costs associated with optimal capacity supply. The conditions are satisfied when (1) capacity is adjustable in continuous increments, (2) capacity can be expanded at constant marginal cost, and (3) trip costs are homogeneous of degree zero in usage and capacity. This "self-financing" theorem is an application of the so-called Product Exhaustion Theorem, a particular case of Euler's Theorem, which states that with constant returns to scale in production and marginal cost pricing, the value of outputs equals the value of inputs. The theorem has been shown to extend to each road individually in a full network and consequently to the network in aggregate, provided each link is optimally priced and all capacities are optimized (Yang and Meng, 2002), to dynamic models (Arnott, De Palma and Lindsey, 1993), and in present-value terms when adjustment costs and depreciation are allowed (Arnott and Kraus, 1998).[1]

Empirical evidence suggests that conditions (2) and (3) may hold at least approximately in a range of circumstances; estimates of the ratio of long-run average and marginal costs are often relatively close to unity (Small, 1992, Sections 3.4, 3.5; see also Section 2 this paper). Under such conditions, optimal profits or deficits under optimal road design and pricing will be relatively small, but unlikely zero. Condition (1) typically does not hold for a single road because the number of lanes is discrete. Yet capacity can be varied by widening lanes, by resurfacing, or by regrading or straightening a stretch of road. And capacity may be almost perfectly divisible at the scale of a road network (Lindsey and Verhoef, 2000).

Therefore, the theorem may be relevant for practical policymaking. First, application would help in achieving an efficient road system in terms of optimal capacities and pricing. Furthermore, it firmly reduces the need to

[1]A convenient feature of first-best pricing—on all links of a network—is that the toll formulae for optimal link tolls are independent of traffic conditions on other links, unlike second-best toll formulae (e.g., Verhoef, 2002). Similarly, under first-best pricing link capacities can be optimized without reference to what is going on elsewhere on the network. First-best capacity choice thus requires less information than second-best capacity choice.

use tax revenues from other sources for financing roads. This may improve efficiency further because other taxes are often distortionary. It may also help in overcoming problems of public acceptability of road pricing. The resulting scheme may be perceived as fair (only the users of a road pay for its capacity) and transparent (there are no hidden transfers surrounding the financing of roads). Last, the theorem's application may lead to improved transparency of political decisions regarding infrastructure expansion. If the technical assumptions are fulfilled, road capacity should be expanded when short-run optimal congestion pricing yields revenues per unit of capacity that exceed the unit (capital) cost of capacity.[2] Although comparable rules can be formulated for discrete units of capacity such as lanes, they will be more complex due to the inherent "integer problem." Therefore, the market would indicate whether expansion is socially warranted, generally improving the transparency and credibility of cost–benefit analyses.

Despite these advantages, the self-financing principle has not been applied because the idea of network-wide optimal congestion pricing was much more an academic curiosity than a realistic policy option until recently. However, increases in congestion, the decreasing scope and increasing costs of further capacity expansions in congested areas, and the development of technologies enabling electronic toll collection have changed the viability of congestion pricing in road transportation. Investigating the applicability of the self-financing theorem for practical applications in greater detail, the aim of this paper, is important due to the limited social acceptance of congestion pricing per se and the greater acceptance expected if roads are financed exclusively by their specific tools.

First-best congestion pricing must be applied on all links of the road network for the theorem to hold; however, many practical policy proposals foresee implementation of prices on only a limited number of links. Examples include toll cordons, pay-lanes, and area schemes such as those recently introduced in London. Verhoef (2002) explored the implications for pricing per se under such second-best circumstances. This paper addresses some implications of second-best congestion pricing for the applicability of the self-financing theorem. The second-best cases addressed include pricing on a subset of links, pricing through undifferentiated kilometer charges that is equivalent to a fuel tax under our simplifying assumptions, and pricing through fixed annual vehicle taxes. We also analyze the optimal use of the latter two traditional tax instruments when link-based tolls can also be established. For that reason, the model allows for endogenous car ownership, which is particularly relevant for assessing the effects of changes in annual taxes.

[2]To see why, observe that for a given demand function, both the short-run optimal congestion price (i.e., for a given capacity) and the road use per unit of capacity are decreasing in capacity. Short-run optimal toll revenues per unit of capacity therefore exceed the unit cost of capacity with a below-optimal total capacity.

Especially in Europe, so-called variabilization of vehicle taxes, replacing annual taxes by use-based taxes, is often considered the only viable way of introducing marginal cost pricing in road transport due to acceptability considerations. However, some analysts fear that abandoning annual taxes would lead to increased car ownership and an associated outward shift of the demand for road use so that the scheme's eventual impact on congestion may be limited. Therefore, it is interesting to consider the optimal use of annual taxes when optimal link tolls are in place with endogenous car ownership. Finally, we study how insights might change when a benefit arises from raising such taxes because other taxes are distortionary or when total tax revenues are exogeneously constrained and capacity cannot be freely adjusted.

The paper is organized as follows. Section 2 provides a brief literature review. Section 3 introduces the model and discusses the calibration of its parameters for the numerical version. Section 4 reports our main findings. Section 5 concludes the paper.

2. PRIOR LITERATURE

Mohring and Harwitz (1962) were the first to establish that, under the conditions set out earlier, the revenues from optimal congestion tolling are equal to the cost of providing the optimal capacity of the road. In fact, they derived a more general result, namely that the ratio between toll revenues and capacity cost equals the elasticity of total capacity costs with respect to capacity. Although the Mohring-Harwitz result is now generally considered one of the cornerstones of transportation economics, apparently it was largely unnoticed at the time. Strotz (1965) reached similar results a few years later.[3] The initial lack of attention to the result is possibly explained by doubts about its practical significance for transportation policy. For example, one may doubt whether the cost function is appropriately specified, whether the static model of road congestion is appropriate, and whether congestion tolls can be successfully introduced and sufficiently varied over time. Eventually, these problems received attention and we discuss the main results of more recent studies below.

Keeler and Small (1977) considered the implementation of the Mohring-Harwitz analysis to urban expressways from an empirical perspective. The homogeneity of degree zero of the transport cost function was not assessed empirically, but assumed on the basis of "considerable empirical evidence" (p. 3) and they measured the cost of capacity as the sum of construction, land acquisition, and maintenance costs. They found no evidence of (dis)economies

[3]Strotz's analysis was based on Mohring and Harwitz (1962), and Strotz, Mohring, and Harwitz were all at Northwestern University at that time. Strotz (1965) references Mohring and Harwitz (1962) and Strotz attributes the similarity between his first parable and the Mohring-Harwitz framework to the help provided by Mohring in formulating it.

of scale in construction cost. Therefore, their study confirms the Mohring and Harwitz conjecture that the conditions under which the self-financing result holds are plausible. Kraus (1981) used engineering data in a simple network setting and concluded that there are probably increasing returns to scale. His best estimates show optimal tolling would cover 84 percent of capital cost. He interprets his findings as "not that different from those of Keeler and Small" (p. 20) who found constant returns to scale. Small, Winston, and Evans (1989) suggest there are probably increasing returns to scale in the construction of single roads of a given type because the costs of shoulders and median strip are independent of capacity. Integration of such roads into networks by intersections and other means would probably cause decreasing returns to scale. Their base-case numerical model has practically constant returns to scale. More recently, Levinson and Gillen (1998) report a point estimate for the ratio between long-run average and marginal cost of 0.92 for autos, but 1.45 for single trucks and 1.96 for combination trucks, suggesting mild diseconomies for passenger cars and considerable economies for trucks.

Stahl (1981) developed an analytical model in which capacity cost depends on the number of road users. He states that cost will be exactly recovered under optimal pricing in this framework if the cost function is homogeneous of degree one, but argues that available evidence suggests that these conditions are not satisfied in reality.[4] In particular, he points to "evidence of substantial increasing returns to scale in highway pavement thickness" (p. 18). However, Newbery (1988, 1989) reconsidered this issue and concluded that "if there are constant returns to scale in roads construction (for roads of given strength), and if there are strictly constant returns to road use (in the sense that heavy vehicles distribute themselves uniformly over road width), then the optimal road user charge (congestion charge plus road damage charge) will recover all road costs (maintenance and interest on capital) even if there are substantial economies of scale in road construction" (Newbery, 1989, p. 167). The result requires that there are constant returns to scale in constructing roads with a given strength at different capacities. If this is the case, economies of scale in strengthening roads are unimportant. The optimal policy response to road damage externalities is a flat charge per "equivalent standard axle load" and Newbery (1988) established a self-financing property of this tax that is analogous to the Mohring-Harwitz result with respect to the congestion externality. Small, Winston, and Evans (1989) considered issues of congestion and road damage in a multiproduct framework. Even though they found substantial economies of scale in providing pavement, they conclude that the road system as a whole comes close to self-financing because of the diseconomies of scope

[4]Stahl (1981, p. 18). Stahl's model also contains external effects other than congestion. These external effects should be homogeneous of degree zero in the number of road users and capacity in order to have self-financing of capacity under optimal pricing and capacity choice.

involved in the simultaneous production of road capacity for passenger cars and trucks. The presence of these diseconomies means that the joint supply of infrastructure for cars and trucks causes additional cost in comparison to the provision of separate infrastructure for both types of traffic.

Another issue that has received attention concerns the market for land. Contrary to the standard assumptions, the supply function of land for road capacity expansion may be rising when the road is large enough to drive up the price of land. In this case, the distinction between returns to scale, a property of production functions, and economies of scale, a property of cost functions, becomes important (Small, 1999). A general equilibrium analysis by Berechman and Pines (1991) showed that so-called imputed profits will have the same sign as the degree of returns to scale of the production function, where imputed profits are based on a cost measure for land obtained by multiplying the amount of land used by its shadow price. However, Small (1999) observed that these imputed profits correspond to actual profits only for a factor price-taker. Instead, a road authority often possesses market power in the land market. Small (1999) showed that the sign of actual profits from highway operation under first-best marginal cost pricing will still be determined by the degree of scale of the actual cost function, which differs from the degree of returns to scale of production with a rising supply curve for land. This holds for both cases with and without price discrimination by a monopsonistic road authority in the land market. The critical condition for exact self-financing under marginal cost pricing thus involves the degree of economies of scale of the cost function and not the degree of returns to scale of the production function.

The Mohring-Harwitz result was originally obtained in the context of the static model of road congestion pioneered by Pigou (1920) and Knight (1924). Arnott, de Palma, and Lindsey (1993) showed that the self-financing result also holds for Vickrey's (1969) dynamic bottleneck model when road users are homogeneous. A remarkable aspect of their analysis is that it holds "*independent of the form of the pricing system employed*. If a road system should be self-financing when a sophisticated tolling system is employed, it should also be self-financing when only a flat parking fee is applied" (p. 173, italics in original). Their result stems from the fact that the flat pricing schedule, although second-best, nevertheless confronts all drivers with marginal social costs in the second-best optimum. When this is not the case, self-financing may break down, as illustrated by Mohring (1970), for the case where the same toll level applies both during peak and off-peak periods in a static model. Also Bichsel (2001) showed that the self-financing result does not hold if there are two groups of users that use the road at different times and the toll is restricted to be uniform.

An additional issue concerns heterogeneity of road users. Arnott and Kraus (1998) considered the bottleneck model with heterogeneous users and concluded that marginal cost pricing and the associated self-financing property would still be feasible with an anonymous congestion toll if the

heterogeneity concerns unobservable differences such as the value of time and if tolling is unconstrained, that is, the toll can vary arbitrarily over time.

Last, two additional studies bear a close relationship to the issues studied in this paper although they are not directly related to the question of self-financing. De Borger (2001) developed a model for analyzing transport pricing in which car ownership is endogenized and explained simultaneously with the demand for trips or kilometers. Although De Borger's model refers to an unspecified externality that could be congestion, he does not incorporate road capacity, and therefore does not consider its financing. A related study by Mayeres and Proost (1997) examined optimal tax and public investment rules in the context of an applied general equilibrium model with congestion; however, they do not refer to the Mohring-Harwitz result and provide no information about the ratio between capacity cost and toll revenues in the optimal situation. One of the interesting aspects of their analysis is that the optimal congestion taxes and congestion levels are almost unaffected by the degree of inequality aversion used by the regulator (p. 277). This suggests that distributional concerns are of minor importance in the design of an optimal policy concerning congestion and road infrastructure.

3. THE MODEL

The model should have a number of characteristics to address the research questions identified in Section 1. First, a network larger than a single road is needed to study second-best charges confined to only a subset of links. Second, car ownership should be endogenized to study the role of fixed annual vehicle ownership taxes in the context of "variabilization" and to prevent these from entering as a perfect lump-sum tax. Given these requirements, we develop a model as simple and transparent as possible. Therefore, its numerical version should be considered much more as a mathematical system that allows us to study the main issues in a consistent equilibrium setting than as an attempt to represent any realistic network, although its calibration deploys some empirical evidence. First we describe the analytical modeling framework and then the numerical model.

The Analytical Framework

The presentation of the analytical framework is subdivided into the demand side, the supply side, equilibrium, and social welfare indicators.

Demand Side. We assume a set of potential users for every origin-destination (OD) pair in the network. A potential user can contribute to the demand for only one OD-pair at most, a restrictive assumption in general, but probably more acceptable in the context of commuting; most people have only one residential and one work location. However, the model does not account for group switching; a user cannot switch to another OD-pair due to policy intervention. An individual's demand function for road use, conditional on car

ownership, is not perfectly inelastic. If the generalized price of road use (including monetized travel times and variable taxes) increases, the number of road trips per unit of time decreases, perhaps reflecting that people would more often take an alternative transport mode, travel outside the peak time, or work from home, but not change residential or work location.[5] Individuals within a group are identical, with the exception of one characteristic, their relative inclination to road use. This heterogeneity is introduced to prevent car ownership as a function of policy variables from becoming constant and positive over a certain range and constant and equal to zero outside that range.

For simplicity, we assume that an individual i's inverse demand function for road use is linear

$$D_i = \bar{d} - d_i \cdot q_i \tag{1}$$

where subscripts for OD-pairs are suppressed, q_i refers to the equilibrium number of trips demanded, $-d_i$ gives the slope of the individual's inverse demand function that varies across individuals to reflect heterogeneity in terms of the inclination to road use, and $\bar{d}$ gives the intercept assumed to be equal across individuals for simplicity. Heterogeneity is thus such that if $d_i = 0.5 \cdot d_j$, individual i consumes twice as many trips as individual j for all generalized cost levels, as long as both own a car. From Equation (1), it is straightforward to derive that with a generalized price level for road use p, the consumer would choose

$$q_i = \frac{\bar{d} - p}{d_i}$$

and would hence enjoy a "gross" consumer surplus (not accounting for the costs of vehicle ownership) of

$$CS_i^G = \frac{1}{2} \cdot q_i \cdot (\bar{d} - p) = \frac{(\bar{d} - p)^2}{2 \cdot d_i} \tag{2}$$

The price level p reflects the price associated with the use of the road conditional on car-ownership. It is assumed to be equal to the sum of the monetized travel time, the travel time t multiplied by the common single value of time vot, plus link-based congestion charges τ encountered, plus the kilometer charge τ_{km} multiplied by the length of the trip. If the unit of distance is set equal to the distance traveled in one unit of time without congestion so that the trip length becomes equal to the free-flow travel time t^{fftt}, we can write

$$p = vot \cdot t + \tau + \tau_{km} \cdot t^{fftt} \tag{3}$$

[5] All such alternatives are implicitly assumed to be efficiently priced.

Note that we define t, τ, and t^{fftt} over the entire trip in Equation (3), and thus ignore route choice and summations over links to avoid notational clutter.

In addition to the variable price p, a traveling individual will incur fixed costs due to the ownership of a car p_f. This is the sum of per-unit-of-time resource costs c_f and the fixed "annual" ownership tax τ_f ("annual" in quotation marks, because the numerical model is calibrated so as to describe a single morning peak). These are equal across individuals

$$p_f = c_f + \tau_f$$

We assume that individual i will own a car, indicated by the dummy δ_i taking the value of 1, if the gross consumer surplus from its use in Equation (2) is at least equal to p_f

$$\delta_i = \begin{cases} 1 \text{ if } CS_i^G \geq p_f \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

We also assume that the vehicle is used exclusively for traveling in the peak and on the network considered. The net consumer surplus for individual i can then be written as

$$CS_i^N = \begin{cases} CS_i^G - p_f & \text{if } \delta_i = 1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Therefore, an individual's net and gross consumer surpluses are zero when he or she prefers not to own a car, consistent with the fact that these surpluses derive from road use. The marginal willingness to pay for road use D_i gives for every trip the surplus that would be enjoyed from costless road use at $p = 0$ above the surplus that would be enjoyed from the most preferred alternative. We can ignore surpluses from alternatives such as public transportation, cycling, and other modes in the welfare analyses if we assume that these alternatives are optimally priced. We make that simplifying assumption.

From Equations (4) and (2), the critical value of d_i, d^*, can be derived for which individual i is indifferent between not owning a car and owning one and using it optimally

$$d^* = \frac{(\bar{d} - p)^2}{2 \cdot p_f}$$

Under these assumptions, the existence of a continuous and smooth aggregate inverse demand function requires that users form a continuum defined over d_i. The type of density function assumed for d_i will have potentially significant effects for the results in the numerical model. One way of selecting an appropriate density function recognizes explicitly that each type will imply a different ratio between what can be called the "short-run aggregate demand elasticity" with respect to p, holding car ownership fixed, and the "long-run aggregate demand elasticity," accounting for changes in car ownership. A plausible constant ratio of 2 between these two measures is found for a density function

$$n(d_i) = \nu \cdot \sqrt{d_i}$$

where $n(d_i)$ denotes the density of users with a slope of their individual inverse demand equal to d_i, and ν is a parameter. With this density function, the aggregate demand function can be written as

$$\begin{aligned} Q &= \int_0^{d^*} \nu \cdot \sqrt{d_i} \cdot \frac{\bar{d} - p}{d_i} \mathrm{d}d_i \\ &= \nu \cdot (\bar{d} - p) \cdot \left[2 \cdot \sqrt{d^*} - 2 \cdot \sqrt{0}\right] \\ &= 2 \cdot \nu \cdot \frac{(\bar{d} - p)^2}{\sqrt{2 \cdot p_f}} \end{aligned} \tag{6}$$

implying a long-run elasticity equal to

$$\varepsilon = -4 \cdot \nu \cdot \frac{(\bar{d} - p)}{\sqrt{2 \cdot p_f}} \cdot \frac{p}{Q}$$

The perceived short-run aggregate demand function holding d^* fixed can be derived from the middle line in Equation (6) and reads

$$\tilde{Q} = \nu \cdot (\bar{d} - p) \cdot 2 \cdot \sqrt{d^*} \tag{7}$$

with d^* treated as fixed, implying a short-run elasticity equal to

$$\begin{aligned} \tilde{\varepsilon} &= -2 \cdot \nu \cdot \sqrt{d^*} \cdot \frac{p}{Q} \\ &= -2 \cdot \nu \cdot \frac{(\bar{d} - p)}{\sqrt{2 \cdot p_f}} \cdot \frac{p}{Q} \end{aligned}$$

which is, indeed, half the value of ε.

Last, the inverse aggregate demand function can be found after some manipulation of Equation (6) and reads

$$D = \bar{d} - \frac{\sqrt[4]{p_f} \cdot \sqrt{Q}}{\sqrt[4]{2} \cdot \sqrt{\nu}} \tag{8}$$

Note that Equations (6) and (8) fully capture car ownership decisions. The aggregate consumer surplus measures that can be derived from these functions correspond to the net consumer surplus as defined in Equation (5). This is most easily verified by observing that the aggregate equilibrium consumer surplus that can be calculated from $\tilde{Q}$ in Equation (7)—or from its inverse—is a summation over all users of the gross surplus in Equation (2). The demand function Q in Equation (6) and its inverse D in Equation (8) only incorporate demand from an individual i below a certain price p_i at which his or her gross

consumer surplus is equal to total fixed costs p_f. For any equilibrium, the difference between the consumer surpluses in Equations (6) and (7) is therefore exactly equal to the fixed costs incurred by all users using the road in that equilibrium.

Supply Side. The model considers static, steady-state congestion on a network for which the travel time on a link depends on the equilibrium flow on that link alone and its capacity. There are no direct link interactions such as spill-backs at bottlenecks. Cost functions are consistent with the technical assumptions underlying the self-financing theorem. For link travel-time functions, this is, for instance, the case with the well-known Bureau of Public Roads (BPR) function, which implies that for link l the generalized price is

$$p_l = vot \cdot t_l^{fftt} \cdot \left[1 + b \cdot \left(Q_l \cdot cap_l^{-1}\right)^k\right] + \tau_l + t_l^{fftt} \cdot \tau_{km} = c_l + \tau_l + t_l^{fftt} \cdot \tau_{km}$$

where b and k are parameters typically set at 0.15 and 4, respectively; cap_l is a measure of the link's capacity; Q_l gives the equilibrium use level for the link; and c_l is the generalized travel cost for link l. The generalized price is homogeneous of degree zero in use and capacity as required for a model calibrated to produce exact self-financing in its first-best optimum.[6]

For the cost of providing capacity, constant economies of scale require that the unit price of capacity is constant for a link so that the total cost for a link can be written as

$$C_l^{cap} = cap_l \cdot c_l^{cap}$$

Network Equilibrium. We use the standard deterministic Wardropian network equilibrium concept. This means that in equilibrium, users from a given OD-pair will only use minimum generalized price routes provided the equilibrium generalized price is below the reservation price ($\bar{d}_j$ in Equation (8), with j denoting OD-pairs) and that there are no routes available with a strictly lower equilibrium generalized price level. A general formal mathematical treatment is suppressed as it would duplicate standard expositions as provided in Verhoef (2002), among many others, while introducing much notational clutter. A simple case is presented below.

When using inverse aggregate demand functions as in Equation (8) for every OD-pair, the equilibrium obtained simultaneously describes equilibrium use of the network given car ownership and car ownership given the network

[6]Many alternatives to this classic BPR function have been proposed and investigated in the traffic engineering literature (e.g., Smith, Hall, and Montgomery, 1996 and Singh, 1999). Our choice for the classic BPR function is not based on any belief that it should be the most realistic function. The main advantages are its simplicity, and the fact that it is probably the most often used static speed-flow function in economic studies into road pricing (e.g., Liu and McDonald, 1998; Small and Yan, 2000; among many others), which makes our model in that respect comparable to prior studies. The qualitative conclusions from our analysis do not depend on the specific congestion function used.

equilibrium. Furthermore, the application of the Wardropian equilibrium principle implies that only one single level of p_j will prevail in equilibrium for an OD-pair j. However, the shares of travel time costs and tolls in the equilibrium trip price may of course vary between different routes used for that OD-pair, for example, when a pay-lane and parallel untolled lanes are simultaneously in use.

Social Welfare. Social surplus measures are used for the definition of social welfare. The structure of the model enables a distinction between a number of welfare indicators that we define in this subsection. First, consistent with the previous discussion, one can distinguish for each OD-pair j between gross and net consumer surplus

$$CS_j^G = \int_0^{Q_j} \tilde{D}_j(x) - p_j \mathrm{d}x$$

(9)

$$CS_j^N \equiv CS_j^G - N_j \cdot p_f = \int_0^{Q_j} D_j(x) - p_j \mathrm{d}x$$

where $\tilde{D}_j$ is the inverse of the function $\tilde{Q}_j$ as defined in Equation (7), Q_j is to be interpreted as the equilibrium demand, N_j is the number of car owners for OD-pair j, and p_f is assumed not to vary over OD-pairs.

Note that the net consumer surplus in Equation (9) is related to the long-run inverse aggregate demand function, incorporating car-ownership decisions, in a standard way. This means that, apart of course from the results explicitly pertaining to car ownership, all other results to be derived are robust in the sense that they would also apply with fixed car ownership and inverse aggregate demand functions D_j equal to those used now. There would then be no reason to consider the functions $\tilde{D}_j$ because these would coincide with the functions D_j.

Because all prices in the model could incorporate taxes, Equation (9) implies a "variable" social surplus—meaning that road capital costs are not included—of

$$S^v = \sum_{j=1}^{J} \left(CS_j^N + \phi \cdot N_j \cdot \tau_f \right) + \sum_{l=1}^{L} \phi \cdot \left(\tau_l + t_l^{fftt} \cdot \tau_{km} \right)$$

where ϕ gives the "shadow price of public funds" assumed to be exogenous and constant. This shadow price is equal to unity when tax revenues are valued equally high as consumer surplus and may exceed unity when tax revenues are used to reduce other distortive taxes or set below unity when these revenues are used in an inefficient way. Note that a total number of J OD-pairs and L links is assumed to apply for the network; the total number of potential and active routes—overall, and per OD-pair—need not be specified in this general treatment.

FIGURE 1: The Network Used for the Numerical Model.

Finally, the overall welfare measure to be employed W can now be defined as

$$W = S^v - \phi \cdot \left(\sum_{l=1}^{L} cap_l \cdot c_l^{cap} \right)$$

It is thus assumed that the same shadow price of public funds applies to the capital cost of infrastructure provision, which reflects our assumption that the infrastructure is publicly owned and operated.

A Numerical Model

We use a numerical model to study the research questions identified in Section 1. The model consists of a small static network with three links (1–3) and three OD-pairs linking three nodes (A–C) depicted in Figure 1. Travelers for OD-pairs AB and AC have two routes to choose from, (using either link 1 or 2), while all users for OD-pair BC use only link 3. In all exercises, only equilibria will prevail with both links 1 and 2 used, and with all OD-pairs possessing a positive demand. We assume that a vehicle is owned solely to be used on this small network and during the period that is implicitly described by the static model. Although admittedly unrealistic, the assumption is inherent to the model's static nature and necessary to obtain individuals' vehicle ownership decisions consistent with their behavior on the network.

The equilibrium conditions for such interior equilibria on this simple network read

$$c_1(Q_{AB}^1 + Q_{AC}^1) + \tau_1 + t_1^{fftt} \cdot \tau_{km} - D_{AB}(Q_{AB}^1 + Q_{AB}^2) = 0 \tag{10a}$$

$$c_2(Q_{AB}^2 + Q_{AC}^2) + \tau_2 + t_2^{fftt} \cdot \tau_{km} - D_{AB}(Q_{AB}^1 + Q_{AB}^2) = 0 \tag{10b}$$

$$\begin{aligned} &c_1(Q_{AB}^1 + Q_{AC}^1) + \tau_1 + t_1^{fftt} \cdot \tau_{km} + c_3(Q_{AC}^1 + Q_{AC}^2 + Q_{BC}) \\ &\quad + \tau_3 + t_3^{fftt} \cdot \tau_{km} - D_{AC}(Q_{AC}^1 + Q_{AC}^2) = 0 \end{aligned} \tag{10c}$$

$$\begin{aligned} &c_2(Q_{AB}^2 + Q_{AC}^2) + \tau_2 + t_2^{fftt} \cdot \tau_{km} + c_3(Q_{AC}^1 + Q_{AC}^2 + Q_{BC}) \\ &\quad + \tau_3 + t_3^{fftt} \cdot \tau_{km} - D_{AC}(Q_{AC}^1 + Q_{AC}^2) = 0 \end{aligned} \tag{10d}$$

$$c_3(Q^1_{AC} + Q^2_{AC} + Q_{BC}) + \tau_3 + t_3^{fftt} \cdot \tau_{km} - D_{BC}(Q_{BC}) = 0 \quad \text{(10e)}$$

where Q^1_{AB} identifies the use for OD-pair AB on link 1 and similarly for OD-pair AC and link 2. An equilibrium for given tax levels is found by solving the Equations (10a) through (10e). Because of the linear dependence in the system, a sixth equation $Q^1_{AB}/Q^1_{AC} = Q^2_{AB}/Q^2_{AC}$ is added to distribute travelers from OD-pairs AB and AC proportionally over links 1 and 2.

Table 1 shows the base-case parameters and policy variables. The demand side parameters were set so as to obtain an equilibrium in which links 1 and 2 are relatively heavily congested with travel times just exceeding twice the free-flow travel times (see also Table 2), and link 3 is only mildly congested, while equilibrium short-run demand elasticities are in a plausible range between −0.3 and −0.4.

All links have a free-flow travel time of half an hour. A value of *cap* equal to 1,750 for the BPR cost function implies a doubling of travel times at a use level of around 2,800 vehicles per hour. This is a high-end estimate of the flow at which travel times double for a single highway lane and the maximum flow on a lane is reached (e.g., Small, 1992, Fig. 3.4, p. 66). The latter, however, is not defined for BPR functions. Furthermore, the implied speed at a flow of 2,000 vehicles per hour is 95.5 km/hr, which is slightly above the estimates provided by Smith, Hall, and Montgomery (1996, Tables 2 and 4) that range from 81 to 93 km/hr. The difference appears justified by the fact that the free-flow speeds in their estimates, ranging from 93 to 111 km/hr, are also slightly below the 120 km/hr assumed here.

The hourly unit prices of capacity of € 6 were determined by dividing the estimated average yearly capital cost of one highway lane kilometer in The Netherlands (€ 0.2 million) by 1,100 (220 working days times 5 peak hours per working day, assuming 2 peaks) and next by 1,750 (the number of units of capacity corresponding with a standard highway lane), and finally multiplying by 60 (the number of kilometers corresponding with a free-flow travel time of half an hour). The calibration procedure thus implicitly assumes that small changes in capacities induce no effects on travel times outside the peak hours considered in the model and that off-peak travel absent from the model can indeed be fully ignored when optimizing capacities.

TABLE 1: Base Parameters and Policy Variables

OD-pair AB	*OD-pair AC*	*OD-pair BC*	*Link 1*	*Link 2*	*Link 3*	*General*
$\bar{d}_{AB} = 30$	$\bar{d}_{AC} = 50$	$\bar{d}_{BC} = 17.5$	$t_1^{fftt} = 0.5$	$t_2^{fftt} = 0.5$	$t_3^{fftt} = 0.5$	$vot = 7.5$
$\nu_{AB} = 15$	$\nu_{AC} = 3.5$	$\nu_{BC} = 17.5$	$c_1^{cap} = 6$	$c_2^{cap} = 6$	$c_3^{cap} = 6$	$c_f = 10$
			$b_1 = 0.15$	$b_2 = 0.15$	$b_3 = 0.15$	$\phi = 1$
			$k_1 = 4$	$k_2 = 4$	$k_3 = 4$	
			$\tau_1 = 0$	$\tau_2 = 0$	$\tau_3 = 0$	$\tau_f = 0$
			$cap_1 = 1{,}750$	$cap_2 = 1{,}750$	$cap_3 = 1750$	$\tau_{km} = 0$

TABLE 2: Base Case: Key Characteristics

OD-pair AB	*OD-pair AC*	*OD-pair BC*	*Link 1*	*Link 2*	*Link 3*	*General*
$Q_{AB} = 3{,}379$	$Q_{AC} = 2{,}267$	$Q_{BC} = 1{,}345$	$Q_1 = 2{,}823$	$Q_2 = 2{,}823$	$Q_3 = 3{,}612$	$W = 17{,}909$
$N_{AB} = 1{,}264$	$N_{AC} = 1{,}438$	$N_{BC} = 294$	$p_1 = 7.56$	$p_2 = 7.56$	$p_3 = 4.39$	$C^v = 58{,}616$
$p_{AB} = 7.56$	$p_{AC} = 11.95$	$p_{BC} = 4.39$	$t_1 = 1.01$	$t_2 = 1.01$	$t_3 = 0.59$	$C^f = 29{,}954$
$\varepsilon_{AB} = -0.67$	$\varepsilon_{AC} = -0.63$	$\varepsilon_{BC} = -0.67$				$C^{cap} = 42{,}000$
						$TR = 0$
						$G = -42{,}000$

Note: t_l indicates travel time on link l; C^v the total generalized (variable) travel costs; C^f the total costs of vehicle ownership; C^{cap} the total capacity costs; TR total tax revenues; and G the government budget.

The BPR parameters b and k have their conventional values. A value of time of € 7.5 corresponds to average estimates for The Netherlands. The fixed cost of car ownership of € 10 implies, when the car is used only for commuting, that the yearly fixed costs of capital and depreciation would be € 2,200, which seems a reasonable order of magnitude for an average car. The shadow price of public funds is set at unity in the base case. Finally, all taxes are set equal to zero for the base-case "no toll" equilibrium.

Table 2 summarizes the key characteristics of the base equilibrium. It is clear that the more heavily used links 1 and 2 make up the more congested part of the network; each link has a free-flow travel time of 0.5. Furthermore, the government faces a deficit of € 42,000.

4. SIMULATION RESULTS

This section presents the results of the modeling exercises that we carried out to answer the various questions raised in the introduction. The section starts with a brief discussion of the technical approach followed for finding second-best optima.

An Extended Algorithm for Finding Second-Best Optima in Transportation Networks

Despite its small scale, the model includes eight potential policy variables: three link tolls, a kilometer charge, a fixed vehicle ownership charge, and three link capacities. Most policy regimes addressed below study only a subset of these. Nevertheless, the task of finding a second-best optimum can be cumbersome, particularly when the number of policy variables exceeds two. Due to interactions between the policy variables, a procedure in which each policy variable is optimized given the level of the other variables will not converge to a second-best optimum after one round and may thus become very time-consuming. At the same time, the model's dimensions are too large to allow for the derivation of insightful analytical optimality conditions

that would help in easily finding a second-best optimum. Moreover, these conditions would, of course, be different for each subset of available policy instruments considered.

We handled this problem by developing a general algorithm for finding second-best optima. The algorithm is a generalization of the algorithm presented in Verhoef (2002) that considers the problem of finding second-best optimal toll levels for a congested network under the constraint that not all links can be tolled, which is a special case of the type of second-best problems considered in this paper.

The backbone of the algorithm is formed by the following Lagrangian

$$\Lambda = W + \sum_{r=1}^{R} \lambda_r \cdot constr_r$$

where r denotes the relevant routes defined as routes that for an OD-pair j offer the lowest possible generalized price and hence can be used in the equilibrium in the network during an iteration (see below), λ_r denote route-specific Lagrangian multipliers, and $constr_r$ is a constraint that equates marginal benefits for the relevant OD-pair to the generalized price for the route considered as given by the left-hand sides of Equations (10a) through (10e) for the current network.

The algorithm can then be summarized as follows:

Step 0: set starting values for the available policy variables
Step 1: calculate the network equilibrium given the exogenous values of the available policy variables
Step 2: solve the system of linear equations that is defined by the two sets of equations that follow below for the variables λ_r for all relevant routes and all available policy instruments π_k (note that with R relevant routes and K available instruments, this gives a system of $R+K$ equations in $R+K$ unknowns):

1. $\partial\Lambda/\partial Q_r - \partial\Lambda/\partial\lambda_r = 0$ for all relevant routes r, evaluated in the network equilibrium determined in step 1
2. $\partial\Lambda/\partial\pi_k = 0$ for all policy instruments π that are available for optimization, evaluated in the network equilibrium determined in step 1

Step 3: update the values for the policy instruments by setting π_k equal to a weighted sum of its previous value and the newly predicted value in step 2
Step 4: check for convergence of all available π_k; terminate if convergence is reached, otherwise return to step 1.

The algorithm bypasses the problem that the full problem defines a set of $2 \cdot R + K$ simultaneous nonlinear equations in $2 \cdot R + K$ unknowns (each available π_k, each relevant λ_r, and each relevant Q_r), the solution to which generally

cannot be found with the mathematical software used. The algorithm performed rather efficiently, although the appropriate weighting procedure for "old" and "newly predicted" levels of π_k was not easily determined. A pragmatic trial-and-error approach was employed, where the trade-off concerned speed of convergence on the one hand and instability of the convergence process on the other. Instability was particularly relevant for sets of policy instruments including both taxes and capacities. With weights of 50 percent for newly predicted tolls and 5 percent of newly predicted capacities, most exercises converged within a few minutes on a standard PC. All second-best optima found were subsequently checked by varying all available policy instruments by plus and minus 5 percent, keeping the other instruments at their predicted second-best optimal level. This yielded lower values for the objective in all cases and the relative reductions (objective values in all cases exceeding 99.5 percent of the predicted second-best optimal value) indicated flatness of the objective function around the predicted second-best optimum. Both findings indicate that the algorithm indeed finds second-best optima as required.

First-Best Configuration

Mathematically speaking, the first-best optimum is a special case of the broader set of second-best optima and the algorithm described earlier could be used equally well to identify it. A peculiarity of the first-best optimum is that all multipliers λ_r for the relevant routes will individually be equal to zero (see also Verhoef, 2002). Table 3 provides the key characteristics of the first-best optimum. The table shows that we make at least one unrealistic assumption for the calculation of the first-best optimum that capacity can be adjusted as if it were a continuous variable and that capacity investments are reversible so that all links indeed can obtain a lower capacity in the first-best optimum than in the base case. The latter assumption could have been easily avoided with a different base case. We make this assumption because it enables us to consider the overall optimum for the system. The results suggest making these assumptions is worthwhile.

First, the results confirm that the self-financing theorem applies to full networks. The government budget equals zero in the optimum. As in Yang and Meng (2002), this result holds both at the level of the full network and for every individual link, which cannot be verified in Table 3.

Next, in the optimum, two taxes appear redundant. The first, τ_{km}, is not surprising because any nonzero level could be corrected for by adjusting the link tolls accordingly and obtaining the same network equilibrium. The other redundant tax is τ_f. Provided road use is taxed optimally, prospective car owners face the optimal incentive of purchasing a car when confronted with the resource costs c_f. After paying optimal road taxes for the optimized use level, the marginal user will enjoy a gross surplus that is equal to the resource cost of owning a vehicle. All other users' ownership adds to social welfare; their gross benefit exceeds the sum of the private and external costs of owning

TABLE 3: First-Best Optimum: Key Characteristics

OD-pair AB	*OD-pair AC*	*OD-pair BC*	*Link 1*	*Link 2*	*Link 3*	*General*
$Q_{AB} = 80.1\%$	$Q_{AC} = 62.9\%$	$Q_{BC} = 33.5\%$	$Q_1 = 73.2\%$	$Q_2 = 73.2\%$	$Q_3 = 51.9\%$	$W = 188\%$
$N_{AB} = 71.7\%$	$N_{AC} = 49.8\%$	$N_{BC} = 19.3\%$	$cap_1 = 97.0\%$	$cap_2 = 97.0\%$	$cap_3 = 44.0\%$	$C^{cap} = 70.5\%$
$p_{AB} = 131\%$	$p_{AC} = 166\%$	$p_{BC} = 226\%$	$p_1 = 131\%$	$p_2 = 131\%$	$p_3 = 226\%$	$TR = 29{,}620$
			$t_1 = 65.9\%$	$t_2 = 65.9\%$	$t_3 = 114\%$	$G = 0.00$
			$\tau_1 = 4.93$	$\tau_2 = 4.93$	$\tau_3 = 4.93$	$\tau_f = 0$
						$\tau_{km} = 0$

Note: Percentages denote values relative to the base equilibrium.

and using the car. All nonusers would enjoy a gross surplus from individually optimized use that would fall short of this sum.

Third, the results show that compared to a somewhat arbitrary, but not unrealistic, base case where congestion is not evenly spread over the network and excess capacity in some areas exists, the implementation of joint pricing and capacity policies may increase the generalized price of transport more strongly in the originally mildly congested areas (link 3 in our model) than in the heavily congested areas (links 1 and 2). The reason is that the optimal capacity reduction is greater. When such capacity reduction is not possible in practice, a comparable result would be approached in the long run if demand over the entire network grew steadily over time. The capacity on link 3 would then not be increased for a much longer time, while congestion and optimal tolls would increase over time.

Fourth, an interesting feature is that the optimal congestion tax obtains the same value on all links in the network. With the value of time, the length of the links, and cost of capacity units equalized, this result follows immediately from the constant returns to scale assumption. However, we emphasize that the concern that second-best congestion pricing suffers from the inability of toll differentiation over place may lose its relevance in the long run, provided capacities are optimized throughout the network, congestion occurs throughout the network, and values of time and costs of construction are indeed constant over space (see below).

A final issue worth addressing is that the relative reduction of car ownership exceeds that of road use for all OD-pairs. This can be explained by the fact that the users priced off the road will be those who have the highest value of d_i, hence those with the smallest inclination to use the road. The sensitivity of car ownership is of course (much) greater than what would seem realistic at first. It is therefore important to emphasize the main modeling features responsible. First, the model ignores the possibilities that a vehicle could be used for trips other than for the individual's OD-pair and outside the peak considered. Another feature is that the timeframe considered is the very long run during which car ownership is fully adjusted. This is reflected by the assumption that p_f can be fully saved by terminating car ownership and by the fact that no second-hand car market is modeled. Under these two assumptions, any driver priced off the road will indeed give up car ownership and the result arises because these are the drivers with a relatively low inclination to use the road. Finally, the car ownership effects are magnified by the unrealistic assumption that all tax levels in the benchmark equilibrium are zero. For example, if column 7 in Table 4 were taken as the benchmark with the same road capacities but (optimized) annual taxes in place, a policy change toward first-best regulation would have been found to increase total car ownership by 7 percent instead of the significant decrease reported in Table 3. Combined with the initially low congestion level on link 3—which, given the assumed value of c^{cap} indicates considerable initial overinvestment in its capacity—these factors explain the drastic optimal reductions in car ownership.

TABLE 4: Various Second-Best Optima: Key Characteristics

	(1) $\tau_1, \tau_2, \tau_3, cap_1, cap_2, cap_3$	(2) τ_1, τ_2, τ_3	(3) cap_1, cap_2, cap_3	(4) $\tau_{km}, cap_1, cap_2, cap_3$	(5) τkm	(6) $\tau_f, cap_1, cap_2, cap_3$	(7) τ_f	(8) τ_2	(9) τ_2, cap_2	(10) τ_2, cap_1, cap_2
τ_1	4.93	5.47	0	0	0	0	0	0	0	0
τ_2	4.93	5.47	0	0	0	0	0	2.70	1.47	5.07
τ_3	4.93	1.21	0	0	0	0	0	0	0	0
τ_{km}	0	0	0	9.86	7.48	0	0	0	0	0
τ_f	0	0	0	0	0	8.04	7.27	0	0	0
cap_1	97.0%	100%	134%	97.0%	100%	110%	100%	100%	100%	0%
cap_2	97.0%	100%	134%	97.0%	100%	110%	100%	100%	165%	217%
cap_3	44.0%	100%	66.9%	44.0%	100%	57%	100%	100%	100%	100%
G	0%	35%	100.3%	0%	37%	56%	73%	85%	103%	48%
ω	1	.61	.37	1	.53	.71	.37	.15	.29	.60

Notes: Percentages denote values relative to the base equilibrium; ω is the index of relative welfare improvement, the increase in W compared to the base equilibrium as a fraction of the increase in the first-best optimum.

Some Second-Best Policies

The results discussed above demonstrate that under first-best conditions, the self-financing result carries over to full networks and to situations where car ownership is endogenized, provided prospective car owners make a rational decision whether or not to own a vehicle. The result has potentially far reaching policy implications; it opens the way to an efficient and transparent system of simultaneously financing and regulating road networks. However, as discussed in the introduction, the assumption of first-best policy feasibility may be considered rather hypothetical. The question then arises to what extent the result would apply under different types of second-best regulation. Table 4 shows some key results for various policy scenarios that were investigated to get some further insight into this question. Column 1 repeats the key results for the first-best policies discussed, and the other columns depict those that we consider the most realistic alternative second-best configurations.

Column 2 presents the results for link-specific tolls while keeping capacities fixed. In that case the optimal toll for link 3 is lower than under first-best policies, which is consistent with the fact that congestion will be less severe because capacity will not be reduced. The fact that the tolls on links 1 and 2 are higher than under first-best policies may be surprising because these links' capacities will be reduced under first-best regulation, albeit slightly. The explanation is that under first-best regulation the optimal trip price on link 3 becomes much higher, discouraging trips by users from OD-pair AC and hence reducing congestion on links 1 and 2 compared to the second-best case (2). The next-to-last row shows that with unchanged capacity costs and the implementation of pricing the government deficit will of course shrink by 65 percent. The final row shows that the implementation of pricing alone will lead to an efficiency gain equal to 61 percent of what can be achieved under first-best regulation, indicated with the index $\omega = 0.61$.

Column 3 shows the results for the mirror case, where capacities can be adjusted while tolls are kept equal to zero. The associated second-best capacities exceed those in the base case for links 1 and 2, but fall short of it for link 3, which seems plausible given the relatively heavy initial congestion on links 1 and 2 and the mild congestion on link 3. All capacities exceed first-best capacities. This result is consistent with, but not fully explained by, the observation that a marginal reduction in toll levels from the first-best optimum will lead to an upward adjustment in optimal capacities. The positive direct effect, the reduction in travel costs for the users, will always dominate the negative indirect effect, the negative net social benefits from the induced increase in road use (e.g., Arnott and Yan, 2000; see also Wheaton, 1978; Wilson, 1983; and d'Ouville and McDonald, 1990, on optimal capacity choice under suboptimal pricing). The welfare gain from this policy amounts to 37 percent of first-best gains and the government runs a deficit.

The relative welfare gains from pricing (column 2) and capacity adjustments (column 3), $\omega = 0.61$ and $\omega = 0.37$ respectively, nearly sum up to unity. This suggests nearly perfect additivity of welfare gains from both policies for the numerical model as opposed to sub- or superadditivity, where the gains from joint implementation would be below or above the gains from implementation in isolation. Nevertheless, if capacity choice is irreversible, the nearly perfect additivity of course does not mean that there would be little lost when the policies are implemented sequentially rather than simultaneously. Apart from the obvious fact that a sequential implementation implies suboptimal welfare gains before both instruments are optimized, the maximum welfare gains from capacity adjustments in isolation require these to be set at levels above optimal levels. Therefore, it is typically not optimal to set capacities at second-best optimal levels as indicated in column 3 in Table 4 if subsequent pricing is anticipated. An advantage of starting with pricing would be that prices can be adjusted more easily than capacities at least from a technical viewpoint.

One of the unanticipated results concerns the second-best policy described in column 4 in which capacities can be adjusted and a flat kilometer charge can be applied. This policy results in a first-best efficiency gain and the government budget is perfectly balanced. As explained above, with the value of time and unit costs of capacity equalized over the network, speeds and first-best charges per kilometer will also become equalized over the network at least insofar that speeds are homogeneous of degree zero in use level and capacity. Indeed, one can easily confirm that the value of τ_{km} in column 4 implies link tolls equal the values of τ_1, τ_2 and τ_3 in column 1; the length of each link is 0.5. In contrast, columns 5 and 2 illustrate that if capacities cannot be adjusted and only τ_{km} can be set, efficiency will typically fall short of that under differentiated tolls, $\omega = 0.53$ versus $\omega = 0.61$. Thus a technical or political constraint on differentiation of per kilometer tolls need not lead to efficiency losses, provided capacities can be optimized.

This equivalence breaks down whenever values of time or capacities are not constant over the network. An interesting question then becomes how much will be lost by imposing the constraint that the per kilometer charges be equal over the network in such a case. To shed some light on this issue, the simulations were rerun for the case where c_3^{cap} was lowered from € 6 to € 3 (the results are not shown in Table 4). Capacity is thus considerably cheaper in the less congested area, which is realistic because land prices are likely lower. First-best regulation then entails an optimal toll of € 2.83 for link 3, in contrast to € 4.93 under base-case parameters, with a capacity equal to 67 percent of the initial value, as opposed to 44 percent. The optimal tolls for links 1 and 2 remain unaltered at € 4.93, but their optimal capacities now increase to 103 percent of the initial value, as opposed to 97 percent, due to the increased demand for OD-pair AC following the cost reduction on link 3. The variation in optimal toll levels indicates that a flat kilometer charge in combination with capacity adjustments will no longer succeed in replicating the

first-best. However, the relative efficiency of this second-best policy—involving $\tau_{km} = €\ 8.10$, cap_1 and cap_2 at 107 percent of the initial values, and cap_3 at 63 percent—nevertheless realizes 98 percent of the achievable welfare gains under this adjusted parameterization. Therefore, even with capacity costs varying over a network, an inability to differentiate per-kilometer tolls over a network may induce far smaller efficiency losses than anticipated if capacities can be adjusted. Moreover, self-financing may still nearly hold; the government deficit is dwarfed to 3.6 percent of the initial deficit in this run. Of course, the very high relative efficiency may be partly due to the remaining symmetry in the network other than through capacity costs. However, the equal lengths of links is not likely to contribute to this result because a change in the length would proportionately alter the optimal link toll for a given flow and would therefore leave the optimal kilometer charge for that link unaffected. Relative efficiency may however be reduced if, for some reason, cost functions would differ over links.

Columns 6 and 7 in Table 4 consider the effects of fixed annual taxes τ_f with and without simultaneous capacity adjustments. An intuitive result that optimal capacities in column 6 are higher than under first-best regulation is explained by the absence of user charges and the resulting higher equilibrium use levels. For the same reason, the relative efficiency of τ_f alone in column 7, with $\omega = 0.37$, falls short of that of use charges in columns 2 and 5, with $\omega = 0.61$ and $\omega = 0.53$, respectively. Also, with capacity adjustments in column 6, $\omega = 0.71$ is below unity, its value in columns 1 and 4. At the same time, the government deficit remains larger than under use charges. In absence of capacity adjustments, this is the joint result of a lower equilibrium number of car owners and lower average taxes per user; recall that users typically make more than one trip in equilibrium (see Table 2). With capacity adjustments, a third factor is the higher second-best optimal capacity. Therefore, optimized annual taxes appear less attractive than optimized use taxes in our model under the chosen parameterization from both perspectives of efficiency and government finance.

Last, columns 8–10 concern pay lanes in link 2 with and without capacity adjustments. The relatively low second-best optimal and limited efficiency gains when capacity is fixed at 50 percent of the total capacity between nodes A and B and therefore not adjusted, $\omega = 0.15$, agrees with earlier findings (for example, Verhoef, Nijkamp, and Rietveld, 1996; Verhoef and Small, 2004). The low value derives from congestion spillovers created on the untolled parallel road that push down the optimal second-best toll level and the associated efficiency gains. Column 9 shows that when the capacity of the pay lane is also optimized, the associated efficiency gains increase. The capacity in the numerical example is increased by 65 percent and ω nearly doubles to 0.29; however, the government budget worsens. The tolls collected on the pay lane fall short of the additional capacity costs after optimal expansion. Of course, this is not necessarily the case. If the initial capacity coincidentally was nearly optimal, the simultaneous implementation of a second-best toll

and optimization of the capacity would have led to an improvement of the government budget, assuming a positive second-best toll (see Verhoef, Nijkamp, and Rietveld, 1996 for the possibility of negative second-best tolls for pay lanes).

A related question is whether full financing of an entire network is possible when only pay lanes are in place. Our numerical results confirm the intuitive answer that this would generally not be the case and that substantial deficits would be expected. Typically, many drivers will avoid all tolls by using only the free lanes and those who pay a toll are charged below their direct marginal external costs. Nevertheless, full financing may occur coincidentally in exceptional cases. An example could be constructed on the basis of column 9 in Table 4, which shows the final second-best case where both cap_1 and cap_2 can be optimized simultaneously with τ_2. Optimization would lead to a complete removal of link 1 and expansion of link 2 to 108.5 percent of the initial joint capacity of links 1 and 2 together, hence 217 percent of link 2's initial capacity alone. This result nicely illustrates the inefficiency of leaving some parallel capacity unpriced; optimality requires the size of this part of the capacity to be reduced to zero in exchange for increases in priced capacity. Now consider the financing under this regime. The toll revenues on link 2 exceed its capacity costs by nearly 3.5 percent; the toll exceeds the marginal external congestion costs on link 2 because the external costs of users from OD-pair AC on link 3 are partly recovered through the charge on link 2. This suggests that with a very small link 1 and very low unit capacity costs for link 3, a balanced or even a small positive government budget would be theoretically possible under second-best optimal pricing and capacity choice, provided a certain share of users of unpriced links is present on the pay lane and capacity costs for unpriced links are sufficiently low. Although the practical relevance is probably negligible, clearly the theoretical point is that exact self-financing, or even a budget surplus, on a full network with some unpriced links and second-best optimal tolls and capacities is not impossible.

Absence of Congestion on Some Roads

For self-financing under optimal regulation of a full network to hold, it is required that in the optimum congestion exists and hence, optimal tolls are positive on all roads that have positive unit capacity costs. In reality, this may not always be true, for instance, when the assumption of capacity as a continuous variable becomes binding and some minimum capacity exists (e.g., a single lane) for which no congestion occurs in the optimum, while the costs of supplying this capacity are positive. To illustrate the breakdown of self-financing under these circumstances, we ran a simulation in which link 3 was assumed to be uncongested by definition and an arbitrary capacity and associated costs were assigned to this link. As expected, the result was that under first-best regulation self-financing still holds exactly for links 1 and 2, with $\tau_1 = \tau_2 = 4.93$ of course still holding. The optimal τ_3, however, becomes

equal to zero because of the absence of congestion on link 3, yielding a government deficit equal to the assumed capacity costs for the uncongested link 3.

An Above-Unity Shadow Price of Public Funds ϕ

A second reason first-best regulation might not lead to exact self-financing even though the cost functions satisfy the required technical assumptions occurs when the shadow price of public funds ϕ differs from unity. The use of an exogenous ϕ to reflect that tax revenues may be used to lower existing distortionary taxes elsewhere in the economy (on labor, for example) can of course be criticized on various grounds. The precise value of ϕ would depend on exactly how the tax revenues are used. In addition, its value in reality will not be constant. Furthermore, partial equilibrium models such as ours typically ignore many aspects that would affect the true value of ϕ and that may often make it lower than what might be expected on intuitive grounds; relevant mechanisms to consider include the so-called tax-interaction, complementarity, and tax-shifting effects (see e.g., Lindsey and Verhoef, 2001). Such shortcomings are important to bear in mind, but do not mean that an exogenous ϕ could not be used to study the impacts of general tax revenue raising objectives upon optimal pricing and capacity choices for road networks.

If ϕ is set above (below) unity, first-best regulation will lead to a budget surplus (deficit). A second consequence is that the optimal annual tax τ_f will no longer generally be equal to zero. The reason is that the implied second objective of raising (or avoiding) tax revenues as efficiently as possible would typically require all available taxes to be adjusted so as to minimize the overall distortions introduced; however, τ_{km} remains redundant for the same reason as before. The sign of τ_f, however, will not generally be that of $\phi - 1$. An increase in τ_f namely induces two relevant effects. The first is the direct effect of increasing the revenues from those users who remain in possession of a vehicle, suggesting the sign would be the same. The second is the effect of reducing the tax revenues from annual and use taxes from those users who give up car ownership in response to the higher annual tax. This would work in the opposite direction. The latter effect may become particularly important because use and annual taxes can be set simultaneously. Starting from $\tau_f = 0$ and positive use taxes, a marginally lower annual tax attracts more users. Raising the use taxes such that the same individual remains the marginal car owner, approximately requiring a marginal increase in use taxes such that his total tax sum remains unchanged, would mean additional revenues from the nonmarginal users who drive more and therefore pay more in additional use taxes than they gain from the lower annual tax.

Which of these effects will dominate in reality is an empirical matter that largely depends on the elasticity of car ownership with respect to annual taxes. Our numerical model provides an illustration of the possible dominance

of the second effect, which is consistent with the relatively high elasticity of car ownership discussed earlier. At $\phi = 1.1$, we find a modest negative optimal $\tau_f = -0.56$, the budget losses of which are more than compensated for through the optimal use taxes $\tau_1 = \tau_2 = 6.10$ and $\tau_3 = 5.42$ (all use taxes were equal to 4.93 under base parameters).[7] These taxes exceed the marginal external costs (again equal to 4.93 on each link), which is reminiscent of Oi's (1971) result that with heterogeneous consumers, a two-part pricing monopolist can generally increase profits by setting the price per usage above marginal cost. In contrast, optimal capacities are smaller than under base parameters, 89 percent of initial capacities for links 1 and 2 and 40 percent for link 3, which is not only consistent with the lower use levels following the higher use taxes, but also reflects that capacity costs, too, are weighed with ϕ and hence have become higher.

Budget Constraints

Apart from the theoretical objections against the use of a shadow price of public funds ϕ to introduce budgeting issues in the analysis, a practical objection could be that government budgets may often be allocated exogenously to the details of the policies implemented. Therefore, a final point that we would like to address concerns the impacts of an exogenously determined budget, where a balanced budget appears a natural choice, rather than resulting endogenously from the optimization of the available instruments. Clearly, a balanced-budget constraint will not be binding in our model if all instruments can be used; otherwise, one simply finds the unconstrained optimum discussed earlier, given that we set $\phi = 1$. Therefore, we will consider the case where capacities are given at their base-case levels and the regulator can set the use tolls and annual tax to maximize efficiency under a budget constraint. Apart from an equality constraint that the budget be balanced $B = 0$, we will also consider the inequality constraints $B \geq 0$ where no deficit is allowed, perhaps motivated by considerations of intersectoral fairness, and $B \leq 0$ where no surplus is allowed, motivated by considerations of fairness for road users. A convenient way of presenting the main results is by varying c_{cap} such that both cases are covered where an unconstrained optimization through the setting of τ_1, τ_2, τ_3, and τ_f (the levels of which are independent of c_{cap}) would lead to budget surpluses, and where it would lead to deficits. Figure 2 presents the results.

The upper panel shows the optimal levels of the four taxes under the equality constraint and the lower panel shows the efficiency gain relative to that under unconstrained optimization through the setting of τ_1, τ_2, τ_3, and τ_f with capacities fixed as presented in column 2 of Table 4. This index will be

[7] If the demand for road use is sufficiently elastic, it is conceivable that also use taxes would fall if ϕ is increased.

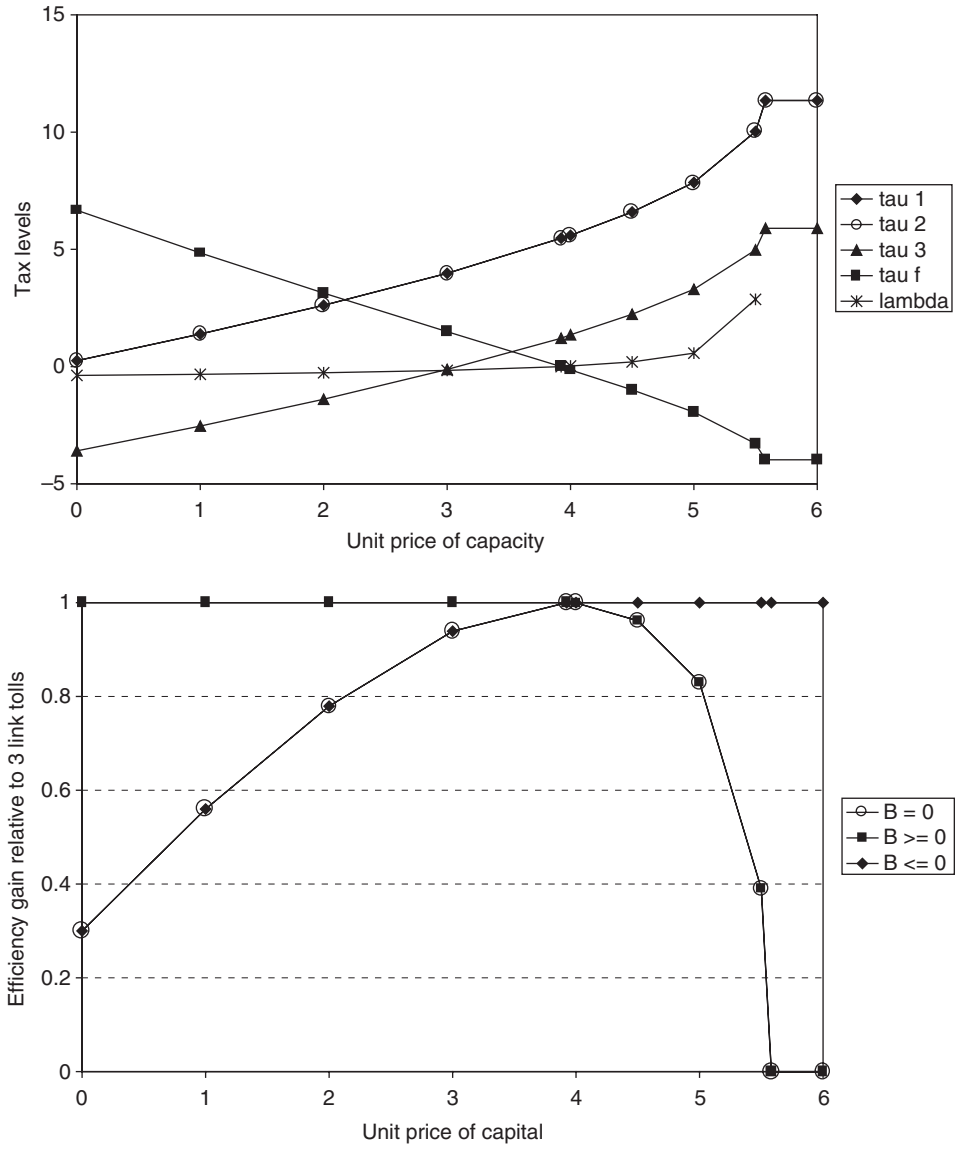

FIGURE 2: Relative Efficiency of Second-best Link Tolls τ_1, τ_2 and τ_3 and Annual Taxes τ_F, for Given Capacities and under Three Types of Budget Constraints.

Notes: At $c_{cap} = 3.924$, optimal use of τ_1, τ_2 and τ_3 (and τ_f optimized at 0) results in a zero budget. Beyond $c_{cap} = 5.581$, B = 0 is not attainable even under profit maximizing use of τ_1, τ_2, τ_3, and τ_f.

indicated with an index $\omega^{\#}$ below and is calculated by applying $\phi = 1$ to government revenues and expenses.

At $c_{cap} = 3.924$, an unconstrained use of τ_1, τ_2, τ_3, and τ_f happens to result in a balanced budget. In this case, τ_f is optimized at zero for the same reasons as in the unconstrained optimum. The concave pattern of $\omega^{\#}$ with the maximum at $c_{cap} = 3.924$ clearly shows that the welfare losses of imposing a balanced-budget constraint rise more than proportionally with the absolute value of the surplus or deficit that would occur under unconstrained pricing. For higher levels of c_{cap}, additional revenues are required to obtain a balanced budget, which is realized most efficiently in our model by subsidizing vehicle ownership (τ_f is negative), while simultaneously raising the link tolls above their unconstrained optimal levels—both to an extent increasing in c_{cap}. The interpretation is the same as for a shadow price of public funds ϕ exceeding unity as discussed earlier. The constraint $B \leq 0$ is not binding in this region so its $\omega^{\#}$ remains equal to unity.

The opposite results are found for $c_{cap} < 3.924$. Noteworthy in this region is that the toll on the relatively uncongested link 3, τ_3, becomes negative for sufficiently low values of c_{cap}. Furthermore, a constraint that total tax revenues should be zero on the left end of Figure 2 does not mean that $\omega^{\#}$ becomes zero. Namely the constraint does not mean that all taxes individually must be equal to zero and some efficiency gains remain possible ($\omega^{\#} = 0.3$).

Finally, on the right end of the diagrams beyond $c_{cap} = 5.581$, a balanced budget is not attainable even under profit-maximizing pricing and the diagrams assume that profit-maximizing pricing is in place in this region regardless of the exact value of c_{cap}. Consequently, if capacities are fixed but are set sufficiently far off optimal levels, self-financing may even become infeasible. Although not easily discerned from the lower panel, profit-maximizing pricing in fact induces a small efficiency loss compared to no pricing in this region; $\omega^{\#}$ has become equal to -0.00013. The closeness to 0 is coincidental.

5. CONCLUSIONS

This paper has developed a simple road network model with endogenous car ownership to study various aspects of the Mohring-Harwitz (1962) result that are relevant for practical application of the principle. A number of findings stand out.

First, optimal per-kilometer congestion tolls and optimal speeds become equal over the network, provided capacities are optimized throughout the network, congestion occurs throughout the network, values of time and costs of construction are constant over the network, and the function that relates travel times to the ratio of use over capacity is the same over the entire network. As a result, a flat kilometer charge in conjunction with optimal capacity policies is capable of reaching the optimum and concerns over the

inability of this instrument to differentiate over space may become less relevant in the longer run when capacities can be optimized. Clearly, from a dynamic perspective, the ability of toll differentiation over time will remain an important requisite for optimal congestion tolling mechanisms. Although a flat kilometer charge becomes truly second-best as soon as unit capacity costs vary over the network, our numerical results suggest that the associated welfare losses may remain very small provided capacities and the kilometer charge remain optimized. A 50 percent reduction in unit capacity costs in the lightly congested area resulted in a reduction of only 2 percent in the efficiency gains from flat kilometer charges compared to differentiated tolls, both with capacities optimized.

As a corollary, a second result of interest is that the implementation of the principle may very well lead to a larger increase in trip costs as experienced by drivers in initially less congested areas than in more heavily congested areas. This may be at odds with intuitive expectations and reflects that often an initially low congestion level may be interpreted equally well as an initial excess capacity.

Third, under second-best pricing on only a subset of links, self-financing no longer necessarily occurs. The numerical results however demonstrate that the efficiency from pay lanes may increase significantly if their capacities are optimized. This is even more evident if the capacities of untolled lanes can also be optimized simultaneously and reductions are possible; however, the consequence is that the untolled lanes are then eliminated and the term "pay lane" would be misplaced. Furthermore, although a government deficit under second-best optimal pay-lane policies is very likely, a balanced budget or even a surplus cannot be excluded on theoretical grounds. However, this requires that a certain number of drivers from an untolled upstream or downstream serial link use the pay lane.

Next, if roads have some minimum technical capacity such as the cheapest possible lane, self-financing under optimal pricing over a full network may break down because some roads with positive capacity costs may have optimal tolls equal to zero. In addition, if capacities cannot be adjusted, imposing a balanced-budget constraint on pricing may lead to substantial efficiency losses that increase more rapidly as optimal short-run pricing implies larger deficits or surpluses.

Finally, unless a nonunitary shadow price of public funds applies, optimal pricing involves a zero fixed ("annual") tax in our model (in which no externalities from car ownership per se, such as through parking, are present). Optimal road pricing provides optimal incentives for car-ownership decisions, provided vehicles are optimally priced. However, when tax-revenue-raising objectives are also relevant, optimal ownership taxes become a relevant instrument, though the objective of raising revenues as efficiently as possible may be served better by a negative annual tax accompanied with increases in use taxes than by a positive one.

REFERENCES

Arnott, Richard, André de Palma, and C. Robin Lindsey. 1993. "A Structural Model of Peak-period Congestion: a Traffic Bottleneck with Elastic Demand," *American Economic Review*, 83, 161–179.

Arnott, Richard and Marvin Kraus. 1998. "Self-financing of Congestible Facilities in a Growing Economy," in D. Pines, E. Sadka, and I. Zilcha (eds.), *Topics in Public Economics: Theoretical and Applied Analysis*. Cambridge: Cambridge University Press, pp. 161–184.

Arnott, Richard and An Yan. 2000. "The Two-Mode Problem: Second-Best Pricing and Capacity," *Review of Urban and Regional Development Studies*, 12, 170–199.

Berechman, Joseph and David Pines. 1991. "Financing Road Capacity and Returns to Scale under Marginal Cost Pricing," *Journal of Transport Economics and Policy*, 25, 177–181.

Bichsel, Robert. 2001. "Should Road Users Pay the Full Cost of Road Provision?" *Journal of Urban Economics*, 50, 367–383.

Button, Kenneth J. and Erik T. Verhoef, eds. 1998. *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility*. Cheltenham: Edward Elgar.

De Borger, Bruno. 2001. "Discrete Choice Models and Optimal Two-Part Tariffs in the Presence of Externalities: Optimal Taxation of Cars," *Regional Science and Urban Economics*, 31, 471–504.

Keeler, Theodore E. and Kenneth A. Small. 1977. "Optimal Peak-Load Pricing, Investment and Service Levels on Urban Expressways," *Journal of Political Economy*, 85, 1–25.

Knight, Frank H. 1924. "Some Fallacies in the Interpretation of Social Costs," *Quarterly Journal of Economics*, 38, 582–606.

Kraus, Marvin. 1981. "Scale Economies Analysis for Urban Highway Networks," *Journal of Urban Economics*, 9, 1–22.

Levinson, David M. and David Gillen. 1998. "The Full Cost of Intercity Highway Transportation," *Transportation Research*, 3D, 207–223.

Lindsey, C. Robin and Erik T. Verhoef. 2000. "Congestion Modelling," in D. A. Hensher and K. J. Button (eds.), *Handbook of Transport Modelling, Handbooks in Transport 1*. Amsterdam: Elsevier/Pergamon, pp. 353–373.

———. 2001. "Traffic Congestion and Congestion Pricing," in D. A. Hensher and K. J. Button (eds.), *Handbook of Transport Systems and Traffic Control, Handbooks in Transport 3*. Amsterdam: Elsevier/Pergamon, pp. 77–105.

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, 352–366.

Mayeres, Inge and Stef Proost. 1997. "Optimal Tax and Public Investment Rules for Congestion Type of Externalities," *Scandinavian Journal of Economics*, 99, 261–279.

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

Mohring, Herbert and Mitchell Harwitz. 1962. *Highway Benefits: An Analytical Framework*. Evanston, IL: Northwestern University Press.

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

———. 1989. "Cost Recovery from Optimally Designed Roads," *Economica*, 56, 165–185.

Oi, Walter Y. 1971. "A Disneyland Dilemma: Two-Part Tariffs for a Mickey Mouse Monopoly," *Quarterly Journal of Economics*, 85, 77–96.

d'Ouville, Edmond L. and John F. McDonald. 1990. "Optimal Road Capacity with a Suboptimal Congestion Toll," *Journal of Urban Economics*, 28, 34–49.

Pigou, Arthur C. 1920. *Wealth and Welfare*. London: Macmillan.

Singh, Rupinder. 1999. "Improved Speed-flow Relationships: Application to Transportation Planning Models," paper presented at the 7th Transportation Research Board conference on application of transportation planning methods, Boston.

Small, Kenneth A. 1992. *Urban Transportation Economics: Fundamentals of Pure and Applied Economics*. Chur, Switzerland: Harwood.

———. 1999. "Economies of Scale and Self-financing Rules with Non-competitive Factor Markets," *Journal of Public Economics*, 74, 431–450.

Small, Kenneth A., Clifford M. Winston, and Carol A. Evans. 1989. *Road Work*. Washington, DC: Brookings Institution.

Small, Kenneth A. and Jia Yan. 2000. "The Value of 'Value Pricing' of Roads: Second-Best Pricing and Product Differentiation," *Journal of Urban Economics*, 49, 310–336.

Smith, W. Spencer, Fred L. Hall, and Frank O. Montgomery. 1996. "Comparing Speed-Flow Relationships for Motorways with New Data from the M6," *Transportation Research*, 30A, 89–101.

Stahl, Dale O. 1981. "Economic Analysis of Transportation Pricing, Tax and Investment Policies," *Transportation Research Record*, 791, 14–20.

Strotz, Robert H. 1965. "Urban Transportation Parables," in Julius Margolis (ed.), *The Public Economy of Urban Communities*. Washington, DC: Resources for the Future, pp. 127–169.

Verhoef, Erik T. 2002. "Second-Best Congestion Pricing in General Networks: Heuristic Algorithms for Finding Second-best Optimal Toll Levels and Toll Points," *Transportation Research*, 36B, 707–729.

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

Verhoef, Erik T. and Kenneth A. Small. 2004. "Product Differentiation on Roads: Second-Best Congestion Pricing with Heterogeneity under Public and Private Ownership," *Journal of Transport Economics and Policy*, 38, 127–156.

Vickrey, William S. 1969. "Congestion Theory and Transport Investment," *American Economic Review*, 59, 251–260.

Wheaton, William C. 1978. "Price-induced Distortions in Urban Highway Investment," *Bell Journal of Economics*, 9, 622–632.

Wilson, John D. 1983. "Optimal Road Capacity in the Presence of Unpriced Congestion," *Journal of Urban Economics*, 13, 337–357.

Yang, Hai and Qiang Meng. 2002. "A Note on 'Highway Pricing and Capacity Choice in a Road Network under a Build-Operate-Transfer Scheme,'" *Transportation Research*, 36A, 659–663.