Transport pricing beyond the social optimum

Verhoef, Erik T.

published in
Handbook on Transport Pricing and Financing
2023

DOI (link to publisher)
10.4337/9781800375550.00009

document version
Publisher's PDF, also known as Version of record

document license
Article 25fa Dutch Copyright Act

citation for published version (APA)
Verhoef, E. T. (2023). Transport pricing beyond the social optimum. In A. Tirachini, D. Hörcher, & E. T. Verhoef (Eds.), *Handbook on Transport Pricing and Financing* (pp. 39-58). (Research Handbooks in Transport Studies series). Edward Elgar Publishing Ltd.. https://doi.org/10.4337/9781800375550.00009

3. Transport pricing beyond the social optimum[1]

Erik T. Verhoef

3.1 INTRODUCTION

Much of the economics literature on transport pricing assumes that the network operator controlling tolls and capacities is what is sometimes referred to as a "benevolent dictator," seeking and imposing policies that are designed to maximize overall social surplus. There are good reasons for making this assumption: it is a construct that helps us identify the efficient outcome in which social surplus is maximized, and in that way identifies a natural benchmark for policies. A sizeable amount of literature on second-best pricing has made abundantly clear that, indeed, the first-best optima that really achieve that benchmark are better not thought of as realistic policy scenarios, since the various constraints that are considered in these analyses of second-best pricing are far from exotic theoretical notions, but instead represent restrictions that policymakers face on a daily basis. Think of, for example, pre-existing distortionary taxes, the inability to differentiate tolls in an optimal fashion over all users, and the inability to price all links of or all users on a network, etc.

Nevertheless, for such second-best pricing problems, it is still typically the objective of maximizing social surplus that is assumed to apply, irrespective of the additional constraints that are considered. This would still match the situation where the network under consideration is in the hands of the aforementioned benevolent dictator, even though this dictator now faces constraints. In reality, however, the possession of and control over networks may be organized differently. In this chapter, we will consider two important practical deviations from the benevolent dictator model and will look at operators who pursue objectives other than the maximization of overall social surplus, investigating the welfare implications relative to the situation considered so far. To maintain focus, we will consider the case of congestion pricing for road infrastructure in the analytic expositions, but where relevant we will also treat other modes in a verbal, qualitative manner.

The first alternative institutional setup that we will consider in fact introduces the fundamental economic question of whether a public authority should set congestion tolls in the first place. As soon as road pricing becomes technically possible, then why not switch to a complete market approach and allow a private company to operate the road? Such privatization of roads is often suggested as an economically attractive way of dealing with congestion. It will, however, be shown why privatization is certainly not a magical cure for inefficiencies in road transport.

The second case of interest concerns the situation where there are multiple governments at stake. In particular, we will discuss the theme of tax competition, which can be expected to arise when different governments represent different populations and tax bases overlap. Because transport by definition involves mobile agents, the chances of tax competition becoming relevant is far from imaginary. At the same time, tax competition is a topic that has a much broader relevance than to only transportation markets. So, the issues we discuss and insights we derive and present are much more widely applicable than for the case of transportation markets alone.

The final topic we address in this chapter concerns the case where a single public regulator is – or feels – constrained to heavily weigh in acceptability next to social surplus in the design and implementation of pricing instruments. Indeed, there is extensive literature assessing the paradoxical situation where policies that maximize social surplus – road pricing – meet with social resistance. An important factor in this debate is the simple observation that people dislike the imposition of new or higher taxes, irrespective of the notion that the receipts will or can be used by the government to finance valuable public goods or lower taxes elsewhere in the economy. This has led to literature on rewards, or subsidies, as well as (budget-neutral) tradable permits, as alternative price-based instruments to manage traffic congestion. These we review in Section 3.5.

3.2 THE FIRST-BEST REFERENCE

While Chapter 2 in this *Handbook* introduces the theory of externality pricing in a mostly graphical way, for this chapter we will complement this with an analytical approach, which is better suited for the exposition of the cases we will be discussing. We begin by discussing the first-best case. The insights will be fully consistent with those from the graphical exposition in the previous chapter, but having the analytical expressions for this case will be helpful when comparing the later cases in this chapter. For the matters of interest in this chapter, it is sufficient to centre the discussion around the simplest possible setup where we consider a static model for a single road that is used by homogenous travellers. Obviously, extensions to more elaborate cases are of great academic and societal interest, but understanding the roles played by factors such as heterogeneity, dynamics, uncertainty or network interaction becomes easier if we have a basic case fully developed and understood.

3.2.1 Optimal Road Pricing for Congestion Externalities

We begin by specifying the objective for the first-best case, which we assume will be social surplus S defined as the difference between total benefits B from the consumption of road trips and the total costs involved. For completeness, we distinguish two types of cost: variable generalized user cost C, which includes the valuation of travel time and therefore has an upward-sloping average cost when congestion prevails, and the cost of supplying road capacity C_{cap}, which is assumed to be given in the short run:

$$S = B - C - C_{cap} \tag{3.1}$$

The Marshallian benefit measure that is used to determine social surplus is intimately related to the Marshallian demand function (which for a given income depicts the relation between market price and quantity demanded), and follows as the integral of inverse demand $D(N)$:

$$B(N) = \int_0^N D(n)\mathrm{d}n \tag{3.2}$$

where N represents the number of travellers and D the marginal willingness to pay (in that way, D is indeed the inverse of the demand function that gives quantity N as the function of generalized price p). In our derivations, it will be convenient to work with the average and also the marginal user cost functions, *ac* and *mc*, which by definition satisfy:

$$C(N) = N \cdot ac(N) = \int_0^N mc(n)\mathrm{d}n \tag{3.3}$$

Note that Equation (3.3) assumes that there are no fixed user costs: with zero travellers, there is no user cost. If fixed user costs were relevant, the analysis would not change, but the interpretation of C would become the total variable user cost. Taking the derivatives of each of the three terms in Equation (3.3), we get another triple equality:

$$C'(N) = ac(N) + N \cdot ac'(N) = mc(N) \tag{3.4}$$

The longer middle expression is central to understanding the essence of congestion pricing. It reflects that when the next user enters a congested road, two types of user costs are created. The first concerns the costs borne by the new user. These are simply equal to the prevailing average costs on the road. Note here that individual road users are treated as infinitesimally small so that average costs can for this purpose be simply evaluated at quantity N. Second, however, the travel time for all other users will increase. And this part of the marginal cost causes this to differ from the average cost. Because these additional costs are not borne by the person who creates them – the last user added – these costs constitute the *marginal external cost* (*mec*) of a trip:

$$mec = mc - ac = N \cdot ac' \tag{3.5}$$

The expression on the right gives the natural interpretation of the marginal external cost: the number of other road users (N), multiplied by the impact of a marginal user on average (per user) cost (*ac*).

As already explained in Chapter 2 in this *Handbook*, the optimal road price in this case is equal to this marginal external cost. To see why, we can compare the first-order condition for optimizing Equation (3.1):

$$\frac{\partial S}{\partial N} = D(N) - ac(N) - N \cdot ac'(N) = 0 \tag{3.6}$$

with the equilibrium condition that with road pricing, the equilibrium is at the quantity where the marginal willingness to pay equals the generalized price or the sum of average cost plus the road price:

$$D(N) - ac(N) - r = 0 \tag{3.7}$$

Subtracting Equation (3.7) from Equation (3.6) shows us that the optimal road use in our model can be realized by setting an optimal congestion charge, equal to the marginal external cost in the optimum:

$$r^* = N^* \cdot ac'(N^*) \tag{3.8}$$

where superscripts * reflect that we are evaluating at the optimum (we will be dropping that reminder from this point onwards for brevity). Equation (3.8) tells the same story as the diagram discussed in Chapter 2 in this *Handbook*: the optimal road price in this setting is equal to the marginal external congestion cost evaluated in the optimum, consistent with the insights from Pigou (1920). If we had included other external costs in the objective in Equation (3.1), notably environmental externalities, by the same logic their marginal value would have become a second component in the optimal road price in Equation (3.8): it would have been part of the first-order condition in Equation (3.6), but not of the equilibrium condition in Equation (3.7), and therefore would have entered the difference between the two on the right-hand side of Equation (3.8).

Since the road capacity and the costs involved in supply are fixed in the short run, it does not enter our optimal pricing result directly. Indirectly, the choice of capacity does affect the optimal road price: it affects the curvature of $ac(N)$, and the conventional operationalization of capacity in this type of model is such that a larger capacity is associated with a lower and flatter ac function, meaning also that mc will be lower and flatter, so that the optimum would involve a larger N and typically a lower r when capacity is increased. This is the indirect effect: the equilibrium value of r will change, but its optimality condition (or: the optimal tax rule) in Equation (3.8) remains the same.

The reader familiar with the literature on the dynamic pricing of traffic bottlenecks (Vickrey, 1969; Arnott, de Palma and Lindsey, 1993) may wonder whether the tolls for the bottleneck model can somehow be reconciled with the Pigouvian toll in Equation (3.7). The answer is affirmative, but the reasoning behind it may not be immediately straightforward. It rests on the notion that in the model's first-best optimum, the time-invariant generalized price is equal to the time-invariant marginal cost of adding one more user to the dynamic equilibrium. The generalized price consists of two time-varying components: the schedule delay cost, which is the only resource cost that a user incurs as travel delays are eliminated in the optimum, and the time-varying toll, which is not a resource cost but a transfer. This toll is therefore equal to the difference between the cost that is privately incurred, and the marginal cost. And that difference, indeed, is the time-varying marginal external cost.

3.2.2 Optimal Capacity Choice

For the long-run problem, road capacity also becomes a variable of choice; we will call it *cap*. This requires the addition of capacity as an argument in the two cost components in our objective in Equation (3.1). Since the first-order condition in Equation (3.5) and the equilibrium condition in Equation (3.6) will remain unaltered (these are partial derivatives), also the optimal tax rule in Equation (3.8) remains the same. That is, for completeness, we should add the argument *cap* also in the optimal tax rule, to remind us that the derivative of ac with respect to N also depends on *cap*:

$$r = N \cdot ac'\left(N, cap\right) \tag{3.8'}$$

where for convenience we keep using the prime to denote the partial derivative with respect to N. We do, however, get a second optimality condition, namely for optimal capacity choice:

$$\frac{\partial S}{\partial cap} = -N \cdot \frac{\partial ac(\cdot)}{\partial cap} - \frac{\partial C_{cap}}{\partial cap} = 0 \tag{3.9}$$

The interpretation is quite straightforward: the marginal benefits of capacity expansion in the first term of the middle expression consists of a decrease in average user cost multiplied by the number of users. It should be equal to the marginal cost of capacity expansion: the second term, after the minus sign.

A perhaps somewhat unexpected result that we will show below for private road supply is that the profit-maximizing choice for optimal capacity boils down to the same first-order condition as the one shown in Equation (3.9). This does not mean that the road capacities chosen for social welfare maximization and profit maximization are the same: the different choice of toll affects the resulting value of *N,* at which Equation (3.9) is to be evaluated. Nevertheless, the exact match of the first-order condition remains remarkable.

Equipped with Equations (3.8') and (3.9), we can also derive the famous Mohring-Harwitz (1962) result on the self-financing of roads, which states that under certain technical conditions, the revenues from optimal road pricing – according to Equation (3.8') – are just sufficient to cover the cost of the optimal supply of capacity – according to Equation (3.9). The technical conditions can be represented as follows: (i) capacity can be defined in units such that the congestion technology exhibits constant returns to scale: changing capacity and road use in the same proportion leaves travel times and therefore average generalized user cost untouched; (ii) for this definition of capacity, the supply of capacity takes place against neutral-scale economies: C_{cap} varies in fixed proportion with *cap*; and (iii) capacity can be supplied in continuous increments so that its value can be optimized in a marginal sense. The mathematical derivation of the result is provided in various textbooks, including Small and Verhoef (2007). The intuition is pretty straightforward.

First, observe that in the long-run optimum the short-run average user cost, so in the first-best short-run optimum treating the capacity as given, must be equal to the long-run average user cost. Second, observe that the assumed scale-neutrality of road supply in the assumptions in Equations (3.1) and (3.2) above means that the optimized amount of capacity varies in perfect proportion with the optimized number of users. The result is that the long-run average (per user) capacity cost is constant. Third, observe that, as a result, the long-run average user cost is also constant. Also the sum of these two long-run average cost components is therefore constant; and therefore so is the long-run marginal cost, as it is equal to that same sum. The self-financing result then follows from the notation that both the short-run and the long-run marginal cost functions intersect with inverse demand at the same point in the long-run optimum. The first observation above then implies that the toll must be equal to the average (per user) cost of capacity. And this means that, after multiplying both by the number of users, we find that the total toll revenues will be equal to the total cost of capacity.

Empirical evidence suggests that the required scale-neutrality for exact self-financing may hold at least approximately in a range of circumstances (Small, 1992). The condition of the continuity of capacity typically does not hold for a single road, because the number of lanes is discrete. But capacity can still be varied by widening lanes, or by resurfacing. And at the scale of a road network, capacity may be almost perfectly divisible.

The theorem may thus be highly relevant for practical policymaking. First of all, application of the theorem would help in achieving an overall efficient road system, both in terms of capacities and in terms of pricing. Second, the need to raise tax revenues from other sources for the financing of roads may be firmly reduced. Efficiency may then increase even further, when these other taxes are distortionary. Moreover, problems with the public acceptability of road pricing may be reduced, because the resulting scheme may be perceived as "fair" (only the users of a road pay for the capacity, but do not pay anything more than that) and transparent (there are no "hidden" transfers surrounding the financing of roads). And finally, application of the theorem may lead to improved transparency in political decisions on infrastructure expansion. It can easily be demonstrated that if the technical assumptions are fulfilled, the capacity of a road should be expanded whenever currently optimal congestion pricing yields revenues per unit of capacity that exceed the unit (capital) cost of capacity. The market would thus indicate whether or not expansion is socially warranted, which will generally help improve the transparency and credibility of cost–benefit analyses.

Up until this point, we have assumed that the network operator controlling tolls and capacities is what is sometimes referred to as a "benevolent dictator," seeking and imposing policies that are designed to maximize overall social surplus. As stated in the introduction to his chapter (Section 3.1), there are good reasons for making this assumption: it is a construct that helps us identify an efficient outcome, in which social surplus is maximized. Nevertheless, in reality, the possession of and control over networks may be organized differently. In this chapter, we will consider two important practical deviations from the benevolent dictator model and will look at operators who pursue objectives other than the maximization of overall social surplus, investigating the welfare implications relative to the situation considered so far.

3.3 CONGESTION PRICING BY A PRIVATE OPERATOR

The consideration of public roads is in accordance with the practice in most modern societies. Nevertheless, the possibility of private road tolls is receiving increasing attention. One practical motivation for this is simply the lack of sufficient public funds for (rapidly) financing the additional road capacity that some people consider necessary for securing an acceptable level of accessibility to major economic centres. A more fundamental motivation might be that roads have traditionally been supplied publicly because they satisfied the theoretical conditions for *public goods*: non-rivalness and non-excludability. A road's capacity is *non-rival* in consumption before congestion sets in, and it is practically *non-excludable* in consumption unless a toll is charged – which would, however, be inefficient at a zero congestion level (at least, if we ignore other external costs of road transport such as pollution, noise annoyance and accidents). Clearly, when congestion becomes relevant, a road can no longer be classified as a pure public good, and this raises the question of whether it would not be appropriate to privatize the road, and let the free market find its optimum. Surely, a private operator would like to see revenues from its investment in road capacity, and one may wonder whether the required tolls would optimize congestion just as a public toll would do.

3.3.1 Revenue-Maximizing Pricing

To compare private tolling to public tolling, it is instructive to start with the same situation as we did when studying public congestion tolls, namely with a single road of a given capacity. Recall that under these conditions, the public regulator would set a toll equal to the marginal external cost in the optimum as given in Equation (3.8). Would a private operator set the same toll, and thus reproduce the same efficient equilibrium? We can answer this question only after specifying the objective that the private operator would pursue. For a fixed capacity, and hence with fixed costs for the private operator, it is to be expected that the objective would be to maximize total toll revenues, because this will maximize total profit when all costs are fixed. If we denote the private operator's toll with r_p, total revenues R can simply be written as:

$$R = N \cdot r_p \tag{3.10}$$

If N were given because demand was completely inelastic, the solution to the problem of maximizing total revenues would simply be to set the toll infinitely high. Unfortunately – for the private operator – demand will in general not be perfectly inelastic. A higher toll may increase the revenues per road user, but at the same time reduces the number of road users. One way of finding the revenue-maximizing toll under such circumstances is to incorporate the road users' responses to different toll levels in Equation (3.10). We do this by substituting the condition that the toll must be equal to the difference between marginal benefits ($mb = D$) and average cost (ac) in the equilibrium with tolling. To remind ourselves that both D and ac will depend on N, we therefore rewrite (3.10) as:

$$R = N \cdot \left(D(N) - ac(N)\right) \tag{3.10'}$$

The problem faced by the private operator is then to find that particular N for which Equation (3.10') is maximized. This means that we should determine its derivative with respect to N, and set it equal to zero:

$$\left(D(N) - ac(N)\right) + N \cdot \left(D'(N) - ac'(N)\right) = 0 \tag{3.11}$$

Equation (3.11) was found by using the so-called "product rule" of differentiation.[2]

The first term between large brackets will in equilibrium be equal to the toll, and by taking the second term to the right-hand side we can thus find that the revenue-maximizing toll should satisfy:

$$r_p^* = N^* \cdot ac'(N^*) - N^* \cdot D'(N^*) \tag{3.12}$$

(where * now indicates the optimum from the private operator's perspective).

A comparison between Equations (3.12) and (3.8) reveals an interesting result. The private operator in fact does internalize the congestion externality (the first component in the revenue-maximizing toll represents *mec*), but adds to this a positive term that is related to the slope of the demand function. (In Equation (3.12), note that D' itself is negative – it is the slope of the inverse demand function – so that the entire second term, including the minus sign shown, is positive.) The appearance of this second term is easily understood once we realize that the private operator is in fact a monopolist supplying road capacity. The second term then corresponds with our micro-economic intuition that the monopolistic price increases as demand becomes less elastic; or: as the absolute value of D' increases and hence the inverse demand function becomes steeper. Interestingly, the profit-maximizing toll for a congested bottleneck would be consistent with what is found in Equation (3.12), and is equal to the time-varying component that also applies for the public operator, plus a time-invariant demand-related markup that is equivalent to the one in Equation (3.12).

But why would the private operator also internalize the congestion externality? The intuitive explanation is that the internalization of the externality simply exchanges time losses on the road for toll revenues. And time losses do not add to the private operator's profits, whereas toll revenues do. A more precise explanation is, however, that the term appears because the overall demand elasticity faced by the monopolist – which is the crucial variable in the determination of the profit-maximizing price – is in this case not only determined by the slope of the demand function, but also by that of the average cost function. And the required correction in the standard monopolistic pricing rule introduces exactly the *mec* ($N \cdot ac'$) in the monopolist's optimal toll expression, in addition to the demand-related term $N \cdot D'$.[3]

Equation (3.12) thus shows that a private operator would typically charge a toll above the optimal level, unless demand is perfectly elastic and D' is equal to zero. Privatization of roads may therefore indeed lead to tolls that internalize the congestion externality, but they typically do more than that because of the introduction of monopolistic power. This means that tolls will be charged that lead to an over-reduction in road use – just as monopolists in general would have an incentive to charge excessively high prices.

3.3.2 Profit-Maximizing Capacity

We now turn to the case where the private operator would not only set a profit-maximizing toll, but also a profit-maximizing capacity. The objective then should include the capacity cost, and becomes:

$$\Pi = N \cdot \left(D\left(N\right) - ac\left(N, cap\right)\right) - C_{cap} \tag{3.13}$$

It is now easily verified that the first-order conditions mimic Equation (3.11) for the optimal toll, and Equation (3.9) for the capacity. Again, the toll rule does not change when we move from a short-run to long-run analysis, and the interpretation is as before. Probably more surprising is that the investment rule is the same as for the welfare-maximizing problem. As explained above, because the toll is different this does not imply that the resulting capacities will be the same; on the contrary, that will only be the case when the two tolls are the same and that only occurs in the theoretical situation where demand is perfectly elastic. But why are the capacity rules the same? The interpretation rests on the observation that for any traffic volume N that the firm chooses, the toll it can charge varies on a euro-by-euro basis with the average user cost, as their sum should equal to a marginal willingness to pay at that level of N. Reductions in average user cost therefore translate directly and fully into increases in average revenue. And that means that whichever N the firm chooses, profits will be maximized if the sum is minimized for capacity cost, borne by the firm, and user cost, borne by the user but translating fully into foregone revenue for the firm. For the welfare-maximizing problem in Section 3.2, the task of minimizing the sum of those two cost components, given the level of use chosen, is exactly the same.

3.3.3 Franchising of Private Road Infrastructure

Given that the pricing and capacity choice of private operators is generally not in line with social optimality, a relevant question is whether governments should try to intervene through direct regulation of tolls and/or capacities, or whether behaviour can be steered towards more desirable outcomes through the design of auctions that are used to select the firms that are allowed to operate certain toll roads. Contributions by Engel et al. (1997) and Verhoef (2007) consider the efficient design of such auctions from different perspectives and find that this design indeed can have important implications, which is not a surprise for those familiar with auction theory. Engel et al. (1997) focus on ways to avoid renegotiations under demand uncertainty; see also their contribution in Chapter 16 of this *Handbook*. Verhoef (2007) focuses on the optimization of capacity and toll through auction design and shows that, in an otherwise unpriced network and when the scale-neutrality conditions underlying the Mohring-Harwitz result are satisfied, a competitive auction that assigns the franchise to the bidder that offers the highest traffic volume in fact replicates the second-best optimal road in terms of capacity and toll.[4] At the same time, auctions based on other selection criteria, even when sounding plausible at first sight, may be far from the second-best optimum.

3.3.4 Heterogeneous Preferences

There is an important, somewhat hidden consequence of our treatment of profit-maximizing pricing in the context of a simple homogeneous-preferences model. As shown by Edelson (1971) and Mills (1981), while for the welfare-maximizing toll it is the average value of time that matters, for the optimal congestion toll, when there is heterogeneity in the values of time, the congestion component of the profit-maximizing toll will then depend entirely on the value of time of the marginal user(s). One could then even construct examples where the public toll would exceed the private one. This underlines that a profit-maximizing supplier cares about congestion only because it affects the marginal willingness to pay for using the road, and – unlike the welfare-maximizer – not because infra-marginal users are affected.

3.3.5 Networks

Another context in which the different incentives for profit-maximizers versus welfare-maximizers become clearly visible is through their treatment of the utility of travellers who are not subject to the toll, but who are affected because of network spill-overs. An instructive example is the classic two-routes problem, where only one of two parallel lanes is priced so that the toll would, through the diversion of traffic, have the undesirable by-product of intensifying congestion on the untolled lane. While the public operator would set the second-best toll below the marginal external cost on the tolled lane in an attempt to optimally balance the reduction of congestion on the tolled link and its increase on the untolled link, the profit-maximizer would still fully internalize the marginal external cost on the tolled link and add a demand-related markup to that. The markup would be smaller than what is shown in Equation (3.12) though, because the relevant demand sensitivity is enhanced compared to the term D' because of the availability of the untolled link. But the markup would remain non-negative (e.g. Verhoef, Nijkamp and Rietveld, 1996).

3.3.6 Oligopolistic Markets and Congested Facilities

When more than one firm is present, competition of course affects the market outcomes. One example would be when different operators control different links in the same network. Consistent with the findings of Economides and Salop (1992) on the economic effects of integration in network markets, it appears that the way in which different firms are organized in the network is crucial for the equilibrium welfare impacts of their uncoordinated pricing strategies. In a rather stylistic setting, where the welfare impacts are considered by increasing the number of symmetric firms controlling a single corridor in either a purely parallel fashion versus a purely serial fashion, Small and Verhoef (2007) show that for the parallel case, the equilibrium toll would asymptotically approach the socially optimal value as market power evaporates. The intuition, when comparing the outcome with the monopolistic toll of Equation (3.12), is that for each individual firm the demand appears to become more price sensitive with the smaller the portion of capacity it controls. In contrast, for the serial case, the resulting equilibrium toll summed over all operators approaches the prohibitive level at which demand is reduced to zero – a move away from Equation (3.12) in the "wrong" direction from the welfare perspective. The mechanism is similar to that of "double marginalization": each individual firm has the incentive of fully internalizing the congestion externality of the full trip, and each firm applies a markup rule based on the sensitivity of the aggregate demand, so that the same markup – already too high when charged just once – will be charged multiple times.

A quite different setup arises when multiple firms jointly use the same congestible facility. The classic example concerns the joint use of a congested airport by multiple airlines. For example, Brueckner (2002) has shown how in such cases an operator has an incentive to internalize congestion within its own firm, with the consequence that congestion tolls for firms would be lower than the standard toll in Equation (3.8), as only effects on other firms and their passengers remain to be internalized. Chapter 11 in this *Handbook* discusses the literature on airport congestion pricing in detail.

3.3.7 Price Discrimination

A final issue concerns price discrimination. A first intuition may be that when market power leads to welfare losses due to inefficient pricing, a further exploitation of market power through price discrimination would most likely aggravate these losses. This, however, is typically not the case. In fact, under the extreme situation of first-degree price discrimination, where every unit can be sold against the relevant consumer's marginal willingness to pay, the monopolist has the incentive to expand output up to the point where marginal benefit equals marginal cost – which is also the necessary condition for the optimization of social surplus. The reason is that the possibility of price discrimination implies that the price for infra-marginal consumers does not have to be lowered in order to capture the marginal consumer into the market. Marginal revenues then coincide with marginal benefits. Of course, while the resulting equilibrium is efficient, the distributional implications will often be considered undesirable, as the entire consumer surplus is skimmed off and is turned into profits for the firm. In cases where undifferentiated marginal cost pricing would result in losses because of economies of scale and scope, price discrimination may be more acceptable as it allows coverage of total cost while limiting the distortive consequences of pricing above marginal cost. For instance, for public transport, this may be an interesting possibility, although the required sophistication in the price discrimination may well have to go beyond what is possible with more traditional types of travel passes and season tickets (e.g. Hörcher et al., 2018, find these to be rather inefficient in the face of crowding externalities). In other cases, however, the distributional implications will often be considered as undesirable.

For more realistic types of price discrimination, notably second-degree where prices are nonlinear and third-degree where prices can only be varied between groups of consumers, it is not the case that price discrimination would always increase social welfare, but it may very well still do that. Varian (1985), for example, argues that a necessary condition for this to be true is that output increases under price discrimination. Specific to transport, Wang, Lindsey and Yang (2011) establish that second-degree profit-maximizing road pricing, with an access fee on top of a use fee, increases profit and may increase or decrease social surplus, compared to use-based profit-maximizing tolls. This confirms that the welfare impact of allowing price discrimination by profit-maximizers may both be positive and negative, depending on the circumstances.

This issue is of importance for transport, as price discrimination is quite common and is likely to become only more important when online marketing gives transport operators increasing knowledge of individuals' preferences. For public transport, customers can, for instance, easily be separated by age, which suggests that discounts for the elderly – with a typically more elastic demand for public transport – may be part of a profit-maximizing strategy rather than just a friendly gesture. Because transportation products are not storable, it is often also possible to separate by trip motive: inelastic demand by commuting peak travellers can easily be exploited by using discount tariffs outside the peak – although of course not every price differential between peak and off-peak tickets needs to reflect genuine price discrimination, as the marginal cost may also differ when capacities are binding in the peak and non-binding outside the peak. And the use of multiple classes in trains also allows price discrimination. Surely the marginal cost may again differ when first-class seats offer more convenience. But a derived benefit for the profit-maximizing firm is that it will typically be business travellers and higher income groups that will choose first class, and both groups can be expected to be relatively price insensitive. Ticket prices can thus be differentiated more strongly between classes than what would be appropriate on the basis of cost considerations alone.

Also, in aviation, price discrimination is extensively practised. The use of multiple classes on offer, for the same reason as for public transport, gives the possibility to separate by income groups and trip purpose. Another possibility arises because airline tickets are usually bought in advance. Tourists tend to plan their trips earlier and to be more price sensitive than business travellers, which is one reason why it makes sense to offer lower tariffs for early bookings. At the same time, few business travellers would like to run the risk of missing an important meeting because a flight is fully booked, which explains why it may also be sensible to sell the last seats cheaply on a last-minute basis – a sub-market that is certainly for some destinations populated primarily by highly price-sensitive back-packers. Furthermore, most tourists would appreciate spending a weekend at the destination of the trip, whereas experienced professional travellers often do not want to lose a weekend at home every time they make a trip. Most airlines therefore find it profitable to charge different prices, depending on whether a Saturday night is spent at the destination.

3.4 TRANSPORT PRICING BY MULTIPLE GOVERNMENTS

Another quite different reason why the overall objective pursued by the agency setting the toll may not be social welfare occurs when there are different governments, each representing its own constituency. Even if in such a case each government acts so as to maximize social surplus for the inhabitants of its own territory, the combined result of all actions undertaken by all governments is likely not to be maximizing aggregate surplus; that is, social surplus enjoyed over all territories jointly. There can be a variety of mechanisms at work that may cause such failure to achieve the best possible outcome. Different mechanisms may in turn lead to deviations in different directions from what would be chosen by a "global benevolent dictator," i.e. a global government that maximizes social surplus over all territories combined. For example, taxes may be set too high or too low from the overall social perspective – but, obviously, in either case, the aggregate social surplus will be below what is achieved under globally optimal policies. When such instances occur, there will potentially be a strong case for the coordination of policies between governments, since at least in theory it must be possible to distribute the aggregate welfare gains from optimal coordination in such a way that everybody is better off. Still, governments may continue to deviate from what would be globally optimal, and more strongly so the more the voting behaviour in their territory depends on the extent to which the government succeeds, through its policies, in raising local social surplus enjoyed by these voters. Indeed, it may well be partly the pressure from democratic voting in their own jurisdiction that pushes a local government away from policies that would be more in line with global objectives, towards satisfying more local interests.

Quite intuitively, the stronger the links between territories served by different governments, the stronger the chances that the combined result of the policies implemented by the different governments will deviate from what is globally optimal. That is, imagine an extreme situation where territories are completely isolated from each other, in each possible respect. The optimization problems faced by the local governments will then be entirely independent, and independent optimization by each government would produce not only local optima, but also the overall global optimum. But as soon as the optimization problems become mutually dependent, risks of inefficiencies through the absence of policy coordination emerge.

Because transport by definition means that agents involved are mobile and may thus enter territories other than their home territory, the chances of inefficiencies of this type occurring in transportation markets are far from imaginary. Still, it is by no means a challenge facing transport markets alone. For example, one of the seminal contributions to the literature on "tax competition" addressed the case where governments compete to attract (internationally) mobile capital to their region, which can subsequently be taxed to finance the supply of a local public good. Setting a lower capital tax rate may then attract so much more capital to the region that the total tax revenues in fact increase (Oates, 1972). This gives an incentives to set a relatively low tax rate, which is reinforced when competing governments do the same, so that a "race to the bottom" may occur, where relatively low taxes are applied, and public goods are in all regions supplied in lower qualities and quantities than what would be socially optimal.

Anyone who has driven south from the north of Europe to more sunny holiday destinations is likely to recognize that a similar mechanism appears to be at work in Luxemburg, but then in a transport setting. Fuels in Luxemburg are relatively cheap compared to fuel prices in the neighbouring countries, the reason being that a lower tax rate is applied. Behind this lower tax rate is Luxemburg's recognition that this will attract so much more traffic that it boosts total tax revenues, even though the tax revenues per litre are somewhat lower. Phrased differently, with a fuel tax equal to that of the neighbouring countries, demand will be so elastic that it is simply too hard to resist the temptation to lower the fuel tax, and in that way raise revenues.

At the same time, the neighbouring countries – France, Germany, Belgium – face other objectives with their (nationally determined) fuel taxes, above competing for mobile taxpayers with Luxemburg in one corner of their territories, and therefore this situation appears a logical and stable outcome of the process of tax competition.

Indeed, transport is just one example of an economic activity in which tax competition and other types of coordination problems and strategic interactions between different governments can arise, and the literature on tax competition encompasses many different sectors (Wilson, 1999, provides a broad review). At the same time, there are more types of strategic interactions between governments than only the setting of taxes, making the competition between governments in reality often a complex and multidimensional process. Moreover, governments can be organized in different ways and may have different types of hierarchical relations between them. The outcomes of the strategic interactions between governments will generally strongly depend on the type of relation between the governments involved, and on the type of interactions that is at stake. It is therefore useful to present a structured overview of the types of interactions that may exist between jurisdictions, and that is what the next sub-section will do.

3.4.1 Different Types of Policy Interactions: A Taxonomy

Table 3.1, adapted from De Borger and Proost (2012), presents the most relevant types of strategic interactions that can occur between governments in the context of transport policymaking, also identifying the likely implications on policy choices.

A first important distinction is between horizontally versus vertically ordered governments, where the former refers to governments with non-overlapping areas and populations (e.g. neighbouring countries), and the latter to instances where there is such an overlap (e.g. a national government versus the local government of one of its cities). In general, because with vertical interactions there is at least some overlap in voting populations served, one may expect deviations from globally optimal policies to be larger under horizontal interactions. But this certainly does not always have to be the case.

A second distinction is between types of externalities between jurisdictions. A *fiscal externality* occurs when tax setting by one government affects tax revenues for the other government. An *expenditure externality* occurs when expenditures such as infrastructure investments by one government affects the well-being of the population served by the other government. And finally, an *environmental externality* refers to a situation where environmental effects spill over into other jurisdictions.

In this way, six different categories are distinguished in Table 3.1. The first of these, the case of a *horizontal fiscal externality*, can in turn be subdivided into two distinctly different types of interactions. The fact that the likely implications on tax levels are opposite is illustrative of the variety of impacts that strategic interactions between jurisdictions may have, and therefore the care that should be taken when analyzing the economics of such interactions. The first mechanism concerns *tax exporting*, and this refers to the desire to make foreigners pay taxes. Especially when foreigners are captive and have no alternative, this will lead to an upward bias on taxes. A good example of tax exporting in the context of road pricing is the toll on German roads that was scheduled to be introduced in 2016. Where both foreign and German drivers are required to pay a toll by acquiring a ten-day, two-month or annual pass, German drivers would be fully compensated through a simultaneous reduction in annual vehicle taxes. Effectively, therefore, the revenues from the toll system would be entirely due to foreign

Table 3.1 Possible types of interactions in transport markets between jurisdictions

Type	Source	Transport example	Likely implications
Horizontal fiscal externality	**Tax exporting**: make outsiders pay taxes **Tax competition**: attract a mobile tax base	Higher tolls for foreign road users Lower fuel tax to attract foreign drivers (Luxemburg)	Upward pressure on taxes Downward pressure on taxes
Horizontal expenditure externality	**Benefit spill-over** **Expenditure competition**	Underinvestment in infrastructure used by foreigners Overinvestment to attract foreign firms	Downward pressure on investments Upward pressure on investments
Horizontal environmental externality	**Pollution spill-overs**	Trans-boundary pollution, such as global warming from CO_2 emissions	Downward pressure on environmental regulations and taxes
Vertical fiscal externality	**Overlapping tax bases**	National fuel taxes and local parking fees used to raise revenues	Upward pressure on combined tax levels
Vertical expenditure externality	**Expenditure interdependence**	Spending on local roads raises use and therewith national fuel tax revenues	Downward pressure on such expenditures
Vertical environmental externality	**Pollution spill-overs**	Trans-boundary pollution between lower-level jurisdictions positioned under the same higher-level government	Dampening the bias from a purely horizontal environmental externality

Source: adapted from De Borger and Proost (2012).

drivers, and this discrimination is why the plans were foiled by the European Court of Justice (ECJ) in 2019. The second type of horizontal fiscal externality, *tax competition*, becomes relevant when the tax base (the individuals paying the tax) is mobile and can therefore escape paying the tax by adjusting behaviour. The earlier example of fuel taxes in Luxemburg fits this case, and shows that in contrast to what we saw for tax exporting, there will be a downward bias on the tax level.

Also, in the case of a *horizontal expenditure externality*, there may be opposing effects. With a *benefit spill-over*, foreigners also use the facility under consideration, and this may lead to a downward bias in the incentives to invest in its quality and maintenance, especially when foreigners have unpriced access. The mediocre or poor quality of untolled roads that are used relatively intensely by foreigners is a case in point. But under *expenditure competition*, in contrast, jurisdictions try to attract firms or visitors by providing excellent infrastructure, and in this case an upward bias in the incentives to invest in quality and maintenance can be expected. A good example is the oversupply of commercial areas ("bedrijventerreinen") and associated infrastructures, offered by Dutch municipalities in the hope of attracting firms and therewith employment.

Next, with a *horizontal environmental externality*, emissions caused within a certain jurisdiction also affect environmental quality outside that area. The emission of CO_2, affecting global warming, provides an extreme example. To the extent that the local government seeks to promote local social welfare, the environmental effects occurring outside the jurisdiction tend to be overlooked in the formulation of environmental policies. Indeed, for a perfectly global environmental externality such as global warming, the appropriate description of the incentives faced by local governments would be that of a public good (or bad), where the economic incentive would be to free-ride on others' contributions. More in general, the larger the share of effects of emissions occurring outside the jurisdiction, the larger the incentive for the local government to make less than a globally optimal effort to restrict the externality.

The basic mechanisms for the three types of *vertical externalities* distinguished in Table 3.1 are similar, but, as stated, because populations are partly overlapping with the interests of the two governments are typically somewhat better aligned than for the horizontal externalities just discussed. But there remain inefficiencies. For a *vertical fiscal externality*, for example, the joint effect of both governments seeking to distract funds from the same individuals through taxation may mean that the aggregate tax becomes so high that a reduction would increase aggregate revenues. For a *vertical expenditure externality*, a local government may ignore that part of the benefits of induced traffic from an infrastructure investment accrues to the central government in the form of, for example, fuel tax revenues. Counting these as a loss at the local level, rather than as a transfer, would underestimate the aggregate benefits of the project and hence may lead to underinvestment. And finally, with a *vertical environmental externality* the underlying mechanisms are comparable to those for a horizontal environmental externality, with the difference that the higher-level government may now succeed in dampening the bias from a purely horizontal environmental externality.

3.4.2 Diverging Incentives with Horizontal Externalities: An Example

We may finally further illustrate the issues at stake by considering an example with two types of horizontal externalities: fiscal and environmental. Consider a world that consists of two regions only, A and B. There is a road in region A that is used by drivers from both regions, the quantities being N_A and N_B. For both groups, there is an inverse demand function denoted $D_A(N_A)$ and $D_B(N_B)$, and assuming that drivers are otherwise identical we may write the average cost as $ac(N)$ with N=N_A+N_B; i.e. there is congestion and both groups contribute to it in equal amounts. And finally, there is an environmental effect: every trip causes damage e^A in region A, and e^B in region B. Note that the superscripts A and B therefore do not refer to the region of the origin of the driver: that is assumed to be immaterial for the environmental externality caused. Finally, we assume that, if desired, tolls can be differentiated between drivers from regions A and B.

To see the possible impacts of the horizontal externalities upon public policies, let us first write out the local social surplus S_A, and compare it to the global social surplus, S. To start with the latter, this consists of the sum of benefits, over all users, minus user costs for all users, minus all external environmental costs:

$$S = \int_0^{N_A} D_A(n)\mathrm{d}n + \int_0^{N_B} D_B(n)\mathrm{d}n - N \cdot ac(N) - N \cdot (e^A + e^B) \tag{3.14}$$

The local government, in contrast, maximizes local social surplus S_A. This consists of total benefits for local inhabitants, minus user costs incurred by these same drivers, minus locally incurred environmental costs, plus tax revenues extracted from foreign drivers:

$$S_A = \int_0^{N_A} D_A(n)\mathrm{d}n - N_A \cdot ac(N) - N \cdot e^A + N_B \cdot \left(D_B(N_B) - ac(N)\right) \quad (3.15)$$

Maximization of global surplus requires us to take the derivative of Equation(3.14) with respect to both levels of use, and this leads to:

$$\frac{\partial S}{\partial N_A} = D_A(N_A) - ac(N) - N \cdot ac'(N) - (e^A + e^B) = 0 \rightarrow r_A = N \cdot ac'(N) + (e^A + e^B) \quad (3.16a)$$

$$\frac{\partial S}{\partial N_B} = D_B(N_B) - ac(N) - N \cdot ac'(N) - (e^A + e^B) = 0 \rightarrow r_B = N \cdot ac'(N) + (e^A + e^B) \quad (3.16b)$$

Note that to obtain the toll expressions in the second line, we substitute $r = D{-}ac$. Both tolls are set equal to marginal external costs, entirely following the logic outlined in Section 3.2. And because the marginal external costs are independent of the origin of a driver, the tolls are equal, too.

The incentives for the local government maximizing S_A can be identified by maximizing Equation (3.15):

$$\frac{\partial S_A}{\partial N_A} = D_A(N_A) - ac(N) - N_A \cdot ac'(N) - e^A - N_B \cdot ac'(N) = 0 \rightarrow r_A = N \cdot ac'(N) + e^A \quad (3.17a)$$

$$\frac{\partial S_A}{\partial N_B} = -N_A \cdot ac'(N) - e^A + D_B(N_B) - ac(N) + N_B \cdot \left(D_B{}'(N_B) - ac'(N)\right) = 0 \rightarrow r_B = N \cdot ac'(N) + e^A - N_B \cdot D_B{}'(N_B) \quad (3.17b)$$

The tolls that the local government sets reflect that it behaves as the social welfare optimizer from Section 3.2 insofar as local drivers are concerned, and as the profit-maximizer from Section 3.3 towards foreign drivers. Interestingly, this implies that both types of drivers face the full marginal external congestion costs $N{\cdot}ac'(N)$. This is a direct consequence of what we found in Section 3.3: also a profit-maximizer perfectly internalizes congestion externalities, albeit for different reasons than those applying for a welfare-maximizer. For foreign drivers, also the profit-maximizing demand-related markup $-N_B \cdot D'_B$ that we found in Section 3.3 is again added. And finally, for both groups only the locally incurred environmental cost e_A is included, missing out on the damage abroad e_B. We see that, as a consequence, the toll for local drivers is below what would be globally optimal, the difference between Equations (3.16a) and (3.17a) being equal to e_B. For foreign drivers, the toll may be either below or above the globally optimal one, depending on whether e_B is smaller or bigger than $-N_B \cdot D'_B$. This illustrates how the effects of horizontal externalities on tax levels can be varied, and may even change significantly depending on the circumstances.

The above example is instructive in that it shows how tax competition may indeed seriously affect outcomes in transport markets, and also how the interpretation of results may benefit by comparing these to what is known from public pricing and private pricing. Still, this is obviously just one, very stylized example. The literature on tax competition in transport is rich and growing, and has considered many cases and situations, each with their own peculiarities. De Borger and Proost (2012) provide a nice overview. As their review makes clear, the outcomes under tax competition depend on many aspects, including the network configuration (notably, serial versus parallel links); the game-theoretic setup (including the distinction between Nash versus Stackelberg behaviour of governments); and the set of policy instruments considered (for example, pricing and/or capacity choice). So the example just given should be treated in that light: it is just an example.

3.5 PRICING, REWARDING AND BUDGET-NEUTRAL INCENTIVES

The notion of marginal cost pricing as the instrument that achieves the highest efficiency and maximum social welfare is well-established in the literature, but in practice things typically do not resemble the stylized assumptions that need to be made to make the result indeed apply. In the foregoing, we have already briefly touched upon various types of second-best pricing, for which it is generally true that given the constraints that apply, a higher social welfare can be achieved when deviations are made from the classic Pigouvian tax rule equal to marginal external cost. The two distinct types of reason for why second-best pricing is so relevant in practice are: (1) inherent constraints on the pricing instrument itself, such as the inability to differentiate taxes optimally over time, place, user types, etc.; or (2) non-optimal prices in markets that are interacting with the transport market under consideration, a good example being the case considered by Parry and Bento (2001) where commuting involves workers who are subject to distortive labour taxation.

Another type of constraint, and one that has played an important role in preventing wide-scale implementation of road pricing in practice, concerns limited social and political acceptability. An important factor behind the limited acceptability is believed to be the general tendency of people to dislike higher or additional taxes, despite the consequence that the revenues could be used to finance useful public expenses or reduce public debts. This has given rise to a search for more acceptable price incentives, two important ones being rewards and tradable permits.

Rewards have been tested empirically in a number of experiments in The Netherlands, often under the label Spitsmijden (Peak Avoidance). The general insight drawn from such experiments is that participants seem very sensitive to financial incentives that are designed to make them reduce their peak travel behaviour, where reductions in the order of 50% are no exception (e.g. Knockaert et al., 2012). An important caveat is that such projects tend to attract relatively flexible travellers, who expect to be able to earn relatively high rewards, so the results should not be mistaken to be representative. This seems consistent with the findings of Graham et al. (2020), who found relatively small effects from a 25% early peak shoulder reduction of transit fares in Hong Kong. Another caveat is that subsidies to combat external costs are not the first-best solution and in the long run may attract too many travellers to the road and to the scheme. And a third caveat is that a wide-scale and structural application of rewards to steer transport behaviour would require huge public budgets, which would need to be financed from taxes that in itself would be distortionary too.

These considerations have spurred research into budget-neutral price incentives, avoiding a net financial flow from the government to travellers, as with rewarding, or in the reverse direction as with pricing. An important example is tradable permits. Originally proposed as an instrument to deal with environmental externalities (e.g. Dales, 1968) and also used as such in the well-known European Emissions Trading Scheme, it has also been considered as a possible instrument in the containment of road traffic externalities (Verhoef et al., 1997). Modelling exercises have focused on aspects such as the implementation of the instrument in a transport network environment (Yang and Wang, 2011), or under conditions of uncertainty (De Palma et al., 2018). The actual implementation would assume that travellers understand the instrument as intended, and that the market for permits can be designed such that it operates in an efficient way – issues that have been tentatively explored in experimental studies (e.g. Brands et al., 2020). It seems too early to make any definitive statements on the prospects for tradable permits in the regulation of transport externalities, but the theoretical properties of combining the efficiency of incentives with budget-neutrality, and the first experiences in experimental settings seem sufficiently promising to justify the continuation of research on this topic.

3.6 CONCLUSION

This chapter has reviewed the theory of transport pricing under different objectives than the textbook reference of first-best optimization of social welfare. We have discussed profit maximization pricing, pricing by multiple governments and – albeit briefly – price incentives that are constrained to be positive or budget-neutral. We have emphasized that although different objectives naturally lead to non-optimal prices, there remain strong conceptual links with insights from the theory of first-best pricing. For example, a profit-maximizing operator has an incentive to internalize congestion externalities, while our example of tax competition produced pricing rules that balance elements from surplus-maximizing and profit-maximizing pricing.

Despite the beneficial impacts that optimized prices might bring, application to apply these remains scarce, in particular in road transport. Especially then, economic theory can be helpful in identifying ways of providing financial incentives that, under constraints of public and political acceptability, produce maximum efficiency. Budget-neutral instruments such as tradable permits may have an important role to play there.

Important challenges for near-future policymaking, but also transport economics research, including the design, evaluation and implementation of innovative price incentives that employ novel technologies in terms of price differentiation and (first-, second- and third-degree) discrimination to better meet objectives and/or constraints in terms of behavioural change, equity and acceptability. Some current developments make this even more pressing than before. An obvious first one is the increasing awareness of, and commitment to counter, global warming. A second one is the predicted rapid growth in Mobility-as-a-Service facilities and sharing-economy-based services, be it from private operators with their own objectives or from public operators, and be it in monopolistic or in strategic oligopolistic supply. This leads to a much more complex and volatile environment in which prices are formed, and in which socially optimal pricing cannot be expected to arise spontaneously. And third is the continuing concern with urban congestion and its likely return to pre-COVID-19 levels after the pandemic has calmed. Carefully designed pricing schemes to address these matters, fully exploiting the possibilities that contemporary pricing technologies offer, are not the only or full answer to combat the grand challenges in urban transportation – but ignoring the importance and potentials of pricing would undoubtedly make that combat much more so an uphill battle.

NOTES

1. The expositions in some sections of this chapter closely follow presentations in an earlier text that is made freely available as a reader for Bachelor students at VU Amsterdam (Verhoef, 2002). Section 3.2 uses the same notation as used in that reader, but presents the material differently. Sections 3.1 and 3.4 are taken directly from that source, with at most only minor textual adjustments. The same texts served as the basis for a textbook that recently appeared in Chinese as Verhoef, Wang and Hu (2020).
2. The product rule of differentiation tells us that the derivative of a composite function $f(x) = h(x) \cdot g(x)$ with respect to x is equal to $h(x) \cdot g'(x) + g(x) \cdot h'(x)$. The reader may test this rule for the simple function $f(x) = x^2 = x \cdot x$.
3. The standard monopolistic price rule, with the operator's mc = 0, would be: $r \cdot (1 + 1/\varepsilon) = 0$ (where ε is demand elasticity). The demand elasticity ε is generally defined as $\mathrm{d}N/\mathrm{d}r_p \cdot r_p/N$. If ac were constant, the first term $\mathrm{d}N/\mathrm{d}r_p$ would be equal to $1/D'$ (because D' would then give $\mathrm{d}r_p/\mathrm{d}N$). With a non-constant ac, however, $\mathrm{d}N/\mathrm{d}r_p$ becomes equal to $1/(D'-ac')$, because the slope of the ac function makes N less responsive to marginal changes in r_p than what would appear from the demand function alone. Put differently: the actual inverse demand function faced by the monopolist, with r_p rather than p on the vertical axis, can be constructed as the vertical distance between D and ac and thus reads $D-ac$, so that its slope is given by $D'-ac'$. We thus find: $\varepsilon = 1/(D'-ac') \cdot r_p/N$ and, therefore, $1/\varepsilon = (D'-ac') \cdot N/r_p$. We can then finally rewrite the standard rule as $r_p + N \cdot (D'-ac') = 0$. This exactly replicates (12).
4. The fact that the network is otherwise unpriced makes this a second-best situation: the second-best optimal capacity and toll accounts in the best possible way for the distortions occurring elsewhere in the network. Should the network be otherwise optimally dimensioned and priced, then the auction replicates the first-best capacity and toll. That is, in all circumstances does it produce the most efficient solution available.

REFERENCES

Arnott, R., A. De Palma and R. Lindsey (1993) "A structural model of peak- period congestion: A traffic bottleneck with elastic demand." *American Economic Review* 83 (1), 161–179.

Brands, D. K., E.T. Verhoef, J. Knockaert and P.R. Koster (2020) "Tradable permits to manage urban mobility: Market design and experimental implementation." *Transportation Research Part A: Policy and Practice* 137, 34–46.

Brueckner, J.K. (2002) "Airport congestion when carriers have market power." *American Economic Review* 92, 1357–1375.
Dales, J. (1968) "Land, water, and ownership." *The Canadian Journal of Economics/Revue Canadienne d'Economique* 1 (4), 791–804.
De Borger, B. and S. Proost (2012) "Transport policy competition between governments: A selective survey of the literature." *Economics of Transportation* 1, 35–48.
de Palma, A., S. Proost, R. Seshadri and M. Ben-Akiva, M. (2018) "Congestion tolling-dollars versus tokens: A comparative analysis." *Transportation Research Part B: Methodological* 108, 261–280.
Economides, N. and S.C. Salop (1992) "Competition and integration among complements, and network market structure." *Journal of Industrial Economics* 40, 105–123.
Edelson, N.E. (1971) "Congestion tolls under monopoly." *American Economic Review* 61 (5), 872–882.
Engel, E., R. Fisher and A. Galetovic (1997) "Highway franchising: Pitfalls and opportunities." *American Economic Review, Papers and Proceedings* 87 (2), 68–72.
Graham, D. J., D. Hörcher, R.J. Anderson and P. Bansal (2020) "Quantifying the ex-post causal impact of differential pricing on commuter trip scheduling in Hong Kong." *Transportation Research Part A: Policy and Practice* 141, 16–34.
Hörcher, D., D.J. Graham and R.J. Anderson (2018) "The economic inefficiency of travel passes under crowding externalities and endogenous capacity." *Journal of Transport Economics and Policy* 52 (1), 1–22.
Knockaert, J. S. A., Y. Tseng, E.T. Verhoef and J. Rouwendal (2012) "The Spitsmijden experiment: A reward to battle congestion." *Transport Policy* 24, 260–272.
Mills, D.E. (1981) "Ownership arrangements and congestion-prone facilities." *American Economic Review, Papers and Proceedings* 71 (3), 493–502.
Mohring, H. and M. Harwitz (1962) *Highway Benefits: An Analytical Framework*. Evanston, IL: Northwestern University Press. Chapter II: Benefits and the tax system (pp. 70–90).
Oates, W. (1972) *Fiscal Federalism*. New York: Harcourt Brace Jovanovich.
Parry, I.W.H. and A. Bento (2001) "Revenue recycling and the welfare effects of road pricing." *The Scandinavian Journal of Economics* 103 (4), 645–671.
Pigou, A.C. (1920) *Wealth and Welfare*. London: Macmillan.
Small, K.A. and E.T. Verhoef (2007) *The Economics of Urban Transportation*. London: Routledge.
Varian, H.R. (1985) "Price discrimination and social welfare." *American Economic Review* 75, 870–875.
Verhoef, E.T. (2002) *Markets and Governments: Transport Economic Applications*. Reader, Vrije Universiteit Amsterdam.
Verhoef, E.T. (2007) "Second-best road pricing through highway franchising." *Journal of Urban Economics* 62, 337–361.
Verhoef, E.T., P. Nijkamp and P. Rietveld (1996) "Second-best congestion pricing: The case of an untolled alternative." *Journal of Urban Economics* 40 (3), 279–302.
Verhoef, E., P. Nijkamp and P. Rietveld (1997) "Tradeable permits: Their potential in the regulation of road transport externalities." *Environment and Planning B: Planning and Design* 24 (4), 527–548.
Verhoef, E.T., Y. Wang and Y. Hu (2020) *Markets and Governments: Transport Economic Theory and Applications (in Chinese)*. China: Social Sciences Academic Press.
Vickrey, W.S. (1969) "Congestion theory and transport investment." *American Economic Review* 59, 251–260.
Wang, J. Y., R. Lindsey and H. Yang (2011) "Nonlinear pricing on private roads with congestion and toll collection costs." *Transportation Research Part B: Methodological* 45 (1), 9–40.
Wilson, J. (1999) "Theories of tax competition." *National Tax Journal* 52, 269–304.
Yang, H. and X. Wang (2011) "Managing network mobility with tradable credits." *Transportation Research Part B: Methodological* 45 (3), 580–594.