Congesting Pricing in a Road and Rail Network with Heterogeneous Values of Time and Schedule Delay

Vincent van den Berg
Erik T. Verhoef

VU University Amsterdam, and Tinbergen Institute.

Congesting pricing in a road and rail network with heterogeneous values of time and schedule delay

Vincent van den Berg[a,*]**, Erik T. Verhoef**[a,b,#]

a: Department of Spatial Economics, VU University, De Boelelaan 1105, 1081HV Amsterdam, The Netherlands
b: Affiliated with the Tinbergen Institute, Roetersstraat 31, 1018 WB Amsterdam, The Netherlands.
\#: Email: e.t.verhoef@vu.nl
*: Corresponding author: email: v.a.c.vanden.berg@vu.nl,   tel: +31 20 598 6049,   fax: +31 20 598 6004

Version of 24 March 2011

Abstract

We analyse congestion pricing in a road and rail network with heterogeneous users. On the road there is bottleneck congestion. In the train there is crowding congestion. We separately analyse "proportional heterogeneity" that varies the values of time and schedule delay scalarly in fixed proportions and "ratio heterogeneity" that is in the ratio of value of time to value of schedule delay. We analyse first-best pricing and second-best pricing on only road or rail. More "ratio heterogeneity" lowers the relative efficiency of welfare maximisation by pricing only the road. This relative efficiency also decreases with proportional heterogeneity. Conversely, in previous research with two parallel roads, the relative efficiency of "single-link pricing" increased with proportional heterogeneity. This difference is caused by two road being perfect substitutes, while car and train are imperfect substitutes. The welfare gain of congestion pricing in the train is less affected by heterogeneity. However, the relative efficiency of "profit maximisation on rail only" decreases with proportional and ratio heterogeneity. There are marked differences in the distributional effects on road and rail. On the road, pricing is generally *more* beneficial for the user the higher her value of time or schedule delay is. In the train, pricing has no distributional effects or is *less* beneficial the higher a user's value of time or schedule delay is.

1. Introduction

Road congestion is one of the most important problems in urban areas. Indeed, as Kraus (2003) states, most urban economists agree that the pricing below marginal cost of car travel causes the largest waste of all urban policies. Small and Verhoef (2007) find that congestion caused the largest marginal external cost due to car travel in the US in 2005: of the total 65% was from congestion, 6% was environmental, and 25% from accident externalities. To alleviate congestion, one could toll the congestion externalities, expand capacity, or attract drivers to public transport.

In studying transport pricing is it important to take into account the dynamics of departure time choice, and substitute modes, as well as preference heterogeneity in multiple dimensions. Moreover, the distributional effects of a policy are also important as these influence the political acceptability. We study the two-mode problem of congestion pricing in an intercity rail and road network with continuous heterogeneity in values of time and schedule delay, where the two modes are imperfect substitutes.[1] Arnott and Yan (2001) find that amount of substitutability

---

[1] Although we call the two modes rail and car, they can be any two modes that have no congestion interactions. The two modes could be any two of: freeway, (arterial) road, air, (light) rail, and metro; but not bus and car, as these use the same road.

affects the second-best congestion charge when only one mode can be charged, and also affects the gain of this policy.

It is important to control for heterogeneity. Arnott, de Palma and Lindsey (1988) find that just using the means of utility parameters biases the calculated gain of tolling, where the bias can be negative or positive. Small and Yan (2001) and Verhoef and Small (2004) analyse a two roads network with static flow congestion. The gain of first-best pricing and relative efficiency of second-best pricing of one of two parallel links increase with the heterogeneity of the value of time. Relative efficiency is the welfare gain of a policy relative to the first-best gain.

Distributional effects are an important reason why congestion pricing meets resistance. When pricing reduces travel times and increases monetary costs, an individual is better off when her value of time is higher. Since the value of time generally increases with income, this is often taken as implying that the poor lose and the rich gain (e.g. Layard, 1977). However, things might not be as simple as this. In Vickrey's (1973) bottleneck model, demand is fixed. He analyses what we call proportional heterogeneity. In his setting the values of time (α) and schedule delay (β) vary in fixed proportions following the heterogeneous scalar k_i: $\alpha_i = a \cdot k_i$, and $\beta_i = b \cdot k_i$. First-best (FB) tolling is now a strict Pareto improvement: all users gain—except those with the lowest values, who are unaffected. The values of time and schedule delay are the ratio of some marginal utility to the marginal utility of income. A higher income generally leads to a lower marginal utility of income, thus proportionally increasing values of time and schedule delay. Hence, proportional heterogeneity could be viewed as an effect of income differences.

De Palma and Lindsey (2002) study heterogeneity in the value of time with fixed values of schedule delay in the bottleneck model. Tolling is now harmful for users: under fixed demand, all users lose—except those with the highest value of time, who are unaffected. We call this heterogeneity "ratio heterogeneity" because what matters for behaviour is not the value of time itself, what matters is the ratio of value of time to value of schedule delay $\mu_i = \alpha_i/\beta$. This ratio has an interpretation: it reflects the willingness to accept greater schedule delays in order to reduce travel time. Heterogeneity in μ could for example result from differences in possibilities to use in-vehicle time productively or differences in the tightness of scheduling constraints: e.g. having small children versus not, working on an assembly line versus in an office, and travelling to a hospital appointment versus to go shopping.

Additionally, Van den Berg and Verhoef (2011b) show that price sensitivity of demand changes the effects of heterogeneity in the bottleneck model. With proportional heterogeneity, FB tolling lowers the generalised prices for users with high values and as a result they demand more travel, which increases congestion and makes tolling harmful for drivers with low values. This paper finds that also the amount of substitutability of the two modes again changes the effects of pricing, and how heterogeneity influences these effects. Hence, it is important to control for heterogeneity and price-sensitive demand, but also for the substitutability of different modes.

Braid (1996) studies "single-link pricing" with two parallel bottlenecks. The welfare-maximising toll has a time-variant term that equals the congestion externality and a time-invariant term that it is negative. This negative term attracts users away from the unpriced link where social marginal cost is above private cost, and thereby increases welfare. A consequence of all this is that the toll is negative in the shoulder of the peak.

Public transport needs a different specification of user costs than road transport. The basic economic model of public transport is by Mohring (1972). It has no congestion. Bus size is fixed. Finally, users arrive randomly at the stops. Since busses run perfectly on time, the average waiting time is half the headway between bus services. If the operator minimises the sum of his own costs and user costs, the optimal frequency increases with the square root of the number of

users. Therefore, an increase in ridership lowers waiting times and per-user operating costs, even if there are neutral scale economies in the production of bus services. This is opposite to on the road where average cost increases with the number of users due to congestion.

Kraus and Yoshida (2002) develop a public transport model with queuing at the platform to enter a service, possibly for up to several headways between services. Their model is similar to the bottleneck model: e.g. also here congestion pricing eliminates all queuing. Tabuchi (1993), Huang (2000), and Rouwendal and Verhoef (2004) use a parallel road and rail track. Tabuchi (1993) has bottleneck congestion on the road and no congestion in the single train service that arrives at the preferred arrival time t^*. The train operator's costs consist of fixed and per-user cost. If the train fare equals average operating cost, the second-best road toll is the standard time-variant toll plus a *positive* term that pushes users to the train. This differs from Braid (1996), who adds a negative term. The average-cost fare is above marginal costs, and an increase in the number of users lowers average cost, which is why a positive addition to the road toll that pushes drivers to the train is optimal. This shows that similar policies in road transport and public transport can look very different, which makes it important to take into account that there are different modes. Huang (2000) considers two user groups. The road has dynamic bottleneck congestion. The single train service arrives at t^* and has crowding congestion: that is costs associated with being in a crowded train. Rouwendal and Verhoef (2004) also use crowding congestion. However, they use a static road model, and car and train are imperfect substitutes.

We study a parallel road and rail line using a dynamic model. On the road there is bottleneck congestion. In each train service there is crowding congestion. Our train model is dynamic; whereas Tabuchi (1993) and Huang (2000) use a static train model. Moreover, we also consider profit maximisation on only road or rail, which appears to have been ignored for a rail-road network. In reality, rail companies are often private or at least partly privatised; thus, profit maximisation by a rail operator seems important. We only look at the short-term policy of price setting. We ignore long-term policies such as train schedule or road, rail-track and train capacity. Rail-operating costs are a constant amount per user, and there are no fixed costs. We also ignore heterogeneity in the value of crowding congestion.

We find that heterogeneity affects road and rail differently. Without congestion pricing, the mean price of car travel decreases with ratio heterogeneity; while proportional heterogeneity has no effect. The average price of train travel is unaffected by ratio heterogeneity, but decreases with proportional heterogeneity. The relative efficiency of "welfare maximisation while only pricing the road" decreases with proportional heterogeneity. This differs from two parallel bottlenecks in Van den Berg and Verhoef (2011b), where the relative efficiency of single-link pricing increases with proportional heterogeneity. As we will explain, this difference is due the assumption that two roads are perfect substitutes, whereas road and rail are imperfect substitutes.

Section 2 discusses the analytical model for car travel; Section 3 derives the model for rail travel. Section 4 describes the numerical set-up. Sections 5 and 6 analyse the numerical model for respectively proportional and ratio heterogeneity. They also give sensitivity analyses on the amount of heterogeneity. Section 7 has sensitivity analyses on the rail-operating cost structure, the cross-price elasticity, and the value of crowding. Section 8 concludes.

2. Analytical road model

One road and one rail track connect the single origin and destination. To simplify the rail model, we assume that users cannot arrive after the common preferred arrival time t^*.[2] The analytical models ignore free-flow travel time and car-operating costs; the numerical models include these. To simplify the analysis, we analyse the two types of heterogeneity separately: so we first study ratio heterogeneity and then proportional heterogeneity. Van den Berg and Verhoef (2011b) find that the qualitative effect of one type of heterogeneity is the same regardless of the amount of the other type. Still, this does make it hard to examine distributional effects, as these depend on the extents of both types of heterogeneity. Note that we use a deterministic model, and hence there is no uncertainty or risk-aversion.

2.1. Heterogeneity in the ratio of value of time to value of schedule delay and road pricing

Ratio heterogeneity is operationalised as heterogeneity in the value of time with a fixed value of schedule delay: the ratio follows $\mu_i \equiv \alpha_i/\beta$. This heterogeneity should not be viewed as stemming from income differences, as income differences should also lead to heterogeneity in β and not only in α_i. We use subscript i to indicate a group of users with the same value of time, so P_i would be the generalised price (henceforth price for brevity) of users with a value of time of α_i. Our model is adapted from Van den Berg and Verhoef (2011a).

Figure 1 gives an example with two-groups heterogeneity and fixed demand. We use this example, because it explains the effects of heterogeneity most clearly. The full model uses continuous heterogeneity. The left panel gives the development of travel time in the No-Congestion-Pricing (NCP) equilibrium, the right panel the first-best (FB) toll. The curves are iso-price curves, and a higher curve implies a higher price. The High group has a higher value of time and thus a higher ratio $\mu_H \equiv \alpha_H/\beta$. In the NCP case, the groups travel temporally separated. The high μ_H of the High group means that they choose to arrive early to limit travel time, since a high μ_i means that a user cares more about travel time than about schedule delays. FB tolling removes all queuing. Now the groups travel jointly, as there are only scheduling costs and tolls which they are assumed to value equally.

The price includes travel-time costs ($C_T[i,t]$), scheduling costs ($C_{SD}[t]$), and a possible toll ($\tau^c[t]$) (we use superscript c to indicate the car, and differentiate the toll from the train charge ρ^t):

$$P_L^{NCP} = C_T[i,t] + C_{SD}[t] = \beta \cdot \frac{n_L^c + n_H^c \cdot \mu_L/\mu_H}{s} = \beta \cdot \frac{n_L^c + n_H^c \cdot \alpha_L/\alpha_H}{s}, \tag{1a}$$

$$P_H^{NCP} = C_T[i,t] + C_{SD}[t] = \beta \cdot \frac{n_L^c + n_H^c}{s}, \tag{1b}$$

$$P_L^{FB} = C_{SD}[t] + \tau^c[t] = \beta \cdot \frac{n_L^c + n_H^c}{s}, \tag{1c}$$

$$P_H^{FB} = C_{SD}[t] + \tau^c[t] = \beta \cdot \frac{n_L^c + n_H^c}{s}. \tag{1d}$$

Here, n_i^c is the number of car drivers of group i; the total number of drivers is N^c; the capacity of the bottleneck is s; and the value of schedule delay is β.

---

[2] With late arrivals, the effect on prices of the number of early and late train services is asymmetric and the exact form of the price equation depends on the ratio of the values of schedule delay early and late. Arnott and Kraus (1993; 1995), Kraus and Yoshida (2002), and Kraus (2003) also use this simplification. Although the restriction of no late arrival is unrealistic; it does not affect the general results of the bottleneck (as also Arnott and Kraus (1993, footnote 12) argue) or the crowding model.

Without heterogeneity the price is $\beta \cdot N^c/s$. Hence, ratio heterogeneity does not affect the price of the High group. Conversely, it lowers the NCP price for the Low group. The high μ_H means that High users build up the queue more slowly than Low users, lowering the price for Low users. In Figure 1 the argument is that the lower slope of the iso-price curve for High users brings the Low users to a lower iso-price curve. This also means that the High users impose lower externalities: the ratio of marginal effects being $(\partial p_L^{NT}/\partial N_H)/(\partial p_L^{NT}/\partial N_L) = \alpha_L/\alpha_H$ (see also Lindsey, 2004). This in turn means that the weighted average of the marginal external costs is lower with ratio heterogeneity than with homogeneity.

Figure 1: Example of ratio heterogeneity with two discrete groups
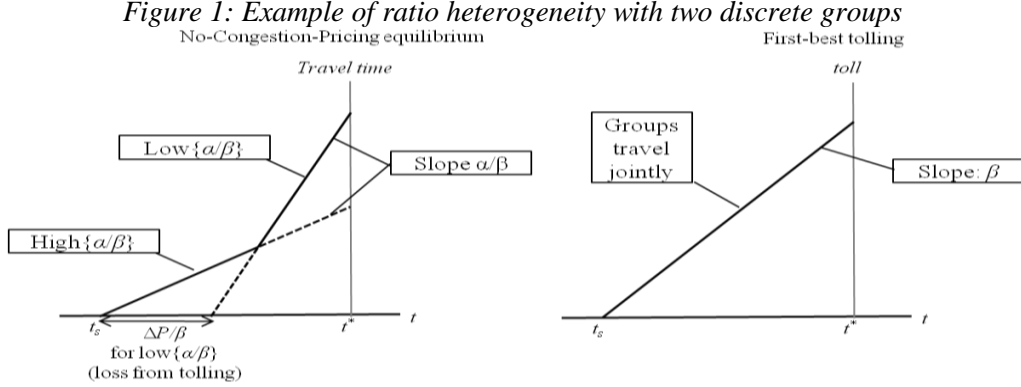
First-best pricing removes all queuing. Now the groups travel jointly, since the slope of the iso-price curve in the right side of Figure 1 is β. The price is now $\beta \cdot N^c/s$ for both groups. Hence, tolling does not affect the High group, but raises the price for the Low group by $\Delta P = \alpha_L/\alpha_H \; \beta \cdot n_H^c/s$. This in turn means that tolling is less welfare increasing with ratio heterogeneity than with homogeneity where prices are unaffected by tolling. With two groups, more heterogeneity means that the ratio α_L/α_H is higher. Thus with more heterogeneity (while keeping the average values fixed), congestion externalities are lower and tolling raises the price for the Low group more. Accordingly, tolling is less welfare enhancing with more heterogeneity.

Arnott, de Palma and Lindsey (1988; 1994) and de Palma and Lindsey (2002) use many groups and fixed demand. Their results are consistent with the foregoing discussion. In the NCP equilibrium, users arrive ordered by the ratio $\mu_i \equiv \alpha_i/\beta$ with the users with the highest ratio arriving furthest from t^*. The price increases with μ_i. The group with the highest ratio has the same NCP price as with homogeneity, while all other groups gain from the heterogeneity. With first-best pricing, all groups travel jointly and face the same price of $\beta \cdot N^c/s$. All users lose due to first-best pricing, except the users with the highest values, who are unaffected.

Van den Berg and Verhoef (2011a) show that price-sensitive demand changes things. Users with low ratios decrease their demand when pricing is introduced, this lowers congestion and the price for all users; making tolling beneficial for users with high ratios. Again, the average NCP externality and price decrease with the heterogeneity of μ. This in turn lowers the gain of tolling. The relative efficiency of tolling one of two parallel bottlenecks also decreases with the amount of ratio heterogeneity. Under continuous heterogeneity, we define an increase in the heterogeneity as an increasing variance of the distribution while the mean remains the same. More heterogeneity could result from more users in the tails of the distribution or the highest and lowest values becoming more extreme.

All this differs from Small and Yan (2001) who use static flow congestion and two parallel links. They find that the gain of first-best pricing and relative efficiency of single-link tolling

increase with the amount of heterogeneity in the value of time. This result is due to self-ordering. Drivers with high values of time use the link with the shorter travel time but higher monetary cost, making the shorter travel time of this link more valuable. Drivers with low values use the link with the longer travel time, making its extra travel time less costly. Accordingly, the type of congestion also determines how heterogeneity affects tolling.

Following Van den Berg and Verhoef (2011a), the price in (2) is the generalisation of (1) for a continuous distribution. In the NCP equilibrium, the distribution of car users follows the probability density function $f^c[\alpha_i]$. The cumulative distribution function is $F^c[\alpha_i]$. The maximum value of time is $\bar{\alpha}$. Scheduling costs equal the left term between brackets, $F^c[\alpha_i]$, multiplied by the term outside the brackets, $C_{SD}[t]=\beta \cdot N^c \cdot F^c[\alpha_i]/s^c$. Although scheduling cost are not directly influenced by α_i; since the arrival t is determined by α_i, the equilibrium C_{SD} does depend on α_i. Queuing costs are the right term multiplied by the term outside the brackets.

$$P_i^{NCP} = C_{SD}[t] + C_T[i,t] = \frac{\beta}{s} N^c \left(F^c[\alpha_i] + \alpha_i \int_{\alpha_i}^{\bar{\alpha}} \frac{f^c[\alpha_j]}{\alpha_j} d\alpha_j \right). \tag{2}$$

Again, as (2) shows, users with a certain μ_i gain when some users with a higher $\mu_j \equiv \alpha_j/\beta$ are replaced by users with an even higher μ_k, but do not suffer if some users with a lower μ_j are replaced by users who have an even lower μ_k. It is this asymmetry that causes total costs to decrease with an increase in ratio heterogeneity.

With FB tolling, users arrive ordered by the value of schedule delay. However, as this value is the same for all, the order is undetermined, and the price in (3) is the same for all users. Here, N^c is the endogenous number of car users in the *new* equilibrium.

$$P_i^{FB} = C_{SD}[t] + \tau^c[t] = \beta \cdot N^c / s \tag{3}$$

Second-best road pricing uses the same formula for the time-invariant toll as the FB toll, but adds a time-invariant toll. With welfare maximisation, the time-invariant part is negative to attract users away from the train where crowding externalities are not priced (see Braid, 1996). With profit maximisation, the term is positive and maximises road toll revenue.

2.2. Proportional heterogeneity and road pricing

Now the relative size of values of time and schedule delay (α_i/β_i) is constant. But all values vary in fixed proportions following the scalar k_i: $\alpha_i=a \cdot k_i$ and $\beta_i=b \cdot k_i$. Our road model is adapted from Van den Berg and Verhoef (2011b). This type of heterogeneity was introduced by Vickrey (1973) and it can be interpreted as stemming from income differences.

Figure 2 gives example NCP and FB equilibria with two groups of users and fixed demand. Now the groups travel jointly in the NCP case, since the slope of the iso-price curve is α/β and this ratio is the same for all. First-best pricing again removes all queuing, and now High users with the high α and β arrive closest to t^*, because they are most willing to pay a toll in order to lower schedule delays. NCP and FB equilibria prices follow

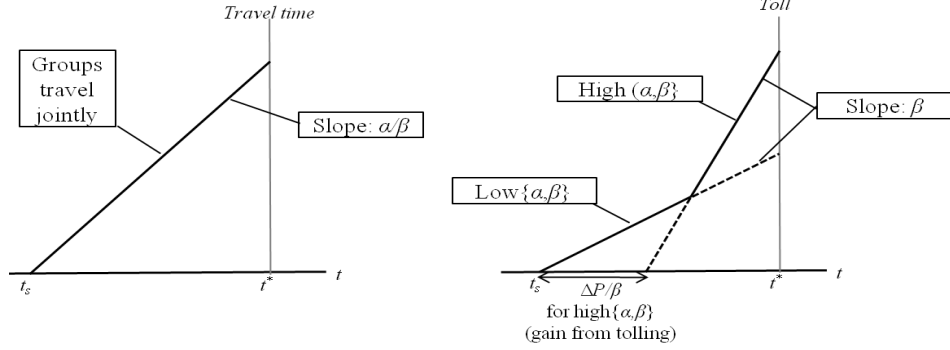$$P_L^{NCP} = C_{SD}[i,t] + C_T[i,t] = \beta_L \cdot \frac{N_L^C + N_H^C}{s}, \tag{4a}$$

$$P_H^{NCP} = C_{SD}[i,t] + C_T[i,t] = \beta_H \cdot \frac{N_L^C + N_H^C}{s}, \tag{4b}$$

$$P_L^{FB} = C_{SD}[i,t] + \tau^c[t] = \beta_L \cdot \frac{N_L^C + N_H^C}{s}, \tag{4c}$$

$$P_H^{FB} = C_{SD}[i,t] + \tau^c[t] = \frac{\beta_L \cdot N_L^C + \beta_H \cdot N_H^C}{s}. \tag{4d}$$

Figure 2: Example of proportional heterogeneity with two discrete groups

The High group gains from tolling, because this group shifts to a lower iso-price curve. For the Low group the price remains the same. The High group gains because users with low values need less of a toll increase over time to prevent queuing—the slope of the toll schedule being *β*—and this lowers the toll that users with high values have to pay.

Vickrey (1973) showed that with continuous heterogeneity the same applies. In the NCP equilibrium, groups travel jointly. With FB pricing, they travel separated and users with the highest values arrive closest to t^*. Under fixed demand, users with the lowest values are unaffected by tolling, all other users gain, and more so the higher their values are. Tolling makes the arrival order more efficient: because drivers with high values of schedule delay now arrive closets to t^*, total scheduling costs decrease. Van den Berg and Verhoef (2011b) show that the gain from tolling increases with a means-preserving increase in the heterogeneity, since the gain from the more efficient arrival ordering increases. Further, the relative efficiency of second-best only tolling one of two parallel bottlenecks also increases. With more heterogeneity, pay-lane users with high values have higher mean values of time and schedule delay, making the travel time and schedule delay savings the pay-lane offers more valuable.

With continuous heterogeneity and no tolling, the NCP price generalises to

$$P_i^{NCG} = C_{SD}[i,t] + C_T[i,t] = \beta_i \cdot N^c / s. \tag{5}$$

The FB price is determined by (6). The distribution of users in a tolling equilibrium follows the probability density function $f^c[\beta_i]$, and the cumulative distribution function is $F^c[\beta_i]$. This distribution is endogenous: it depends on the tolling policy, distribution of users in the NCP equilibrium, and the price sensitivity of each type of user.

$$P_i^{FB} = C_{SD}[i,t] + \tau^c[t] = N^c \left[\beta_i (1 - F^c[\beta_i]) + \int_{\underline{\beta}}^{\beta_i} \beta_j f^c[\beta_j] \, d\beta_i \right] / s \tag{6}$$

Section 3 of Arnott et al. (1988, 1994) has independent heterogeneity in values of time and schedule delay. In our terminology, this means that they have a mix of proportional and ratio heterogeneity. This explains their finding that the welfare gain of tolling can be higher or lower with heterogeneity than with homogeneity: this depends on the relative amounts of proportional

7

heterogeneity—which raises the gain—and ratio heterogeneity, which lowers the gain. Sections 4 and 5 of Arnott et al. (1988, 1994) analyse, respectively, heterogeneity in the ratio of value of schedule delay early and late and in the preferred arrival times. This paper focuses on the two types of heterogeneity in Section 3 of Arnott et al. (1988, 1994).

3. Analytical rail model

In the crowding model of Kraus (1991) the value of time is higher for standing than for seated passengers. Yet, the crowding cost might also depend on the number of users: it is more unpleasant to sit or stand in a fully packed train than in a half-full train. The meta study of empirical estimates by Wardman and Whelan (2011) confirms this. When all seats are taken, the most supported function is a value of time that increases linearly with the number of users. When there are empty seats, the value of time seems constant when most seats are empty; when many seats are taken, some studies find that the value of time increases with the number of users. Wu and Huang (2010) use a value of time that is linear in the number of users; Rouwendal and Verhoef (2004) use a non-linear function. Huang (2000) simplifies the crowding model by using separate crowding and travel time costs. Crowding costs are linear in the number users, and the distinction between standing and seated passengers is ignored. This paper follows Huang's set-up.

Alternative models for train congestion are longer times to board and leave the train or queuing on the platform to enter a train for several headways between services. We prefer crowding congestion. Boarding and leaving times are short relative to total intercity travel time, and thus their costs are minor. Queuing at the train platform for several headways between services (e.g. for 30 minutes) seems rare. Conversely, crowded trains are indeed an often mentioned discomfort. The analytical models ignore travel time and the train fare that covers the marginal operating costs; the numerical models do include these.

3.1. Ratio heterogeneity and rail pricing

Since β is generic there is no product differentiation or self-selection over time, the price for a user is the same in all used train services. The value of time does vary, but train travel time is constant so this does not lead to product differentiation. It is impossible to arrive after t^*. The train that arrives at t^* is service 1; earlier services have a higher index. The headway between train services is given and equals h. So a service is a particular schedule train. The price of service r is the sum of crowding ($C_{CR}[r]$) and scheduling costs ($C_{SD}[r]$), and is independent of the value of time as travel time is zero in the analytical model (in the numerical model it is positive, but since it is constant over time the heterogeneity still does not affect the choice when to travel). Crowding costs in service r are $g \cdot N_r^t$; scheduling costs are $h(r-1)\beta$. The g is the crowding coefficient, N_r^t is the number of users in service r.

Figure 3 shows an example of crowding costs per service that would make the latest 8 services attractive to use. In equilibrium, all services used have the same price. For service 7 to be attractive, it would have to have negative crowding costs. Hence, services 7 and higher are not used, since their price is above the equilibrium price even when it is empty. Service 6 is the earliest service used. It has the highest scheduling costs and the lowest number of users and crowding costs. Service 1 has zero scheduling costs and only crowding costs. Service 2 has $\beta \cdot h$ higher scheduling costs. Hence, its crowding costs must be $\beta \cdot h$ lower. Since crowding costs are $g \cdot N_r^t$, this implies that the difference in the number of users of services 1 and 2 is

$N_1^t - N_2^t = \beta \cdot h / g$. Since these parameters are constant, the difference in number of users of two services j and $j+1$ follows the same formula: $N_j^t - N_{j+1}^t = \beta \cdot h / g$. Accordingly, crowding costs increase linearly over time, which is what the black trend line in Figure 3 indicates.

Figure 3: Crowding costs per service that would make each service equally attractive to use
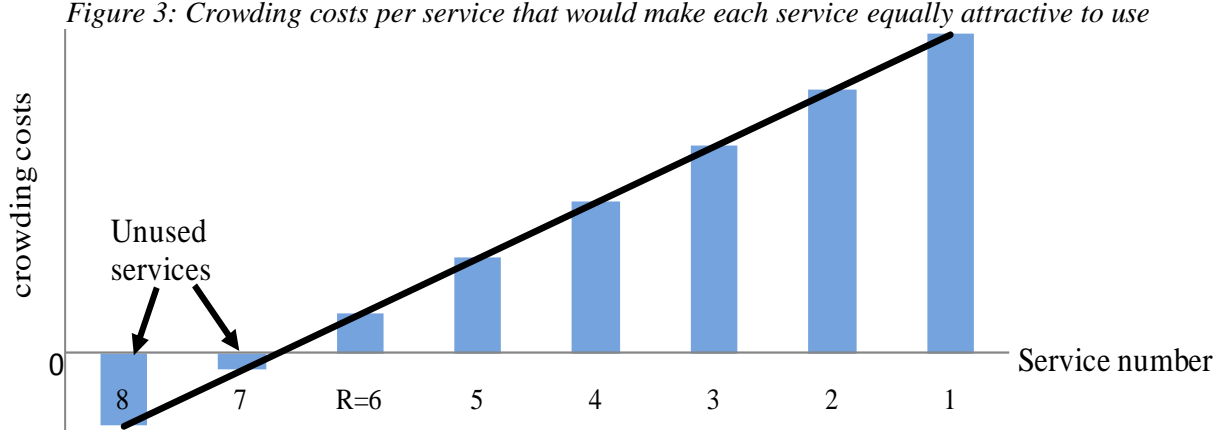
In general the earliest service used will be indicated by R. Note that R is not a policy instrument. There might be earlier services that are unused: for example, in Figure 3, services 7 and 8 are unused. Since by assumption providing services is costless, these empty services do not harm welfare or profits.

Given that $N_i^t - N_{i+1}^t = \beta \cdot h / g$ and that the total number of train users (N^t) equals the sum of users per service, we can shows that the number of total number follows:[3]

$$N^t = \sum_1^R N_r^t = N_1^t + \ N_1^t - \beta \cdot h / g \ + \ N_1^t - 2\beta \cdot h / g \ + ... + \ N_1^t - (R-1)\beta \cdot h / g \ ,$$
$$= N_1^t \cdot R + (1 + 2 + ... + (R-1))\beta \cdot h / g = N_1^t \cdot R + ((R-1)R/2)\beta \cdot h / g$$

which implies that the number of users in service 1 is

$$N_1^t = N^t / R - ((R-1)/2)\beta \cdot h / g. \tag{7}$$

To find R we first calculate for which arrival time the scheduling costs equal the crowding costs of service 1. Since the number of services used is an integer, there is generally no service that arrives at this moment. Service R is then the first service to arrive after this time. Hence, R depends on N^t and the cost parameters following

$$R = \text{Floor}\left[\frac{g \cdot N_1^t}{\beta \cdot h}\right] = \text{Floor}\left[\frac{1 - \sqrt{8g \cdot N^t + \beta \cdot h}}{2\sqrt{\beta \cdot h}}\right], \tag{8}$$

where Floor[x] returns the highest integer not larger than x. We derive the NT price for service 1 in (9) by inserting (7) and (8) into the crowding cost function. This price is also the equilibrium price, since prices are constant over time.

$$P_i^{NCG} = C_{CR}[r] + C_{SD}[r] = g \cdot N_r^t + h(r-1)\beta = g \cdot N^t / R - \beta \cdot h(R-1)/2. \tag{9}$$

---

[3] The substitution of $R(R-1)/2$ for the series $(1+2+...+R-1)$ follows Kraus and Yoshida (2002).

Although the price is the same in all used services, marginal social costs are not. The marginal external cost in service r is $g \cdot N_r^t$. Service 1 has the highest externalities, service R the lowest. The time-variant congestion *levy*, $\rho_r^t[r]$, equals the externality. We use the term *levy* to distinguish the crowding charge from the road toll and train fare (where the fare equals operating costs). The FB price in (10) is the sum of crowding costs, crowding levies, and scheduling costs. The R now follows (11). Note that N^t is endogenous as prices change and demand is price sensitive.

$$P_i^{FB} = C_{CR}[r] + C_{SD}[r] + \rho_r^t[r] = 2g \cdot N^t / R - \beta \cdot h(R-1)/2 \tag{10}$$

$$R = \text{Floor}\left[\frac{1 - \sqrt{16g \cdot N^t + \beta \cdot h}}{2\sqrt{\beta \cdot h}} \right] \tag{11}$$

If road congestion is not priced, we are in a second-best situation. Then, the welfare-maximising time-invariant levy ($\bar{\rho}^t$) is negative (i.e. a subsidy). This lowers the number of car drivers relative to when the term is zero, which increases welfare as marginal social costs on the road are above private costs. In contrast, the private operator adds a positive time-invariant term that maximises its profits. With FB pricing there is no time-invariant addition.

Pricing of congestion externalities increases the prices of train travel. This differs from bottleneck congestion, where tolling lowers prices for drivers with a high ratio and raises them for drivers with a low ratio. Similarly, with static flow congestion or Chu's (1995) dynamic-flow congestion, pricing is also less beneficial as tolling raises prices. With bottleneck congestion there is the pure waste of queuing, and pricing eliminates this. Crowding and flow congestion do not entail such a pure deadweight loss (see also Vickrey, 1969).

3.2. Proportional heterogeneity and rail pricing

The foregoing illustrated that ratio heterogeneity has a little impact on the train model since travel time is fixed. Now we find that proportional heterogeneity has more effect. Users with the highest values use service 1 that arrives at t^*, as they care most about schedule delays. Users with the lowest values use service R. Thus already without congestion pricing users arrive ordered by β, while on the road this only happens with pricing. This self-ordering lowers the mean scheduling cost and thereby the mean price, and more so the more heterogeneity there is. Generally with proportional heterogeneity more services are used than with homogeneity; since users with the lowest values then have lower values of schedule delay and care relatively more about crowding costs, and thus arrive earlier to lower these.

Figure 4: Crowding costs per service that would make each service attractive to use
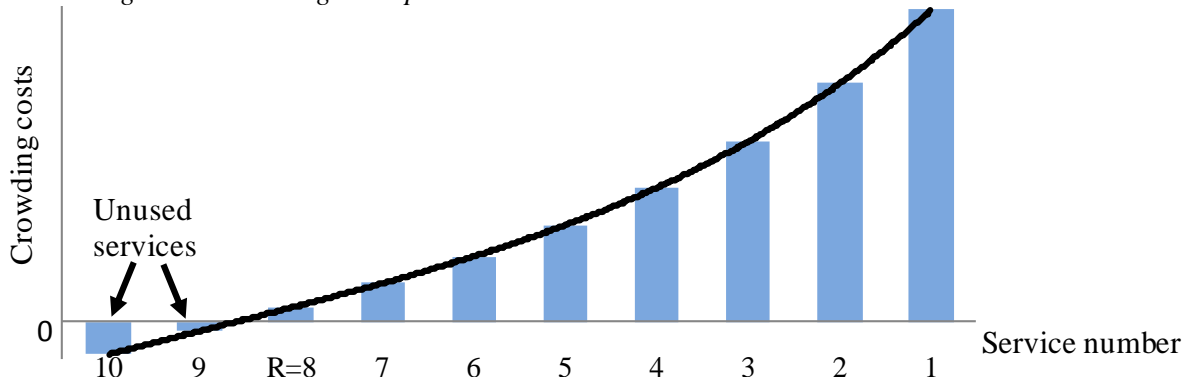
Crowding costs now change non-linearly over time, since users travel separated by value of schedule delay. This is also what the black trend line in Figure 4 indicates. A consequence of this is that we were unable to find closed-form solutions for the number of users per run and the number of services used. With ratio heterogeneity, we used the fixed β to solve the model; now β varies. For each pair of services r and $r+1$ there is a value of schedule delay that implies indifference between the two. For all other users of service r, the price of using service $r+1$ is higher than that of service r, because of the higher schedule delay of service $r+1$. For all other users of service $r+1$, the price of service r is higher. Using these indifferent users it is possible to numerically solve the model. Rail-travel prices now follow

$$P_i^{NCG} = C_C[r] + C_{SD}[r, \beta_i] = g \cdot N_r^t + h(r-1)\beta_i. \tag{12}$$

Crowding externalities are not affected by proportional heterogeneity, and hence service r's crowding levy still follows $g \cdot N_r^t$. The FB price is

$$P_i^{FB} = C_C[r] + C_{SD}[r, \beta_i] + \rho_r^t[r] = gN_r^t + h(r-1)\beta_i + gN_r^t. \tag{13}$$

4. Set-up numerical models

Now we discuss the calibration of the model. We assume that the two modes are imperfect substitutes. The inverse demand for user type i of mode j is

$$D_{ij}^{-1} = A_{ij} + \frac{b_{ij}}{B_j} n_{ij} + \frac{c_{ij}}{C_j} n_{ik}, \tag{14}$$

where n_{il} is the number of type i users on mode l. Demand for mode j by type i users decreases with the price of mode j, and increases with the price of mode k. Hence, both b_{ij}/B_j and c_{ij}/C_j must be negative. If c_{ij}/C_j were positive, the two modes would be complements.[4] Total consumer surplus is the integral of the surplus per type. Welfare is total consumer surplus plus profits.

In the NCP equilibrium, there are 9000 road users and 5000 train users. The bottleneck capacity is 4500 cars per hour; hence the road peak lasts 2 hours. The crowding coefficient, g, is 1/200. The fixed headway, h, between trains is 15 minutes. The mean value of time is €10.50. Free-flow car-travel time is 30 minutes. Fixed train-travel time is 45 minutes. The operating costs of the car are €6.50. The rail fare equals per-user operating costs of €7.50. We have no fixed costs to rail travel, although Section 7 has a sensitivity analysis on the effect of fixed cost.

The average value of schedule delay is €5.00. In the base-case parameterisation of proportional heterogeneity, the NCP equilibrium has a value of schedule delay that is uniformly distributed between €2.00 and €8.00, and the value of time is 2.1 times the value of schedule delay. With ratio heterogeneity, the value of time ranges between €5.50 and €15.50, and the value of schedule delay is fixed at €5.00. Without congestion pricing, the distribution of users on the road and rail follows the same probability density function, this is so with ratio and proportional heterogeneity. Pricing changes the prices, and thus results in different distributions on the two modes.

---

[4] Following Kraus (2003), we impose that the cross-substitution effects are the same for both modes for all types of users (i.e. $\partial D_{ij}/\partial P_{ik} = \partial D_{ik}/\partial P_{ij}$), and there are no income effects. Users are utility maximisers, and consequently $\partial D_{ij}/\partial P_{ij} \cdot \partial D_{ik}/\partial P_{ik} > \partial D_{ij}/\partial P_{ik} \cdot \partial D_{ik}/\partial P_{ij}$ must hold. Under these assumptions, two-good consumer surplus for a type of user is the line integral of the two inverse demands minus total costs for this type of users.

Finally, the models are calibrated so that the weighted average of the own-price elasticity, in the NCP equilibrium, is −0.5 for both car and train; the cross-elasticity of rail-travel demand w.r.t. price of car travel is 0.1.[5]

5. Numerical pricing model with proportional heterogeneity

Table 1 presents the regimes we study. In the NCP regime there is no congestion pricing and the rail fare equals marginal operating cost. With second-best road pricing, the road has a time-variant toll—which prices the externality—and a second-best time-invariant term. With welfare maximisation by "pricing the road only" (CW), this second term is negative and attracts users away from the suboptimally priced train. With "profit maximisation on the road only" (CP), the term is positive. The second-best rail-welfare (TW) and profit (TP) maximisation schemes have similar set-ups. With first-best (FB) pricing, the road toll prices congestion externalities; the train has a levy that equals the crowding externalities and a fare that equals operating cost. In practice, there is only a single train fare, but we distinguish different components for ease of discussion. Because we were unable to find closed-form solutions for the regimes, we present numerical results for the base-case parameterisation and sensitivity analyses on the effect of heterogeneity.

Table 1: Description of the pricing regimes

Policy	Meaning	Rail pricing	Road Pricing
NCP	No Congestion Pricing	Only a marginal rail operating cost fare (v^t)	No-toll
TW	Train Welfare	Welfare-maximising second-best congestion levy + a marginal-rail-operating -cost fare	No-toll
TP	Train Profit	Rail-profit-maximising second-best congestion levy + a marginal-rail-operating -cost fare	No-toll
CW	Car Welfare	Only a marginal-rail-operating -cost fare	Welfare-maximising second-best congestion tolling
CP	Car Profit	Only a marginal-rail-operating -cost fare	Road-profit-maximising second-best congestion tolling
FB	First-best	First-best congestion levy +marginal-rail-operating -cost	First-best congestion toll

5.1. Base case no-congestion-pricing (NCP) equilibrium

This section discusses the pricing regimes in Table 1 under the proportional heterogeneity from Vickrey (1973). Figure 5 (left) shows NCP generalised prices (or prices for brevity) excluding monetary costs and free-flow travel time, the right part includes these. The price of train travel is increasing in β and piecewise linear with kinks at the values that are indifferent between two services, although the latter observation is hard to detect visually. Users to the left and right of a kink use different services. For all users in service 1, the price excluding travel-time costs is the same. Prices of car travel increase linearly with β.

5.2. Congestion pricing and proportional heterogeneity

Table 2 discusses the aggregate results of the different policies. The road-only-pricing CW attains 95.8% of the first-best gain, while the rail-only TW attains only 3.4%. Road tolling lowers prices; rail levying increases prices. A reason for these differences is that the mean NCP (No-Congestion-Pricing) externality on the road is €10.00, whereas in the train it is €6.04. The lower externality is partly due to the lower number of train users. Yet, the main difference between

bottleneck and crowding pricing is that bottleneck congestion has a pure waste from queuing, and tolling removes this, while for the train not all congestion is a pure loss.

Figure 5: Prices in the no-congestion-pricing (NCP) equilibrium excluding monetary costs and free-flow travel time (left) and including (right)
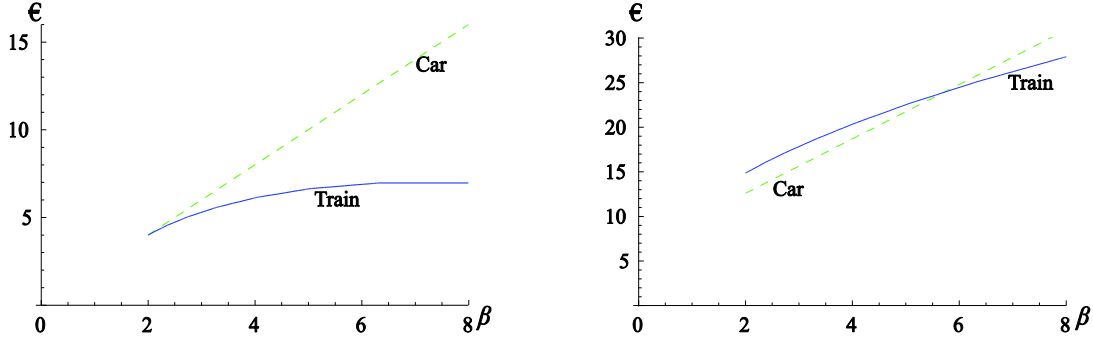

Table 2: Base case effects of the policies with proportional heterogeneity

	NCP	TW	TP	CW	CP	FB
Time-invariant train levy	-	−€2.97	€22.65	-	-	-
Time-invariant road toll	-	-	-	−€0.63	€24.51	-
Mean car-travel price (E[P^C])	€21.75	€21.75	€22.55	€19.70	€41.14	€20.32
Mean train-travel price (E[P^t])	€22.15	€21.98	€ 44.79	€22.04	€22.55	€24.54
Total toll revenue	-	-	-	€34,284	€136,059	€40,062
Total crowding levy revenue	-	€1,232	€61,783	-	-	€14,394
Consumer Surplus	€414,073	€414,788	€321,672	€434,714	€277,826	€416,803
Welfare	€414,073	€416,020	€383,455	€468,997	€413,885	€471,259
Number of car driver (N^c)	9000	8995.6	9,747.6	9447.1	5065.7	9423.8
Number of train users (N^t)	5000	5014.7	2,479.9	4918.7	5724.0	4656.8
Relative efficiency	0	0.034	−0.535	0.958	−0.003	1
Percentage welfare gain	-	0.47%	−7.39%	13.26%	−0.05%	13.81%
Percentage of NCP users who would be better off	-	81.4%	0%	100%	0%	49.5%

Figure 6 shows the change in prices of train travel due to the policies. Figure 7 does this for prices of car travel. The higher β is, the more the price of train travel increases due to congestion pricing. The users with the highest values have to use a service that arrives close to t^*, these services have the highest crowding costs and thus the highest levies. The curves showing the change in the price of train travel due to rail pricing are piecewise linear with upward sloping sections followed by flat sections. The flat sections are for users who use the same service as before, because for them only crowding costs and congestion levies change which they value equally. The sloping sections are for switchers, for whom scheduling costs increase, which depend on β. All users with low values of time and schedule delay switch, since they care relatively little about schedule delays and more about congestion levies.

The second-best TW only prices rail users. It is less harmful for train users than the first-best FB, because it adds a subsidy to the time-variant congestion levy to attract road users. In fact, the TW lowers the prices of train travel for values of schedule delay below €4.85. These users arrive far from t^*, in relatively empty trains, where the total levy is negative.

For car travel, congestion pricing is mostly regressive: the higher the values of schedule delay and time are, the higher the gain from pricing. Conversely, in the train, the higher these values

13

are, the less beneficial pricing is. This is because users with high values have to use a service that arrives close to t^* where crowding cost and levies are high, which raises prices substantially; low values arrive far from t^* when levies are low (or even negative in the second-best TW, which is why this policy lowers the price of users with low values). This suggests that congestion pricing might have a very different political acceptability in different modes.

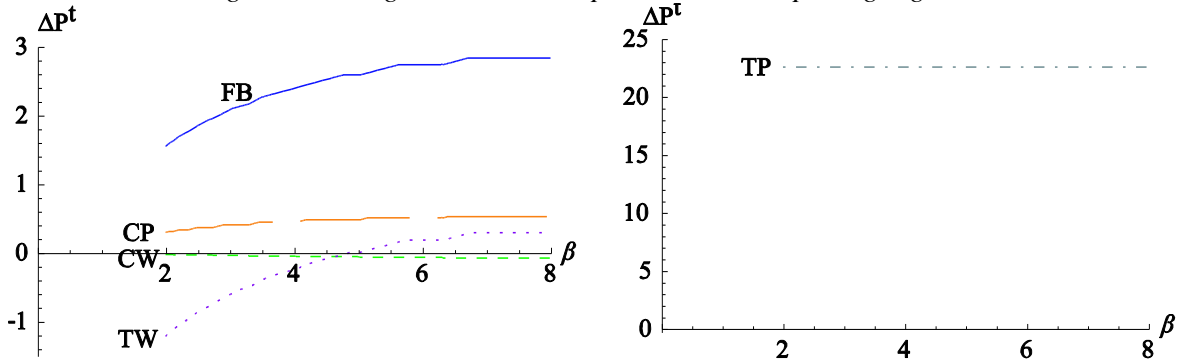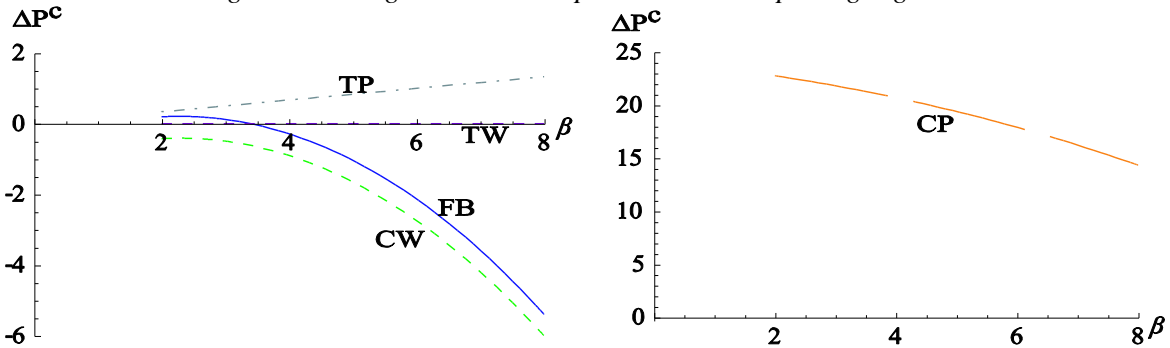*Figure 6: Change in train-travel prices due to the pricing regimes*

Figure 7: Change in car-travel prices due to the pricing regimes

5.3. Sensitivity analysis on the amount of proportional heterogeneity

We measure the amount of proportional heterogeneity by the spread of the uniform distribution of the value of schedule (β), which also determines the spread of the value of time because of the fixed proportions. Without congestion pricing, the average price for car travel is unaffected by the spread. The mean price of train travel decreases with the spread due to the self-ordering of users just as happens on the road with pricing: train users with high values arrive close to t^* when delays are low, users with low values arrive far away when delays are large. Table 3 discusses the NCP equilibria with distributions with spreads of 0, 3, 5, 6, and 7. A spread of 0 means homogeneous users. All distributions of β have a mean of €5.00. Consumer surplus decreases with the spread, but this is a result of the calibration, it does not reflect a meaningful effect of heterogeneity.

Table 3: Proportional heterogeneity and the no-congestion-pricing (NCP) case

Spread of the value of schedule delay	Consumer surplus	$E[P^C]$	$E[P^t]$
0	€418,250	€21.75	€23.19
3	€416,241	€21.75	€22.69
5	€414,814	€21.75	€22.34
6 (Base case)	€414,073	€21.75	€22.15
7	€413,260	€21.75	€21.94

14

Table 4 gives the results of first-best pricing for different amounts of heterogeneity. FB prices of car travel decrease and welfare increases with proportional heterogeneity, because the gain from the more efficient arrival order increases. This is consistent with the single mode case. The FB price of train travel also decreases with the spread. Partly this is because the ordering on β reduces scheduling cost; and the larger the spread is, the larger the reduction. Partly it is because road tolling lowers the average price of car travel, thereby lowering demand for train travel.

Table 4: Heterogeneity in the value of schedule delay and first-best FB pricing

Spread of the value of schedule delay	E[P^C]	E[P^t]	Consumer Surplus	Toll revenue	Levy revenue	Welfare	%ΔW	N^c	N^t
3	€21.06	€25.39	€409,819	€43,004	€15,708	€468,531	12.6%	9258	4672
5	€20.56	€24.84	€414,433	€41,046	€14,842	€470,321	13.4%	9369	4662
6 (Base case)	€20.32	€24.54	€416,803	€40,062	€14,394	€471,259	13.8%	9424	4657
7	€20.08	€24.22	€419,138	€39,076	€13,902	€472,116	14.2%	9479	4651

Table 4 showed that the welfare gain of FB pricing increases with proportional heterogeneity. Table 5 gives percentage welfare gains of the policies. Table 6 reports relative efficiencies: i.e. the welfare gain of a policy from the No-Congestion-Pricing case relative to the first-best gain. The welfare gain of second-best CW road pricing increases with the spread. Still, its relative efficiency decreases slightly. This differs from second-best pricing of two roads that are perfect substitutes in Van den Berg and Verhoef (2011b), where the relative efficiency increases with proportional heterogeneity. The difference lies in that train and car are imperfect substitutes. With perfect substitutes, users with the highest values of time and schedule flock to the priced link, as it is most beneficial to them. With imperfect substitutes, the priced link is used by all values of time and schedule delay, even though for many values the unpriced link has a far lower price and tolling substantially raises the car-travel price. Furthermore, with imperfect substitutes, there are also users with high values who continue to use the unpriced train, even though it has a higher price. Accordingly, for two modes the relative efficiency of CW pricing decreases with the amount of proportional heterogeneity if the degree of substitutability is small enough.

Conversely, the relative efficiency of private CP road pricing increases with proportional heterogeneity. The difference between the CW and CP is because with profit maximisation the number of drivers drops substantially and primarily high-values drivers continue to use the road. Hence, the effect of the imperfect substitutes is less with profit maximisation. Moreover, more proportional heterogeneity increases the mean values of time and schedule delay on the private road, making its travel time and schedule delay savings more valuable. Finally, more heterogeneity lowers the prices of train travel, increasing the competition the private road faces and thereby lowering the mark-up the firm can ask.

The relative efficiency of private TP rail-only pricing increases with the spread. This is predominantly because the FB gain increases, which makes the TP's welfare loss relatively smaller. TP's welfare gain only increases slightly. With a larger spread, the lowest values of schedule delay are lower. This enables users with low values to use earlier train services. This easier shift to earlier services limits the market power of the operator, inducing a lower time-invariant levy. TW second-best rail pricing is hardly affected by proportional heterogeneity.

Table 5: Effect proportional heterogeneity on percentage welfare gains

Spread of β	TW	TP	CW	CP	FB
0	0.62%	-7.41%	11.01%	-1.55%	11.42%
3	0.50%	-7.41%	11.97%	-0.87%	12.56%
5	0.48%	-7.40%	12.85%	-0.34%	13.38%
6	0.47%	-7.39%	13.23%	-0.05%	13.81%
7	0.45%	-7.39%	13.57%	0.25%	14.24%

Table 6: Effect of proportional heterogeneity on relative efficiencies
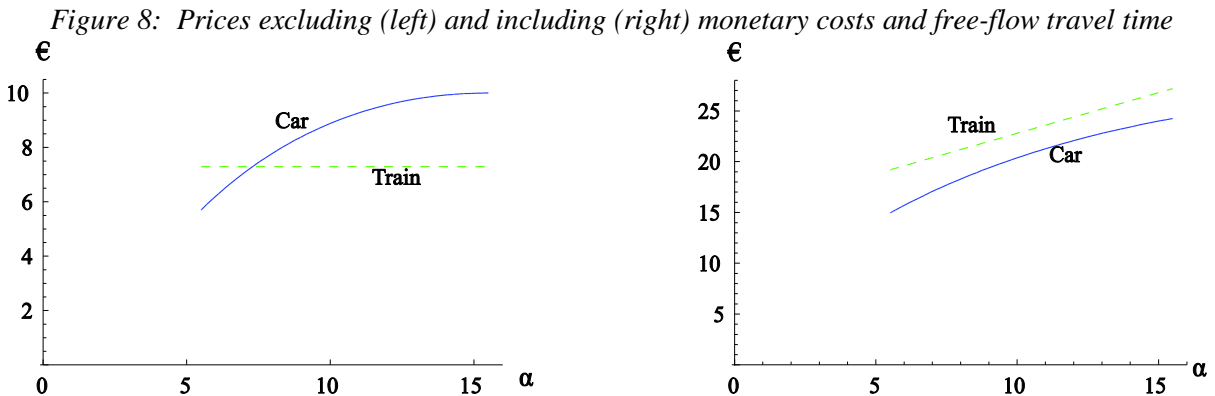
Spread of β	TW	TP	CW	CP
0	0.054	-0.649	0.966	−0.136
3	0.040	-0.590	0.963	−0.069
5	0.036	-0.553	0.960	−0.025
6 (Base case)	0.034	-0.532	0.958	−0.003
7	0.032	-0.519	0.953	0.017

5.4. Conclusions on proportional heterogeneity

On the road without congestion pricing, users arrive ordered by the common ratio α_i/β_i, and hence arrival times are undefined. With tolling, drivers arrive ordered by β, with the highest values arriving closest to t^*. Tolling thus makes the arrival order more efficient, lowering scheduling costs. In the train, this extra efficiency gain does not occur, since users always arrive ordered by β. Due to the extra efficiency gain on the road, the welfare gain of first-best pricing increases with proportional heterogeneity. The relative efficiency of "welfare maximisation by only pricing the road" decreases with proportional heterogeneity; whereas with two perfect substitute roads, the relative efficiency increases. Since crowding-congestion pricing does not alter the arrival order in the train, proportional heterogeneity has little effect on schemes that only price the train.

6. Numerical pricing model with ratio heterogeneity

Now we look at "ratio heterogeneity", which involves the ratio of the value of time (α) to the value of schedule delay (β), with a fixed value of schedule delay: $\mu_i \equiv \alpha_i/\beta$. In the NCP case, the value of time is uniformly distributed between €5.50 and €15.50; the value of schedule delay is fixed at €5.00. Section 2 noted that road externalities decrease with ratio heterogeneity, which lowers the gain from road tolling. The mean price of rail travel is unaffected by ratio heterogeneity as the train-travel time is fixed. The price of train travel for a user is constant across services; with proportional heterogeneity, the price differs over services due to the differences in β_i.

Figure 8: Prices excluding (left) and including (right) monetary costs and free-flow travel time

16

6.1. Base-case no-congestion-pricing and first-best equilibria with ratio heterogeneity

Figure 8(left) gives the NCP prices excluding free-flow travel time and fixed costs, the right side includes these. The price of rail travel increases linearly with the value of time. The price of car travel increases concavely. The price of car travel is the highest for drivers with the highest ratio. In this calibration, prices for train travel are higher than for car travel for all users. Nevertheless, since the two modes are imperfect substitutes, all values of time are represented in the train.

Again, road pricing gives a higher welfare gain than rail pricing. As Table 7 shows, CW pricing attains 96% of the first-best FB gain. The TW policy attains only 2%. The train has about 35% of the users, so this large difference in welfare gain reflects more than just differences in the percentage of users who face congestion pricing.

Congestion pricing is less beneficial to the average user with ratio heterogeneity than with proportional heterogeneity or homogeneity. With proportional heterogeneity, consumer surplus increases due to TW, CW, and FB policies. With ratio heterogeneity, only "pricing the road only" CW increases consumer surplus before the distribution of revenues. The percentage of NCP (No-Congestion-Pricing) users who would face a lower price with pricing is also lower now. In Table 7, no user gains from TW pricing. With proportional heterogeneity in Table 2, all car drivers and 47% of rail users would gain. Now 73% of NCP users would gain due to CW pricing; with proportional heterogeneity all users gain. Finally, for all schemes, the welfare gain is lower with ratio heterogeneity than with proportional heterogeneity.

Table 7: Base case effects of the policies with ratio heterogeneity

	NCP	TW	TP	CW	CP	FB
Time-invariant train levy	-	−€2.02	€23.72	-	-	-
Time-invariant road toll	-	-	-	−€0.68	€22.17	-
Mean car-travel price ($E[P^C]$)	€20.43	€20.47	€21.16	€20.99	€39.40	€21.64
Mean train-travel price ($E[P^t]$)	€23.19	€24.30	€46.87	€23.19	€23.69	€26.13
Total toll revenue	-	-		€37,843	€120,877	€43,603
Total crowding levy revenue	-	€6,588	€64,949	-	-	€15,717
Consumer Surplus	€398,705	€392,906	€303,318	€394,216	€265,242	€374,120
Welfare	€398,705	€399,494	€368,267	€432,059	€386,119	€433,441
Number of car driver (N^c)	9000	9035.5	9757.5	8887.7	4860.4	8859.2
Number of train users (N^t)	5000	4881.6	2475.9	5018.3	5677.6	4726.0
Relative efficiency	0	0.023	−0.876	0.960	−0.362	1
Percentage welfare gain	-	0.20%	−7.63%	8.37%	−3.16%	8.71%
Percentage of NCP users who would be better off	-	0%	0%	30.3%	0%	13.8%

Figures 9 and 10 plot the price changes relative to the NCP case for train and car users. For train travel, the curves are flat lines, because only the cost of the fixed travel time depends on α_i. This contrasts with proportional heterogeneity, where these curves are piecewise linear.

All rail users are affected to the same extent by pricing. Hence, pricing does not have distributional effects between rail users. Conversely, on the road there are distributional effect: the higher the value of time (α) is, the less harmful is road pricing. This is because the travel time saving from tolling is more valuable with a higher α. Rail only pricing raises travel times, and therefore it is more harmful, the higher the value of time.

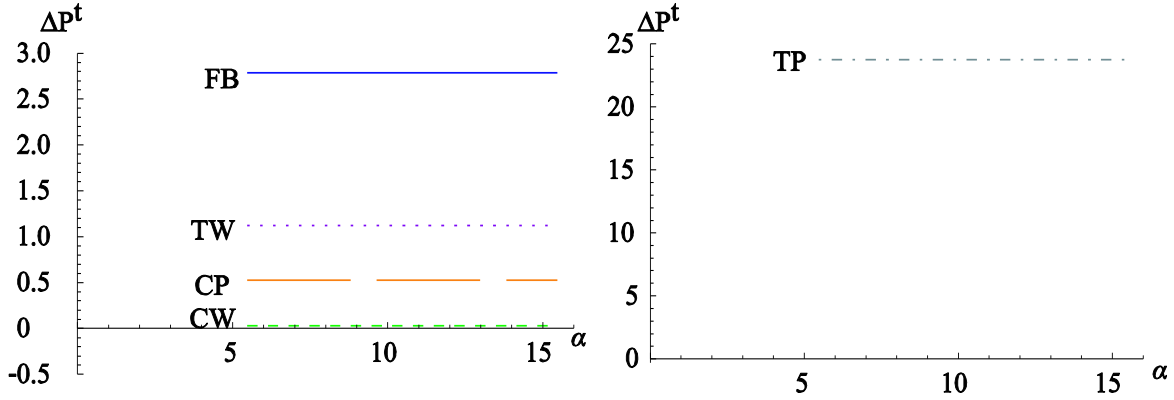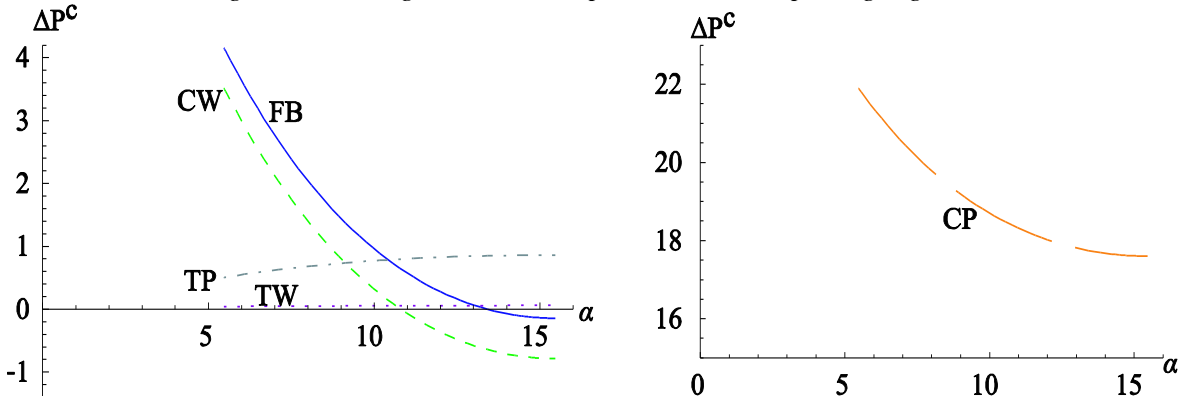*Figure 9: Change in train-travel prices due to the pricing regimes*

Figure 10: Change in car-travel prices due to the pricing regimes

6.2. Sensitivity analysis on the amount of ratio heterogeneity

Table 8 analyses how heterogeneity in the ratio of values of time and schedule delay affects the NCP equilibrium. The base-case spread of uniform distribution of α is 10; the other spreads are 10.8, 9.2, 6, and 0.[6] The mean value of time is always €10.05. Consumer surplus increases with the spread, just as it did with proportional heterogeneity. But this is due to the calibration; it is not a meaningful effect of heterogeneity. The average price of car travel decreases with the spread of α, because the congestion externalities decrease. The mean price of train travel is unaffected, as the train travel time is fixed.

Table 8: Ratio heterogeneity and the no-congestion-pricing (NCP) case

Spread of the value of time	CS	E[P^C]	E[P^t]
0	€418,250	€21.75	€23.19
6	€405,734	€20.78	€23.19
9.2	€400,052	€20.52	€23.19
10 (base case)	€398,705	€20.43	€23.19
10.8	€396,925	€20.34	€23.19

Table 9 studies the effect of the spread of α on the first-best pricing. Different from with proportional heterogeneity, with ratio heterogeneity FB pricing raises the average price for car travel. Still, the price increase for rail travel is larger. This is, again, because congestion pricing is

---

[6] For a deterministic bottleneck equilibrium $\alpha_i > \beta$ must hold for all i. Hence, the maximum spread is 10.99. However, for a spread above 10.8, the numerical model is unstable. Therefore, 10.8 is the maximum spread.

less beneficial with crowding congestion than with bottleneck congestion. FB welfare decreases with the spread; because as Section 2 argued road-congestion externalities and prices decrease which means that there is less to gain from tolling.

Table 9: Ratio heterogeneity and first-best pricing

Spread of the value of time	$E[P^C]$	$E[P^t]$	Toll revenue	Fare revenue	CS	Welfare	%ΔW	N^c	N^t
6	€21.71	€26.09	€44,461	€15,656	€384,481	€444,598	9.6%	8946.0	4712.9
9.2	€21.65	€26.13	€43,770	€15,706	€376,097	€435,573	8.9%	8876.1	4723.5
10 (base case)	€21.64	€26.13	€43,603	€15,717	€374,120	€433,441	8.7%	8859.2	4726.0
10.8	€21.60	€26.13	€43,448	€15,726	€371,768	€430,942	8.6%	8843.4	4727.9

Table 10 gives the welfare gains for different spreads of *α,* and Table 11 the relative efficiencies. The gain from FB pricing decreases with the spread. The relative efficiencies of "road-only pricing" CW and CP also decrease with this heterogeneity. This corresponds with Van den Berg and Verhoef (2011a), where the relative efficiency of "single-link pricing" lowers with this heterogeneity. The welfare gain of TP rail operator's profit maximisation is hardly affected by ratio heterogeneity. Still, its relative efficiency is more negative with ratio heterogeneity, because the first-best gain decreases, and the first-best gain is in the denominator of the relative efficiency. The welfare gain and relative efficiency of TW "welfare maximisation by only pricing the train" marginally decrease with the spread of *α*. With a larger spread, road externalities are less, and thus there is less to gain from attracting car drivers away from the road.

Table 10: Effect of ratio heterogeneity on percentage welfare gains

Spread of α	TW	TP	CW	CP	FB
0	0.62%	-7.41%	11.01%	-1.55%	11.42%
6	0.33%	-7.59%	9.24%	-2.58%	9.58%
9.2	0.23%	-7.62%	8.53%	-3.05%	8.88%
10	0.20%	-7.63%	8.37%	-3.16%	8.71%
10.8	0.19%	-7.65%	8.56%	-3.26%	8.57%

Table 11: Effect of ratio heterogeneity on relative efficiencies

Spread of α	TW	TP	CW	CP
0	0.054	-0.649	0.966	-0.136
6	0.034	-0.792	0.965	-0.269
9.2	0.026	-0.859	0.961	-0.343
10 (base case)	0.023	-0.876	0.960	-0.362
10.8	0.022	-0.893	0.958	-0.381

6.3. Conclusions on ratio heterogeneity

The welfare gains of all pricing schemes decrease with ratio α_i/β heterogeneity. Moreover, the relative efficiency of pricing only the road or train also decreases. With two roads, this heterogeneity has the same effect. This contrasts with Section 4, where the effect of proportional heterogeneity differs between a two-roads and two-modes network.

7. Some further sensitivity analyses

7.1 Fixed cost

Tabuchi (1993) analysed road and rail pricing when the rail operator's costs have a variable and fixed component; our operator only has variable costs. To test the effects of this limitation, we redefined our models such that rail only has fixed cost. The fixed costs are such that the average-operating cost in the NCP equilibrium equals the marginal-operating cost in the rest of the paper.

The effects of congestion pricing, and how heterogeneity affects these, are on the whole the same as with marginal costs. A difference is that with fixed cost and second-best road-pricing CW, a positive time-invariant toll is added instead of a negative one. The "positive externality" due to that more train users lowers average operating cost dominates the negative crowding externality. Accordingly, it is attractive to push users to the train using the positive term. This result was also found by Tabuchi (1993). In the previous sections, the opposite effect (see, e.g., Braid, 1996) held that the priced link has a negative term to account for the congestion spillovers to the unpriced link

7.2 Price elasticities

The effects of changing own-price elasticities are in line with the earlier literature. In particular, private pricing is increasingly harmful with less elastic demand, because the company's market power increases. Changing the cross-elasticity is more interesting. It could be that the relative efficiency of second-best TW train pricing is low because the cross-elasticity is low, which makes it difficult to attract drivers away from the unpriced road.

In the ratio heterogeneity model, the highest cross-elasticity of rail demand w.r.t. the price of car travel consistent with utility maximisation is 0.379 (see also footnote4). In the base case, this elasticity was 0.1. This almost quadrupling raises the TW's relative efficiency by 314% to 0.095. Still, its relative efficiency remains low, and far below the relative efficiency of road-only pricing, which also increases. With proportional heterogeneity, the maximum cross-elasticity is 0.37. Here the relative efficiency for TW increases by 201% to 0.10. Hence, the limited cross-elasticity is not the most important reason for the low gain from rail pricing.

7.3 Crowding costs

Another reason for the low gain of rail pricing could be the relatively low value of crowding. Although it seems likely that crowded trains are indeed a discomfort, we have little information on the value of crowding for our model. As a test we double the crowding coefficient g. This less than doubles crowding costs, as users respond by arriving more spread out over the day. In the ratio heterogeneity model, the mean NCP price increases by €3.27 to €26.46. Yet, the relative efficiency of TW pricing only increases by 11% to 0.025. The relative efficiency of private TP increases from −0.876 to −0.38. With proportional heterogeneity, the doubling of g raises TW's relative efficiency by 17% to 0.04.

All this suggest that the value of crowding congestion does have an effect. But the primary reason for the low gain from rail pricing relative to road pricing seems the difference between bottleneck and crowding congestion. With bottleneck congestion, pricing removes the pure waste that is queuing. The train has no purely wasteful congestion.

9. Conclusion

This paper analysed congestion pricing in a road and rail network, where the train and car are imperfect substitutes. If the two modes are closer substitutes, this generally increases the gain of second-best pricing of one of the two modes, because it becomes easier to attract users away from the link where marginal private costs are above social cost. We separately studied "ratio heterogeneity" in the ratio of value of time to value of schedule delay and "proportional heterogeneity" that varies both values in fixed proportions.

With proportional heterogeneity, road pricing makes the arrival order more efficient, thereby lowering scheduling costs. Accordingly, the welfare gain of road pricing increases with this type of heterogeneity. The relative efficiency of private road-only pricing rises with proportional heterogeneity, while for welfare maximisation the relative efficiency decreases. This contrast with the finding with two roads, where the relative efficiency of "single-link pricing" was found to increase with proportional heterogeneity, because more heterogeneity means that the drivers with high values—for whom the pay-lane lowers schedule delays and travel time—have higher mean values which makes the savings they attain more valuable (Van den Berg and Verhoef, 2011b). The difference lies in that two roads are perfect substitutes, while car and train are not. With perfect substitutes, users with the highest values of time and schedule flock to the priced link. With imperfect substitutes, the priced link is used by users with all values, even though for some the unpriced link has a lower price and tolling substantially raises the price of car travel. Therefore, the beneficial effect on tolling of proportional heterogeneity does not occur if the extent of substitutability is large enough.

On the rail link, users always arrive ordered by their value of schedule. Thus, pricing does not improve the arrival order. The welfare gain of rail-congestion pricing is hardly affected by proportional heterogeneity. Still, the relative efficiency of private rail-only-pricing slightly decreases with this type of heterogeneity.

A general conclusion is that the gain of congestion pricing can be lower or higher with heterogeneity depending on the types and extents of the heterogeneity. Nevertheless, pricing always leads to a welfare improvement. Perhaps more important for policy are the distributional effects, which differ strongly between road and rail. In the train, pricing either has no distributional effects or is more harmful, the higher a user's values of time and schedule delay are. On the road, it is generally the case that the higher a user's value of time or schedule delay is, the more beneficial is pricing. Only private rail pricing is more damaging for a road user, the higher her value of time or schedule delay, because private rail pricing increases congestion on the road, and this is more costly with a higher value.

An interesting extension of our model would be heterogeneity in the value of crowding. Then, users with high crowding values would arrive early in relatively empty trains, and low-value users would arrive close to the preferred arrival time. Hence, crowding externalities will be lower with heterogeneity, thus lowering the gain of rail pricing. Accordingly, the effect of value-of-crowding heterogeneity in the rail model might be similar to that of ratio heterogeneity in the bottleneck model. Then it would also be interesting to introduce first and second class coaches, as users could self-select by value of crowding, thereby lowering total costs. If we used a crowding multiplication of the value of time following the empirical literature instead of separate crowding and travel time costs, the effect of heterogeneity on rail and road pricing might also be similar.

Another interesting extension is different objectives for the rail operator. For example, that the operator has fixed costs and maximises welfare under a minimum profit constraint. Alternatively, the company could be a profit maximiser but face fare or maximum profit regulations.

Acknowledgements

References

Arnott, R., de Palma, A., Lindsey, R., 1988. Schedule delay and departure time decisions with heterogeneous commuters. Transportation Research Record 1197, 56–67.

Arnott, R., de Palma, A., Lindsey, R., 1994. The welfare effects of congestion tolls with heterogeneous commuters. Journal of Transport Economics and Policy 28(2), 139–161.

Arnott, R, Kraus, M., 1993. The Ramsey problem for congestible facilities. Journal of Public Economics 50(3), 371–396.

Arnott, R, Kraus, M., 1995. Financing capacity in the bottleneck model. Journal of Urban Economics 38(3), 272–290.

Arnott, R., Yan, A., 2000. The two-mode problem: Second-best pricing and capacity. Review of Urban and Regional Development Studies 12, 170–199.

Braid, R.M., 1996. Peak-load pricing of a transportation route with an unpriced substitute. Journal of Urban Economics 40(2), 179–197.

Chu, X., 1995. Endogenous trip scheduling: the Henderson approach reformulated and compared with the Vickrey approach. Journal of Urban Economics 37(3), 324–343.

de Palma, A., Lindsey, R, 2002. Congestion pricing in the morning and evening peaks: a comparison using the Bottleneck Model. In: Proceedings of the 39th Annual Conference of the Canadian Transportation Research Forum: 2002 Transportation Visioning - 2002 and Beyond, Vancouver, Canada, May 2002, 179-193, Canada, 9-12 May 2004, 179–193.

Huang, H.-J., 2000. Fares and tolls in a competitive system with transit and highway: the case with two groups of commuters. Transportation Research Part E 36(4), 267–284.

Kraus, M., 1991. Discomfort externalities and marginal cost transit fares. Journal of Urban Economics 29(2), 249–259.

Kraus, M., 2003. A new look at the two-mode problem. Journal of Urban Economics 54(3), 511–530.

Kraus, M., Yoshida, Y., 2002. The commuter's time-of-use decision and optimal pricing and service in urban mass transit. Journal of Urban Economics 51(1), 170–195.

Layard, R., 1977. The distributional effects of congestion taxes. Economica 44(175), 297–304.

Lindsey, R., 2004. Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes. Transportation Science 38(3), 293–314.

Mohring, H., 1972. Optimization and scale economies in urban bus transportation. American Economic Review 62(4), 591–604.

Rouwendal, J., Verhoef, E.T., 2004. Second best pricing for imperfect substitutes in urban networks. In G. Santos (Ed), Road pricing: Theory and evidence. Research in transportation economics (Vol. 9). pp. 27–60. Oxford: Elsevier.

Small, K.A., Verhoef, E.T., 2007. The Economics of Urban Transportation. London, Routledge.

Small, K.A., Yan, J., 2001. The value of "value pricing" of roads: second-best pricing and product differentiation. Journal of Urban Economics 49(2), 310–336.

Tabuchi, T., 1993. Bottleneck congestion and modal split. Journal of Urban Economics 34(3), 414–431.

van den Berg, V.A.C., Verhoef, E.T., 2011a. Congestion tolling in the bottleneck model with heterogeneous values of time. Transportation Research Part B 45(1), 60–70.

van den Berg, V.A.C., Verhoef, E.T., 2011b. Winning or losing from dynamic bottleneck congestion pricing? The distributional effects of road pricing with heterogeneity in values of time and schedule delay. Journal of Public Economics, in press: doi:10.1016/j.jpubeco.2010.12.003

Verhoef, E.T., Small, K.A., 2004. Product differentiation on roads: constrained congestion pricing with heterogeneous users. Journal of Transport Economics Policy 38(1), 127–156.

Vickey, W.S., 1969. Congestion theory and transport investment. American Economic Review (Papers and Proceedings) 59(2), 251–260.

Vickrey, W.S., 1973. Pricing, metering, and efficiently using urban transportation facilities. Highway Research Record 476, 36–48.

Wardman, M., Whelan, G.A., 2011. 20 Years of rail crowding valuation studies: evidence and lessons from British experience. Transport Reviews, in press.

Wu, W.-X., Huang, H.-J., 2010. Equilibrium and modal split in a competitive highway/transit system under different road-use pricing regimes. Workingpaper of Beijing University of Aeronautics and Astronautics, version of 2 december 2010.