TI 2010-016/3
Tinbergen Institute Discussion Paper

Why Congestion Tolling could be Good for the Consumer: The Effects of Heterogeneity in the Values of Schedule Delay and Time on the Effects of Tolling

Vincent van den Berg
Erik T. Verhoef

VU University Amsterdam, and Tinbergen Institute.

Tinbergen Institute

The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Why congestion tolling could be good for the consumer:

The effects of heterogeneity in the values of schedule delay and time on the effects of tolling

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

a: Department of Spatial Economics, VU University, De Boelelaan 1105, 1081HV Amsterdam, The Netherlands
#: email: everhoef@feweb.vu.nl
*: Corresponding author: email: vberg@feweb.vu.nl, tel: +31 20 598 6049, Fax: +31 20 598 6004

Abstract

In studying congestion tolling, it is important to account for heterogeneity in preferences of drivers, as ignoring it can bias the welfare gains. We analyse the effects of tolling, in the bottleneck model, with continuous heterogeneity in the value of time and schedule delay. The welfare gain of a time-variant toll increases with heterogeneity in the value of schedule delay. With heterogeneity, tolling makes the arrival ordering more efficient, and this lowers scheduling costs. If there is not much more heterogeneity in the value of time than in the value of schedule delay, then first-best tolling decreases the generalised price for most users. In our model, first-best tolling is not most detrimental for the lowest values of time and schedule delay: it raises prices more for users with an average value of schedule delay and a slightly larger value of time. Further, the lowest values of time are among those who gain most from a public pay-lane.

JEL codes: R41, R48, H23

Keywords: biases in calculated welfare effects, bottleneck model, distributional effects, heterogeneity in the value of schedule delay, heterogeneity in the value of time, second-best tolls

1. Introduction

Traffic congestion is an important problem in many societies. A possibility to alleviate this issue is to toll the congestion externalities caused by drivers. In analysing tolling, it can be dangerous to ignore heterogeneity: Arnott et al. (1988) find that ignoring heterogeneity, by just using the mean parameters, biases the welfare effects of tolling. Arnott et al. (1994) use two group heterogeneity. They find that first-best tolling has distributional effects: it raises prices of the lowest value of time, while prices for the highest value of time are unaffected or decrease. Lindsey (2004) notes that the congestion costs a user imposes on other users decreases with her value of time. Hence, heterogeneity not only causes tolling to have distributional effects, it also affects the congestion externalities and the welfare effects of tolling.

Vickrey (1973) analyses the case where the values of time, schedule delay early and schedule delay late vary proportionally (i.e the ratios of the three parameters are the same for all drivers). All drivers except those with the lowest values are better off with first-best (FB) tolling. Xiao et al. (2009) use the same assumption for the parameters as Vickrey (1973). They find that one-step (coarse) tolling has higher welfare gain with heterogeneity set-up and is Pareto-improving.

Small and Yan (2001) and Verhoef and Small (2004) use static flow congestion. They find that public pay-lane's relative efficiency increases with heterogeneity in the value of time. The highest values of time use the pay-lane. The pay-lane's mean value of time increases with heterogeneity, making the pay-lane's travel time savings more valuable. Relative efficiency is the welfare gain of a policy from the no-toll situation relative to the public first-best toll's gain.

Conversely, Van den Berg and Verhoef (2009) note that, with bottleneck congestion, a pay-lane's relative efficiency decreases with the heterogeneity in the value of time. Further, with a more heterogeneous value of time, no-toll equilibrium congestion externalities are smaller. Hence, there is less to gain from first-best and pay-lane tolling, and their welfare gains are lower.

This is a pre-print version of the article: van den Berg, V., Verhoef, E.T., 2011. 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, 95(7-8), 983-992, available from http://dx.doi.org/10.1016/j.jpubeco.2010.12.003.

We analyse how *first-best public tolling* and *public* and *private pay-lanes* are affected by heterogeneity in value of schedule delay and in the relative size of the value of time to the value of schedule delay (μ). A type i driver faces a schedule delay if she does not arrive at her preferred arrival time (t*). The value of earlier arrival (*schedule delay early*) is β_i and of later (*schedule delay late*) it is γ_i. There is no heterogeneity between the values of schedule delay early and late. Type i's value of time (α_i) is $\mu_i \cdot \beta_i$. Van den Berg and Verhoef (2009)'s model is a special case of our model where the value of schedule delay is homogeneous. Vickrey (1973)'s model can be interpreted as a version of our model with the relative size of the value of time (μ) fixed.

Arnott et al. (1998) indicate that, in an equilibrium with a queue, users arrive ordered on μ_i. The larger μ_i is, the further i arrives from t*. With an optimal time-variant toll, the entire queue is eliminated. Then, users arrive ordered on their value of schedule delay, with the *highest-β-users* arriving closest to t*. Before the tolling, there were users with high values of schedule delay that were forced to arrive far from t*, whereas with tolling they arrive close to t*. Hence, tolling makes the arrival ordering more efficient, and this decreases average scheduling costs.

We find that the gain from this reordering increases with the heterogeneity in the value of schedule delay. The average generalised price can be lower with first-best tolling than without tolling. For this to be the case, the value of time distribution must not be too much more heterogeneous than the value of schedule delay distribution. The idea that tolling can be good for the average consumer is surprising, as the common thought is that tolling is harmful for consumers. The distributional effects of tolling in our model can also be surprising. The lowest values of time and schedule delay are among those who gain most from the public pay-lane. With a private pay-lane, the low values of schedule delay lose relatively little; the free-lane users with larger values, and even some pay-lane users, face larger price increases.

The next section describes the demand and generalised price equations. Section 3 analyses the no-toll (NT) and tolling equilibria with M discrete user groups. Then, the case of continuous heterogeneity is studied. For this, Section 4 describes the numerical model set up. Section 5 studies the NT equilibrium, Section 6 the first-best public (FB) toll and Section 7 the pay-lanes. Section 8 gives the sensitivity analysis and Section 9 concludes. Table 1 summarises the policies.

Table 1: Abbreviations of the policies

Abbreviation	Description
NT	No toll equilibrium
FB	Welfare maximizing public first-best time-variant toll
PL	Welfare maximizing public pay-lane, with a time-variant toll
PPL	Profit maximizing private pay-lane, with a time-variant toll

2. The demand and generalised price functions

One road connects the origin and destination. All users have the same preferred arrival time (t^*), which is normalised to arrival time (t) is zero. Because the bottleneck's capacity (s) is lower than demand, a queue develops. Drivers face a trade off between *travel time* costs and *scheduling* costs, due to arriving before (*schedule delay early*) or after (*schedule delay late*) t^*. We use square brackets to indicate that something is a function of the variables listed inside brackets. We use round brackets for arithmetic. Equation (1) gives the generalised total price for a type i driver.

$$P_i[t] = CT_i[t] + CSD_i[t] + \upsilon + \tau[t] = \mu_i \cdot \beta_i (T_D[t] + T_f) + Max(-\beta_i t, \gamma_i t) + \upsilon + \tau[t] \qquad (1)$$

The price is the sum of *travel time costs* ($CT_i[t]$), *schedule delay costs* ($CSD_i[t]$), toll ($\tau[t]$) and operating costs (υ). Operating costs are the same for all users. Travel time is the sum of free-flow travel time (T_f) and travel delay ($T_D[t]$). The latter follows from the queue length ($q[t]$) by $T_D[t]=q[t]/s$. The analytical models ignore free-flow travel time and operating costs, the numerical models include them. The toll consists of the time-*variant* toll ($\tau_t[t]$) and time-*invariant toll* ($\overline{\tau}$). The value of schedule delay *late* (γ_i) is a linear function of the value of schedule delay *early* (β_i), following $\gamma_i=\eta\beta_i$. The relative size of the value of time is μ_i. The term *type* indicates all users with the same values of time and schedule delay. Types can be continuously or discretely distributed. For a deterministic equilibrium, the inequality $\mu_i>1$ must hold for all users (Arnott et al., 1991).

Equation (2) gives the inverse demand function. $A+A_i$ is the constant in type i's demand function: A is common to all and A_i is type specific. The slope is determined by B and $b_i[\mu_i,\beta_i]$. The $b_i[\mu_i,\beta_i]$ is assumed to integrate to one, for algebraic ease.

$$D_i = A + A_i - \frac{B}{b_i[\mu_i, \beta_i]}\, n_i \qquad (2)$$

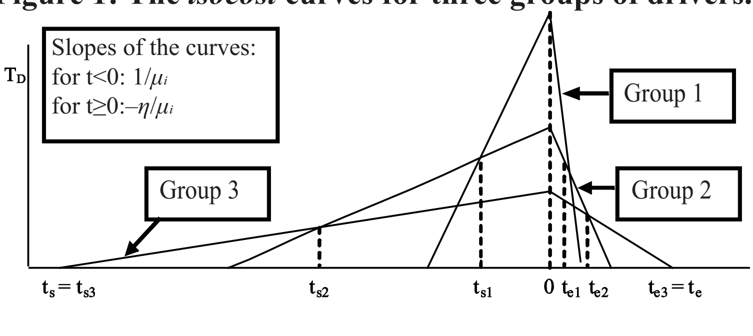
3. Equilibria with discrete heterogeneity

3.1 No-toll (NT) equilibrium with discrete heterogeneity

It is illustrating to start with the case of discrete heterogeneity, as this makes it easier to explain the effects of heterogeneity. This section assumes that each user group has a different μ_i and β_i. If two groups have the same μ_i, they share their NT arrival period and this complicates the mathematical notation. The subsequent sections relax this assumption. The groups are ordered on their μ_i. Arnott et al. (1988) find that, group 1, with the lowest μ_i, arrives closest to t*. Group M drivers, with the highest μ_i, arrive at the greatest distance from t^*.

A group's equilibrium price is constant during the period this group arrives. Outside this period the price is higher. A group's isocost curve gives the combinations of schedule delay and queuing time for which prices are constant over time. Obviously, a different isocost curve applies for a different cost level. The slope of an isocost curve is $1/\mu_i$ before t* and $-\eta/\mu_i$ after. Figure 1 gives the equilibrium isocost curves with three groups. At t_s and t_e, the first and last driver arrive, whereas t_{si} and t_{ei} indicate when group i starts and ends to arrive. If i's equilibrium isocost curve is above the curves of the other groups and not below the x-axes, then at this moment only group i drivers arrive. Hence, group 3 users arrive between t_s and t_{s2}, and between t_{e2} and t_e. Group 1 arrives between t_{s1} and t_{e1}. In equilibrium, the bottleneck operates at capacity during the entire peak. Thus, the peak duration (t_e–t_s) equals N/s. Here N is the number of users. At t_s and t_e group M users arrive and face a zero queue length.

Figure 1: The *isocost* curves for three groups of drivers.



Using the above discussion, group M's price can be derived. Then, the price for group M–1 can be found, and so on for each group. Equations (3) and (4) give the generalised formulas for i's queuing and scheduling costs at t_{si} and t_{ei}. For arrivals closer to t^* a group i user's scheduling cost is lower, whereas the queuing cost is higher. The scheduling costs of i increase with the number of users with a smaller relative size of the value of time (μ_j), while queuing costs increase with the number of drivers with a larger μ_j. Both costs increase with i's value of schedule delay.

$$CSD_i[t_{si}] = CSD_i[t_{ei}] = \frac{\eta}{(1+\eta)} \frac{\beta_i}{s} \sum_{j=1}^{j=i} n_j \tag{3}$$

$$CT_i[t_{si}] = CT_i[t_{ei}] = \frac{\eta}{(1+\eta)} \frac{\beta_i}{s} \left(\mu_i \sum_{j=i+1}^{j=M} n_j / \mu_j \right) \tag{4}$$

$$P_i = CSD_i[t] + CT_i[t] = \frac{\eta}{(1+\eta)} \frac{\beta_i}{s} \left(\sum_{j=1}^{j=i} n_j + \mu_i \sum_{j=i+1}^{j=M} n_j / \mu_j \right) \tag{5}$$

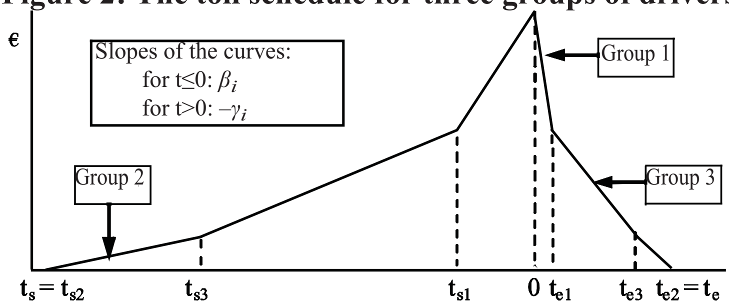
The NT price of i in (5) is the sum of i's queuing and scheduling costs. It increases with μ_i and β_i. Group M's price is $\eta \cdot N \cdot \beta_M/((1+\eta)s)$. This is the same price as in the homogeneous user model with a value of schedule delay of β_M. Group M−1 gains from the heterogeneity. Group M's higher relative size of value of time induces them to build up the queue slower than group M−1 drivers would. Hence, group M drivers impose lower congestion costs than group M−1 users. Group M−2 enjoys a larger price advantage, because group M and M−1 drivers build up the queue slower than group M−2 users.

3.2 Tolling equilibrium with discrete heterogeneity

With first-best tolling, the *time-variant* toll eliminates all queuing and the time-invariant toll is zero. As de Palma and Lindsey (2000) discuss, both a private and a public pay-lane's time-variant toll eliminates all queuing on the pay-lane, since queuing is always wasteful.

A queue eliminating toll changes the arrival ordering: groups now arrive ordered on their value of schedule delay. The *highest-β-users* are group 1 and arrive closest to t^*. The *lowest-β-users* are group K and arrive furthest from t^* (Arnott et al, 1992). Figure 2 gives an example FB toll schedule for the same set-up as Figure 1. Groups 2 and 3 have switched in the arrival order, since group 3 has a higher β_i. Group 1 has the highest β_i and lowest μ_i. Hence, Group 1 arrives closest to t^* with and without tolling. The queue eliminating toll makes all users indifferent, as long as there is no queue, between all arriving moments in the period that their group arrives in the new ordering. Outside that period, their price must be higher. The *time-variant* toll is by definition zero at t_s. Its slope is β_i before t^* and $-\gamma_i$ after.

Figure 2: The toll schedule for three groups of drivers.



Using the above discussion, equations (6) and (7) for i's schedule delay cost and toll can be derived. For type i arrivals closer to t^*, scheduling costs are lower, while the toll is higher. The sum of the scheduling cost and toll is constant during the entire period a group arrives. The resulting price formula is given in (8).

$$CSD_i[t_{si}] = CSD_i[t_{ei}] = \frac{\eta}{(1+\eta)} \frac{\beta_i}{s} \left(\sum_{k=i}^{k=K} n_k \right) \tag{6}$$

$$\tau[t_{si}] = \tau[t_{ei}] = \frac{\eta}{(1+\eta)\, s} \left(\sum_{k=1}^{k=i-1} \beta_k n_k \right) + \overline{\tau} \tag{7}$$

$$P_i = \tau[t] + CSD_i[t] = \frac{\eta}{(1+\eta)} \frac{1}{s} \left(\beta_i \sum_{k=i+1}^{k=K} n_k + \sum_{k=1}^{k=i-1} n_k \beta_k \right) + \overline{\tau} \tag{8}$$

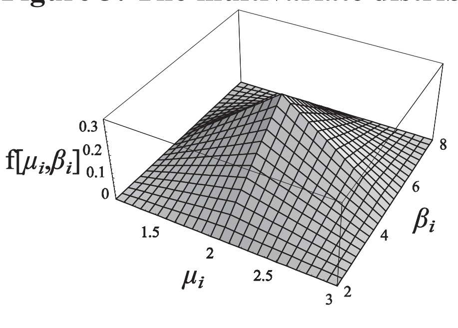
4. Numerical set-up

This section discusses the numerical set up of the base case. The bottleneck's capacity is 3600 cars per hour. The operating costs per trip are €7.3 and free-flow travel time is 30 minutes.

Figure 3 shows the NT density function of μ and β. It is based on two univariate symmetric triangular distributions: $f[\mu_i,\beta_i]=g[\beta_i]\cdot h[\mu_i]$. A symmetric triangular distribution is defined by its minimum and maximum. The minimum value per hour of schedule delay early ($\underline{\beta}$) is €2, the maximum ($\overline{\beta}$) is €8. The minimum relative size of the value of time ($\underline{\mu}$) is 1.01, the maximum ($\overline{u}$) is 3.01. The value of time is always larger than the value of schedule delay early. The mean value of time is €10.05. In this paper we only use weighted averages. The relative size (η) of γ_i to β_i is 3.9. This is the same value as in Arnott et al. (1992).

The inverse demand function is created, following equation (2), so that three goals hold. First, the total number of NT users is 9000. Second, the weighted average of the NT equilibrium elasticity to the total price is –0.4. Total price is the price including free-flow travel time and operating costs. Third, the discussed density function holds in the NT case[1].

Figure 3: The multivariate distribution



5. Continuous heterogeneity no-toll (NT) equilibrium

Now we analyse the continuous heterogeneity no-toll case. This paper first studies the analytical models and then illustrates these by the numerical results as calculated in Mathematica 5.0. There

[1] To achieve these goals we set $b_i[\mu_i,\beta_i]$ equal to the density function ($f[\mu_i,\beta_i]$), and A_i to the NT total price (P_{iNT}). The mean elasticity depends on B and the average total price. The average price is a function of the density function and the total number of users. Hence, we can calibrate the aggregate elasticity with B. If $b_i[\mu_i,\beta_i]=f[\mu_i,\beta_i]$ and $A_i=P_{iNT}$, then equating inverse demands to prices, and rewriting, results in $N_{NT}=A/B$. Hence, we set the aggregate number of users with A. The A and B we use equal 53.1841 and 0.0059094.

are no closed-form solutions for the tolling policies. Hence, in these cases we give the analytical results to the extent that they exist and then describe the numerical solution.

5.1 Analytical model for the no-toll (NT) equilibrium

The continuous heterogeneity price formula proofed a straightforward generalisation of the discrete version. The NT users arrive ordered on their relative size of the value of time (μ_i). The lowest-μ-users arrive closest to t^*. Schedule delays increase and queuing times decrease with the relative size of the value of time, whereas they are independent of the value of schedule delay. All NT users with the same μ_j behave in the same way. Thus, as (9) shows, we aggregate all users with the same μ_j to $m_{iNT}[\mu_i]$, by integrating over β_i the number of the NT drivers.

To get the continuous heterogeneity price equation (10) from the discrete version, we replace the summation signs by integrals and n_i by $m_{iNT}[\mu_i]$ (since there are now multiple types with the same μ_i). Equation (10) is rewritten to equation (11) by replacing $m_{iNT}[\mu_i]$ with $h[\mu_i]\cdot N_{NT}$ (i.e. μ's density function multiplied with the number of NT users). $H[\mu_i]$ is the cumulative distribution function of μ_i and N_{NT} is the number of NT users. We use a multivariate probability density function (PDF) of μ and β that is the product of two independent PDF's: $f[\mu_i,\beta_i]$ equals $h[\mu_i]g[\beta_i]$. The $g[\beta_i]$ and $h[\mu_i]$ are the univariate density functions of β and μ. Our model also works with a multivariate density function that is not based on two unconditional density functions. Then, however, our equations require more integrals and look more cluttered.

$$m_{iNT}[\mu_i] = \int_{\underline{\beta}}^{\bar{\beta}} n_{lNT}\, d\beta_l \tag{9}$$

$$P_{iNT} = CSD_i[t] + CT_i[t] = \frac{\eta}{1+\eta}\frac{\beta_i}{s}\left(\int_{\underline{\mu}}^{\mu_i} m_{jNT}[\mu_j]\, d\mu_j + \mu_i \int_{\mu_i}^{\bar{\mu}} \frac{m_{jNT}[\mu_j]}{\mu_j} d\mu_j\right) \tag{10}$$

$$P_{iNT} = \frac{\eta}{1+\eta}\frac{N_{NT}}{s}\beta_i\left(H[\mu_i] + \mu_i \int_{\mu_i}^{\bar{\mu}} \frac{h[\mu_j]}{\mu_j} d\mu_j\right) \tag{11}$$

Scheduling costs are given by the left term in the brackets in (11) multiplied by the term outside the brackets: by $\eta\cdot N_{NT}\cdot\beta_i\cdot H[\mu_i]/((1+\eta)s)$. The right term in the brackets multiplied gives the scheduling costs. Although schedule delays and queuing times are independent of β_i, prices increase linearly with β_i. A given delay is more valuable with a higher β_i. The value of time equals $\mu_i\cdot\beta_i$. Thus, a given queuing time is more valuable with a higher β_i. As prices depend linearly on β_i, ignoring heterogeneity in β should not bias the average price. In the linear in β case, $E[\beta\cdot k[\mu]]$ equals $E[\beta]\cdot E[k[\mu]]$; while the non-linear $E[k[\mu]/\beta]$ does not equal $E[k[\mu]]/E[\beta]$. Prices increase non-linearly with the relative size of the value of time (μ_i).

5.2 Congestion externalities and heterogeneity

Following Lindsey (2004), equation (12) gives the congestion cost effect of a type j driver on a type i: it is the derivative of i's price to the number of type j users. On all drivers with a larger μ_i than j (i.e. with $\mu_i \geq \mu_j$), type j causes a congestion effect of $(\beta_i/s)\cdot\eta/(\eta+1)$, which is independent of μ_i and μ_j. On all users with a smaller μ_i, j causes a smaller congestion effect. This smaller effect decreases with j's relative size of the value of time (μ_j) and increases with i's μ_i.

$$\partial P_{iNT} / \partial n_{jNT} = \begin{cases} \dfrac{\eta}{1+\eta}\dfrac{\beta_i}{s} & \mu_i \geq \mu_j \\[2ex] \dfrac{\eta}{1+\eta}\dfrac{\beta_i}{s}\dfrac{\mu_i}{\mu_j} & \mu_i < \mu_j \end{cases} \tag{12}$$

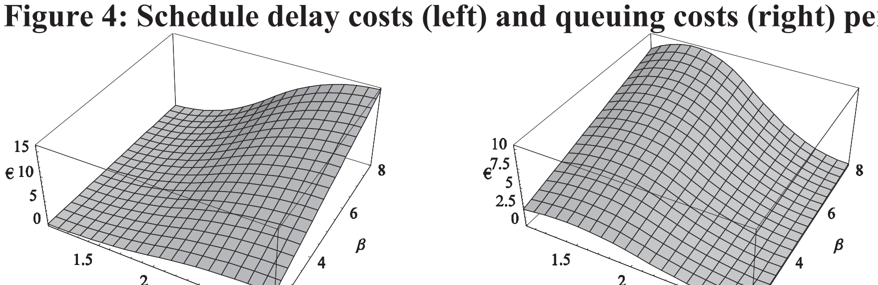
The congestion effect of j also increases with i's value of schedule delay, j's value of schedule delay has no effect. A higher μ_j means that j's isocost curve becomes flatter. This implies that j builds up the queue less quickly. This decreases the price for all users with a smaller μ_i than j, while prices for users with a larger μ_i than j are unaffected.

The marginal external cost of i in (13) is the integral of i's congestion effect on j multiplied by n_j over all values of μ_j and β_l. It is the sum of all congestion effects i imposes. The marginal external cost of i in (13) is the integral of i's congestion effect on j multiplied by n_j, it is the sum of all effects i imposes. $E[\beta]$ is the mean value of schedule delay. Only the mean value of schedule delay affects externalities. The heterogeneity middles out[2], as congestion effects depend linearly on β_i. The larger μ_i is, the smaller i's externality. The *lowest-μ-users* cause the highest externality. For any distribution with the same $E[\beta]$, the highest externality equals the externality of all users in the homogeneous user model of $E[\beta](\eta/(1+\eta))\cdot N_{NT}/s$. Consequently, the mean externality is lower with heterogeneity in μ than without. With more heterogeneity, there are higher *highest-μ-types*, who impose lower externalities than all other types, and/or more *high-μ-users*, who cause low externalities. The externalities of the low-μ-users cannot exceed $E[\beta](\eta/(1+\eta))\cdot N_{NT}/s$. Hence, with a more heterogeneous μ, the mean externality is lower. This result is also found by Van den Berg and Verhoef (2009a) and is hence not discussed further.

Since, the mean externality decreases with heterogeneity, the mean prices also decreases. Hence, if one just uses the mean value of time, the mean price is positively biased.

$$\begin{aligned}
\mathrm{mec}_i &= \int_{\underline{\beta}}^{\bar{\beta}}\int_{\underline{\mu}}^{\bar{\mu}} (\partial P_j / \partial n_{iNT}) n_{jNT}\; d\mu_j d\beta_l \\
&= \frac{\eta}{1+\eta}\frac{N_{NT}}{s}\left(\int_{\underline{\beta}}^{\bar{\beta}}\left(\int_{\mu_i}^{\bar{\mu}} h[\mu_j]\; d\mu_j + \frac{1}{\mu_i}\int_{\underline{\mu}}^{\mu_i}\mu_j h[\mu_j]\; d\mu_j\right)\beta_l g[\beta_l] d\beta_l\right) \\
&= \frac{\eta}{1+\eta}\frac{N_{NT}}{s}\left(1 - H[\mu_i] + \frac{1}{\mu_i}\int_{\underline{\mu}}^{\mu_i}\mu_j h[\mu_j]\; d\mu_j\right)E[\beta]
\end{aligned} \tag{13}$$

Figure 4: Schedule delay costs (left) and queuing costs (right) per user



5.3 Base case numerical model for the no-toll (NT) equilibrium

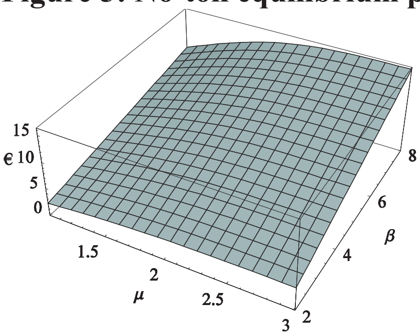
In the numerical NT base case, average queuing and scheduling costs are €3.97 and €4.97. The average total price (including free-flow travel time and operating costs) is €21.27. Total consumer surplus is €239,332. Average travel time is 54 minutes. Minimum travel time equals the free-flow travel time of 30 minutes, the maximum is 77 minutes.

[2] Comparison of line two and three of equation (9) also shows this. In line two we find the term $\beta_l \cdot g[\beta_l]$, which contains the heterogeneity in β. In line three, the term is integrated out to $E[\beta]$, and i's mec is a function of $E[\beta]$, μ_i and the distribution of μ.

Figure 4 shows that scheduling and queuing costs increase linearly with β. Scheduling costs increase non-linearly with μ. Queuing costs decrease with μ for all, but the *lowest-μ-users*. The *lowest-μ-drivers* face the longest queue. Yet, with their low values of time, these long delays are not that costly. Figure 5 depicts the NT prices. Prices increases linearly with the value of schedule delay; they increase non-linearly with μ_i.

Summarising this section, the average externality decreases with heterogeneity in the relative size of the value of time. Hence, the average NT price also decreases. Ignoring heterogeneity in μ, by just plugging in the mean parameter, biases the average price. Since prices depend linearly on β_i, ignoring heterogeneity in β should not bias the average price.

Figure 5: No-toll equilibrium price excluding free-flow travel time and operating costs



6. Continuous heterogeneity and first-best public (FB) tolling

The first-best public toll maximises welfare, which is the sum of total consumer surplus and toll revenues. The optimal time-*variant* toll eliminates all queuing. It slopes upward with t before t^* by β_i and downward by $-\gamma_i$ after t^*. The optimal time-*invariant* toll is zero. FB tolling changes the arrival ordering. Before the toll, users arrived ordered on μ_i. Now they arrive ordered on β_i. All high-β-users now arrive close to t^*, whereas before some had to arrive far from t^* (Arnott et al, 1992). Thus, not only does tolling eliminate the queuing, it also decreases scheduling costs.

6.1 Analytical model for the first-best public (FB) equilibrium

We derive the FB price equation (14) from the discrete version, by replacing the summation signs by integrals. $q_{jFB}[\beta_j]$ is the number of FB users with a value of schedule delay of β_j. It is the integral over μ_i of the number of users with β_j. In the FB equilibrium, all users with the same value of schedule behave in the same way regardless of their μ_i. The toll i faces increases with i's value of schedule delay; because the higher it is, the closer i arrives to t^* where the toll is the highest. The effect on scheduling costs of β_i is ambiguous: schedule delays decrease with β_i, whereas the value of a given delay increases with β_i. Nevertheless, FB prices increase with β_i. Different from in the no-toll equilibrium, prices are now independent of μ_i.

$$P_{iFB} = CSD_i[t] + \tau_t[t] = \frac{\eta}{(1+\eta)}\frac{1}{s}\left(\beta_i \int_{\beta_i}^{\bar{\beta}} q_{jFB}[\beta_j]\, d\beta_j \; + \int_{\underline{\beta}}^{\beta_i} \beta_j\, q_{jFB}[\beta_j]\, d\beta_j\right) \tag{14}$$

6.2 Numerical base case model for the first-best public (FB) equilibrium

There is no closed-form solution of the FB equilibrium. Still, for a given starting distribution of the FB users, there is a solution for prices, total number of users and implied distribution of the number of users. Unless, however, we exactly chose the equilibrium distribution as the starting distribution, the implied distribution and starting distribution are not equal. We can directly calculate NT prices, because we know the NT distribution of users.

Still, this discussion does suggest a simple solution method. Starting with some distribution of FB users (we use the NT user distribution). Using this distribution, we calculate prices and the distribution of demands implied by these prices. This distribution is not equal to the starting distribution, as demands and prices are not in equilibrium. By continuing this procedure until convergence, using the new distribution of demand as the starting distribution, we find the FB equilibrium[3]. The convergence criterion is a maximum absolute difference in the number of users of 10^{-20}% between iterations.

Figure 6 plots the FB schedule delay costs and tolls. The sum of the two (i.e. the price) concavely increases with β. Scheduling costs increase with β for all but the *lowest-β-users*. These lowest-β-drivers do face the largest delays. Yet, with their low values of schedule delay, costs are lower. Figure 7 shows difference between the NT and FB prices, the left panel depicts the change in 3D and the right panel in contourplot. It does this in value of schedule delay and value of time space; because the distributional effect are easier understood in the way than if the results are plotted in β and μ space. The darker the contourplot is, the lower the graph. The levels of the (white) contourlines are indicated in the figure.

Users with a high to intermediate β_i and μ_i gain from FB tolling. In the no-toll case, these users faced high queuing costs. In the FB case, these costs are replaced by substantially lower tolls. Drivers with a low β_i and/or μ_i face higher costs with FB tolling than without a toll.

Figure 6: Schedule delay costs and tolls per user

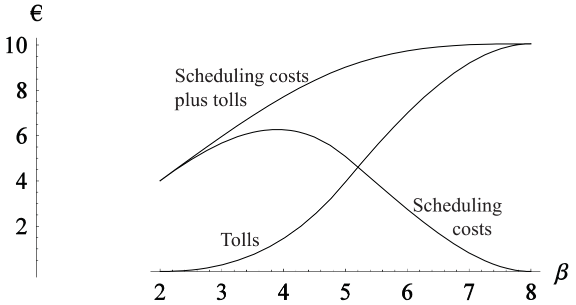
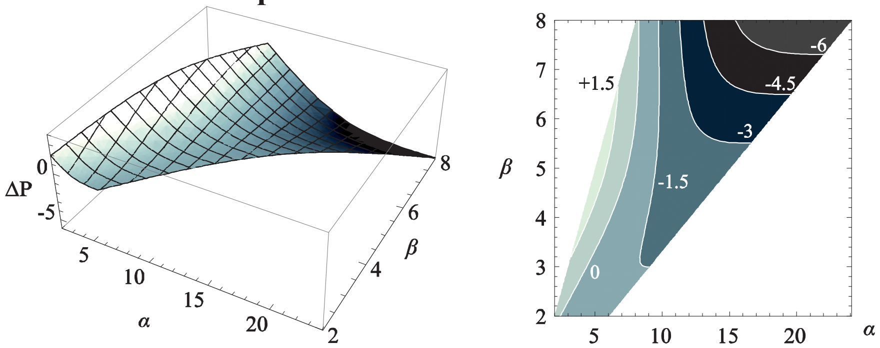


Figure 7: Change in price in 3D (left) and in contourplot (right) in value of schedule delay and value of time space



Note: the contourlines are, from left to top right, change in price is 1.5, 0, –1.5, –3, –4.5 and –6 euros. The conversion to β and α space is easily attained by defining $\mu_i=\alpha_i / \beta_i$.

[3] We approximate the distribution of demand between iterations by a cubic spline with 200 points. This spline is used as the next starting distribution. It is possible to use the demand as the starting distribution and not the spline. Then, however, the starting distribution's equation grows exponentionally complex with the number of iterations. This makes the integrations slow and often causes them to break down.

It is interesting that it is not the lowest value of time who lose most. The *intermediate type* users, with a value of schedule delay early of €5 and a slightly larger value of time, lose more. In the NT case, a lowest value of time and schedule delay user arrived on t^*. With her low value of time, the long queue she faced was not that costly. Her generalised price was the lowest of all users. Her FB price is also the lowest. Now, she faces a zero toll and the longest schedule delay of all FB users. This long delay is, with her low β_i, not that costly. Hence, her price only increases slightly. In the NT case, an *intermediate type* user arrived on t^*. With her higher value of time, this long queue was more costly than for the lowest value of time. With FB tolling, the *intermediate type* arrives halfway the peak. She faces the average toll, and her scheduling costs are slightly higher than those of a lowest value of time and schedule delay user. Thus, the *intermediate type*'s price increases substantially, and her price increase is the largest of all types.

Verhoef and Small (2004) and Van den Berg and Verhoef (2009) analyse continuous heterogeneity in the value of time respectively using static flow and bottleneck congestion. Verhoef and Small (2004) find that with FB pricing all users lose; but the larger the value of time is, the smaller the loss. Their public pay-lane has similar distributional effects as our FB toll: the intermediate values of time lose most, the lowest values of time less, and the highest values gain. In Van den Berg and Verhoef (2009) the highest values of time slightly gain and the lowest values lose most from a FB toll. This differs from our results, thus allowing for heterogeneity in the value of schedule delay has a large effect on the distributional impact of FB tolling.

In the base case, the average FB toll is €4.32. The number of users increases from the NT equilibrium with 0.6% to 9057. Consumer surplus goes up by 1.4% to €242,571. Welfare increases by 17.7% to €281,708. Average travel time decreases by 24 minutes. Average scheduling costs decrease with 13.1%, even though the number of users increases. Thus, FB tolling makes the arrival order substantially more efficient. The average total price (including free-flow travel time and operating costs) decreases with 1.2% to €21.01. Further, 55% of the NT users would now face lower prices. Hence, in our base case numerical model, FB tolling is good for the average consumer. If one only uses the mean value of schedule delay, one misses this point: for instance, with homogeneous users, tolling has no effect on prices.

7. Continuous heterogeneity and the pay-lane

With a pay-lane, a share (ρ) of capacity is made a separate lane, and to use this lane one has to pay a toll. The remainder of road is the untolled free-lane. We also refer to the pay-lane as lane 1 and the free-lane as lane 2. The pay-lane's *time-variant* toll eliminates all queuing on the pay-lane. The private operator adds a time-invariant toll that maximises total toll revenue. The public operator adds a negative time-invariant toll (i.e. subsidy) that maximises welfare.

The subsidy attracts extra drivers to the public pay-lane. Hence, the pay-lane peak last longer than on the free-lane, and schedule delays are higher on the pay-lane. The higher the subsidy is, the lower free-lane travel delays, but the higher total schedule delays. The optimal subsidy is at the point where, for a marginal subsidy increase, the welfare gain from lessening queuing equals the loss from higher schedule delays.

Because of the positive time-invariant PPL toll, the private pay-lane is used by relatively fewer drivers than the free-lane. Hence, the free-lane has the longer peak and schedule delays.

7.1 Analytical pay-lane model

The pay-lane's price equation (15) is basically the same as in the FB case. $q_{j1}[\beta_j]$ is the number of pay-lane users with a value of schedule delay of β_j. The free-lane price equation (16) is basically the same as the NT price formula. The $m_{j2}[\mu_j]$ is the number of free-lane users with μ_j. Finally, there is a critical $\mu^*[\beta_i]$ curve that separates the free-lane and pay-lane users. The users on the $\mu^*[\beta_i]$ curve are indifferent about using the pay-lane or free-lane. All the users that, for their β_i, have a higher μ_i than $\mu^*[\beta_i]$ drive on the pay-lane.

$$P_{i1} = CSD_i[t] + \tau_t[t] + \bar{\tau} = \frac{\eta}{(1+\eta)} \frac{1}{s\,\rho}\left(\int_{\underline{\beta}}^{\beta_i} \beta_j\, q_{j1}[\beta_j]\, d\beta_j + \beta_i \int_{\beta_i}^{\bar{\beta}} q_{j1}[\beta_j]\, d\beta_j \right) + \bar{\tau} \tag{15}$$

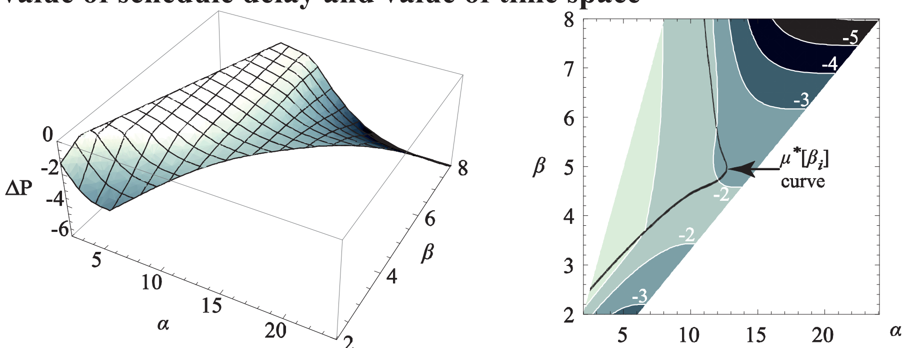
$$P_{i2} = CSD_i[t] + CT_i[t] = \frac{\eta}{1+\eta}\frac{\beta_i}{s(1-\rho)}\left(\int_{\underline{\mu}}^{\mu_i} m_{j2}[\mu_j]\, d\mu_j + \mu_i \int_{\mu_i}^{\bar{\mu}} \left(m_{i2}[\mu_j]/\mu_j\right) d\mu_j \right) \tag{16}$$

7.2 Base case numerical model for the public pay-lane (PL)

The pay-lane equilibria have no closed-form solutions. The numerical solution is more difficult than for the FB toll, because now there are also no closed form solutions for the $\mu^*[\beta_i]$ curve and optimal time-invariant toll ($\bar{\tau}$). The numerical solution method is discussed in the Appendix.

The optimal time-*invariant* toll is −€5.36. Hence, arrivals at the outside of the pay-lane peak receive a subsidy of €5.36. The mean time-*variant* toll is €6.25. Consumer surplus increases 7.0% from the NT case to €256,142. Welfare increases with 8.7% to €260,237. The number of users increases, with 3.4%, to 9309.8. Of these users, 4619.1 use the pay-lane and 4690.7 the free-lane. Thus, although the pay-lane has only a third of capacity, almost equal amounts of traffic use the pay-lane and free-lane. Hence, the pay-lane's mean schedule delay is higher.

Figure 8: Differences between PL and NT prices in 3D (left) and contourplot (right) in value of schedule delay and value of time space



Note: the critical $\mu^*[\beta_i]$ curve is given in the contourplot, all users to the right of the curve use the Pay-lane. The white contourlines are, from left to top right, change in price is: –1, –2, –3, –4, and –5 euros.

Figure 8 shows the difference between PL and NT prices. Again we plot the results is value of time and value of schedule delay space, as this makes it easier to interpret the results. In Figure 8's contourplot, the $\mu^*[\beta_i]$ curve separates the pay-lane and free-lane users. All users to the right of the curve use the pay-lane. Surprisingly, not only the high values of time and schedule delay use the pay-lane, but also the low values of time and schedule delay. Moreover, all *lowest-β-users* use on the pay-lane. This is counterintuitive: one would expect that only the highest values of time and schedule delay would use the tolled lane. The *low-β-users* arrive at the outside of the pay-lane peak. They face *negative* tolls and large schedule delays. With their low values of schedule delays, these large delays are not costly. Hence, they can enjoy the negative tolls and attain a large price decrease. Having the *low-β-users* driving on the pay-lane improves the PL's welfare gain, as the pay-lane's higher schedule delays are imposed on the *lowest-β-users*.

All users are better off under the PL than in the NT case. In the Van den Berg and Verhoef (2009) study, with only heterogeneity in the value of time, the 31% lowest value of time users lost due to the PL. Hence, the fact that, with heterogeneity in β, tolling lowers schedule delays makes consumers better off. The distributional effects of the public pay-lane are surprising.

Although, the users with a high μ_i and β_i gain most; the lowest values of time and schedule delay also gain. The users with an intermediate β_i and a low μ_i gain the least. In contrast, the conventional view is that a pay-lane is bad for the lowest values of time (before revenue recycling), since these users cannot afford the pay-lane and have to use the free-lane.

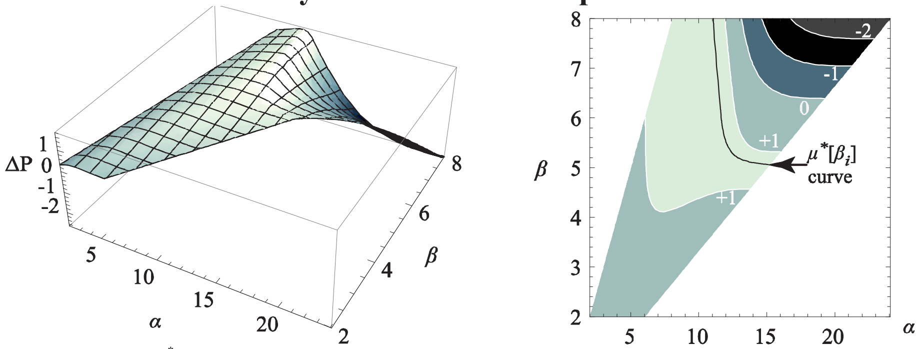
7.3 Base case numerical model for the private pay-lane (PPL)

Figure 9 depicts the differences between PPL and NT prices. Due to the positive time-invariant toll, the peak lasts shorter on the pay-lane than on the free-lane. The PPL's $\mu^*[\beta_i]$ curve, in Figure 9, has a more expected shape than the PL's: only the *high μ and β users* drive on the pay-lane. For the users close to the $\mu^*[\beta_i]$ curve prices increase most. Hence, there are pay-lane users that are hurt considerably by the PPL. For the free-lane users that are further from the $\mu^*[\beta_i]$ curve the PPL is less detrimental. On the pay-lane, as a type is further away from the $\mu^*[\beta_i]$ curve the PPL becomes rapidly more beneficial. Nevertheless, for most pay-lane users the PPL raises the price.

The PPL has some distributional surprises. It is not the *low values of time and schedule* delay that lose most, but the middle group (those who are almost indifferent about using the pay-lane or the free-lane). Further, this last group also contains pay-lane users. There are thus pay-lane users for who the PPL is more harmful than for the lowest-β-users. These results are similar to Verhoef and Small (2004)'s, where also the intermediate users, who are almost indifferent between using the pay-lane and free-lane, lose most, and the lowest value of time loses little.

The time-invariant PPL toll is €3.87 and the mean time-variant toll is €4.34. The number of users decreases with 1.7% to 8846.8, because the PPL raises the price for most users. Nevertheless, the PPL increases welfare by 4.1%. Its relative efficiency is 0.23.

Figure 9: Difference between PPL and NT prices in 3D (left) and contourplot (right) in value of schedule delay and value of time space



Note: the critical $\mu^*[\beta_i]$ curve is also given in the contourplot, all users above the curve travel on the Pay-lane. The values for the difference in price ($P_{iPL}-P_{iNT}$) of the contourlines are, from top to bottom, change in price is: −2, − 1, 0, and 1 euros.

7.4 Concluding the pay-lane models

A public pay-lane lowers prices of all users. Prices with a private pay-lane are higher than no-toll prices for all free-lane users and for some pay-lane drivers. The public pay-lane is not only used by the high values of time and schedule delay, but also by the low values of schedule delay. The *low-β-users* arrive on the outside of the peak. They face large schedule delays and negative tolls (i.e. subsidies). For these users the large schedule delays are not that costly, and thus they can enjoy the subsidies. The lowest values of time and schedule delay are among those who gain the most from the PL. Moreover, these users lose relatively little with the PPL.

8. Sensitivity analysis

This section focuses on the effect of different distribution of μ and β on the policies. We study five cases: *homogeneity*, the *base case*, *less heterogeneity in μ*, *less heterogeneity in β* and a *uniform distribution.* In all cases the mean value of schedule delay early is €5 and the mean relative size of the value of time is 2.01. The *base case'*s spread of the triangular distribution of μ is 2 and of β it is €6. With the *less heterogeneity in μ,* the spread of triangular distribution of μ is reduced to 1. In the case with *less heterogeneity in β,* the spread of β equals €2. With the uniform distribution, we test whether our results depend on the used triangular distribution. For comparability, the uniform distribution has the same variance as the *base case*. Finally, in the *homogeneity* case, all users have the same parameters[4].

8.1 Effect of heterogeneity on the no-toll (NT) case

Table 2 studies the effect of heterogeneity in μ and β on the NT equilibrium. The mean price is lower in the *base case* than in the *less heterogeneity in μ* case. This suggests that the mean price decreases with heterogeneity in the relative size of the value of time. The mean congestion externality decreases with heterogeneity in μ, thereby lowering queuing costs. As section 5 predicted, heterogeneity in β has no effect on average NT prices. The *base case* and *less heterogeneity in β* case have the same average externality and price. This signifies that this heterogeneity has no aggregate effect on the NT equilibrium.

By design, all five cases have the same NT consumer surplus. The advantage of this is that the effect of tolling is more comparable over cases than when surplus would differ across the cases.

Comparison of the *base case triangular* and *uniform* distribution indicates that the choice of NT user distribution has no significant effect on aggregate results. The average total price only differs a cent between the two cases, whereas mean queuing and scheduling costs do not even differ a cent. With both distributions, prices non-linearly increase with μ_i and linearly with β_i.

Table 2: Effect of heterogeneity in the no-toll (NT) equilibrium

	Homogeneity	Base case	Less heterogeneity in μ	Less heterogeneity in β	Uniform distribution[a]
Spread of the μ distribution	-	2	1	2	1.414
Spread of the β distribution	-	6	6	2	4.243
Mean schedule delay cost	€4.97	€4.97	€4.97	€4.97	€4.97
Mean travel delay cost	€4.97	€3.97	€4.44	€3.97	€3.97
Mean total price	**€22.27**	**€21.27**	**€21.74**	**€21.27**	**€21.26**
Mean marginal external cost	€9.95	€8.95	€9.42	€8.95	€8.94
Total NT consumer surplus	€239,332	€239,332	€239,332	€239,332	€239,332

a: This uniform distribution has the same variance and mean as the base case triangular distribution of NT users.

[4] We do not present sensitivity analyses on the price elasticity or pay-lane's capacity share. Van den Berg and Verhoef (2009) discuss these analyses. Their results are in line with the homogeneous user literature. Also in this research, these sensitivity analyses give unsurprising results.

8.2 Effect of heterogeneity on first-best public (FB) tolling

Table 3 shows the results of the sensitivity analysis for FB tolling. Tolling is more beneficial for consumers in the *base case* than with *homogeneity*. The price decreases due to FB tolling in the *base case*; whereas under *homogeneity*, prices are unaffected. Still, the welfare gain is higher under homogeneity, since toll revenues are considerably higher.

With more heterogeneity in μ, externalities are lower, and thus there is less to gain from tolling. Therefore, consumer surplus, toll revenues and welfare are higher with a less heterogeneous μ. In the *less heterogeneity in μ* case, the FB welfare gain is 19.5% of NT welfare, whereas in the *base case* it is 17.7%, and with *homogeneity* 18.7%.

With FB tolling, scheduling costs are lower in *base case* than with the *less heterogeneity in β,* since the gain of a more efficient arrival ordering increases with heterogeneity in β. FB welfare gain is higher in the *base case* than in the *less heterogeneity in β* case. This suggests that the FB welfare gain increases with heterogeneity in β. Conversely, this gain decreases with the heterogeneity in μ. If there is not too much more heterogeneity in the value of time than in the value of schedule delay, then FB tolling can be good for the average consumer[5]. For example, in the *base case*, 55% of the NT users would gain. With homogeneity, FB tolling has no effect on prices. In the *less heterogeneity in μ* case, 66% of NT users would gain from FB tolling, and with the *less heterogeneity in β* case 39%. The share of NT users that would gain seems to increase with heterogeneity in β and decrease with heterogeneity in μ.

FB welfare in the *base case* is lower than with homogenous users. For welfare to be higher with heterogeneity than with homogeneity, the value of schedule delay must be more heterogeneous (for a given NT μ distribution) than the *base case*. An example of such a situation is the *less heterogeneity in μ* case, where FB welfare is higher than with *homogeneity*.

The differences between the results for the *uniform* and *base case* distribution are small. The mean total price differs by a cent. The discrepancy between the total numbers of users is 0.6 user. The percentage welfare gain differs by 0.003 percentage points. A difference is that in the *uniform* case 53% of the NT users would gain, while in the *base case* 55%. These results seem to justify the view that the used distribution form has not an important effect on aggregate results.

In our model, there are always users that lose due to the FB toll. Conversely, in Vickrey (1973), FB tolling is Pareto improving. The difference lies in: (1) we also have heterogeneity in the relative size of the value of time, and this makes tolling more harmful for the average consumer. And (2) we use elastic demand, which means that the types that gain increase their demand, which increases the price for all types.

	Homogeneity	Base case	Less heterogeneity in μ	Less heterogeneity in β	Uniform distribution
Spread of the μ distribution	-	2	1	2	1.414
Spread of the β distribution	-	6	6	2	4.243
Mean schedule delay cost	€4.97	€ 4.32	€ 4.36	€4.70	€4.32
Mean FB toll	€4.97	€ 4.32	€ 4.36	€4.70	€4.32
Mean total price	€22.27	€21.01	€ 21.07	€21.75	€21.00
Number of FB users	9000	9054.6	9122.3	8922.8	9055.3
Toll revenues	€44,770	€39,137	€39,746	€41,970	€39,094
Total FB consumer surplus	€239,332	€242,571	€246,180	€235,352	€242,606

[5] The value of time always has at least as much heterogeneity as the value of schedule delay, because μ_i must be larger than one. If μ is homogeneous and the β heterogeneous, then the value of time and schedule delay are perfectly correlated and are equally heterogeneous.

Welfare under the FB	€284,102	€281,708	€285,926	€277,323	€281,701
Percentage welfare gain from the NT case	**18.7%**	**17.7%**	**19.5%**	**15.9%**	**17.7%**
Percentage NT users that would have a lower price with FB tolling	Price is unchanged	55%	66%	39%	53%

Table 3: Effect of heterogeneity on the first-best public (FB) toll

8.3Effect of heterogeneity on the public (PL) pay-lane

Table 4 gives the sensitivity analysis for the PL. In all five cases, all users gain due to the PL's time-invariant subsidy. In Van den Berg and Verhoef (2009)'s model (which is a constrained version of our model with no heterogeneity in β) there are users that lose from the PPL. This indicates that if the value of schedule delay is almost homogeneous, there will be some users that are disadvantaged by the PL.

The *less heterogeneity in* μ case has a higher welfare gain and relative efficiency than the *base case*, indicating that these measures decrease with the heterogeneity in μ. In the *base case*, the time-invariant toll is higher than in the *less heterogeneity in* μ case. Thus, in the *base case*, the pay-lane has relatively fewer users compared with the free-lane. Hence, free-lane queuing is worse in the *base case*. As Van den Berg and Verhoef (2009) discuss, it is apparently social optimal to allow more wasteful queuing with more heterogeneity, because of the lower congestion externalities, and this decreases the PL's relative efficiency.

The more heterogeneous β is, the more efficient tolling makes the arrival ordering, thereby increases PL's welfare gain. Surprisingly, the PL's relative efficiency also increases with the heterogeneity of β. One would expect that this relative efficiency would decrease with heterogeneity, as the pay-lane imposes the more efficient arrival order on only a part of the users. The public pay-lane has a larger maximum and mean schedule delay than the free-lane. These large delays are faced by the *lowest-β-users*. The *high-β-users* have the lowest delays. With more heterogeneity, the mean value of schedule delay of the pay-lane's *high-β-users* is higher, making the PL's schedule delay savings more valuable. The mean value of schedule delay of the pay-lane's *low-β-users* decreases with heterogeneity in β, making the PL's extra schedule delays less costly. This explains why the PL relative efficiency increases with the heterogeneity in β.

The differences between the *base case* and *uniform* distribution are larger than in the NT and FB cases. Nevertheless, the differences are still minor. The time-invariant toll is a cent lower with the *uniform* distribution; the PL's percentage welfare gain is 0.22 percentage point lower.

	Homogeneity	Base case	Less heterogeneity in μ	Less heterogeneity in β	Uniform distribution[a]
Spread of the μ distribution	-	2	1	2	1.414
Spread of the β distribution	-	6	6	2	4.243
Time-invariant part of the toll	−€6.37	−€5.36	−€5.40	−€5.38	−€5.35
Mean time-variant part of the toll	€7.27	€6.25	€6.29	€6.72	€6.31
Number of users	9302.8	9309.8	9356.4	9205.5	9293.9
Toll revenues	€3917	€4095	€4153	€5680	€4439
Total PL consumer surplus	€255,706	€256,142	€258,692	€250,408	€255,260
Welfare under the PL	€259,623	€260,237	€262,845	€256,089	€259,699
Relative efficiency	**0.453**	**0.493**	**0.505**	**0.441**	**0.481**
Percentage welfare gain from the NT case	**8.48%**	**8.73%**	**9.82%**	**7.00%**	**8.51%**
Percentage NT users that would					

have a lower price with the PL	100%	100%	100%	100%	100%

Table 4: Effect of heterogeneity with the public (PL) pay-lane
a: This uniform distribution has the same variance and mean as the base case triangular distribution of NT users.

8.4 Effect of heterogeneity on the private (PPL) pay-lane

Table 5 shows the final sensitivity analysis for the PPL. It indicates that the PPL welfare gain and relative efficiency decrease with heterogeneity in μ; as in the *less heterogeneity in* μ case these measures are higher than in the *base case*. Van den Berg and Verhoef (2009) suggest that with a more heterogeneous value of time, it is profit maximizing to allow more wasteful queuing and schedule delay on the free-lane, lowering the PPL's relative efficiency and welfare gain.

The PPL's welfare gain and relative efficiency are lower in the *less heterogeneity in* β case than in the *base case*. This indicates that these measures increase with heterogeneity in β. With a more heterogeneous β, the average value of schedule delay on the pay-lane is higher, and on the free-lane it is lower. This makes the schedule delay savings on the pay-lane more valuable and the free-lane's extra schedule delays less costly. Thus, the PPL is more beneficial with a more heterogeneous β. This reasoning follows the same train of thought as Verhoef and Small (2004) used to explain why, with static flow congestion, the pay-lane's relative efficiency increases with heterogeneity in the value of time.

With homogeneity, all users are worse off with a PPL than in the NT case. Conversely, in all heterogeneity cases there are some users who gain. The share of NT users that would face lower prices with the PPL decreases with heterogeneity in β. In opposition to FB tolling, the PL's share that would gain slightly decreases with heterogeneity in μ. This seems to be because with more heterogeneity in the value of time, the free-lane congestion externalities are worse. This makes the free-lane a less attractive good, enabling the PPL operator to ask a higher profit maximizing time-invariant toll. This is also why relatively more users use the pay-lane in the *less heterogeneity in* μ case than in the *base case*, even though the time-invariant toll is higher in the *less heterogeneity in* μ case. Hence, with welfare maximization, the share of users that gains from tolling increases with heterogeneity in μ; with profit maximization it decreases.

In all heterogeneity cases, the lowest values of time and schedule delay lose because of the PPL. Yet, they lose less than the users that are (almost) indifferent between the free-lane and pay-lane. For the lowest values of time and schedule delay, the price heightening decreases with the heterogeneity in β and raises with the heterogeneity in μ.

The differences between the base case triangular and uniform distribution are larger than for the NT or FB equilibrium. Again, however the differences are not over all result changing.

	Homogeneity	Base case	Less heterogeneity in μ	Less heterogeneity in β	Uniform distribution[a]
Spread of the μ distribution	-	2	1	2	1.414
Spread of the β distribution	-	6	6	2	4.243
Time-invariant part of the toll	€3.03	€3.87	€3.97	€3.14	€3.96
Mean time-variant part of the toll	€4.64	€4.34	€4.48	€3.97	€4.37
Number of users	8855.9	8846.8	8855.4	8836.1	8837.4
Toll revenues	€16,201	€17,964	€18,619	€16,288	€18,133
Total PPL consumer surplus	€231,729	€231,279	€231,724	€230,701	€230,790
Welfare under the PPL	€247,930	€249,243	€250,343	€246,989	€248,923
Relative efficiency	**0.192**	**0.234**	**0.236**	**0.202**	**0.226**
Percentage welfare gain from					

the NT case	**3.59%**	**4.14%**	**4.60%**	**3.19%**	**4.00%**
Percentage NT users that would have a lower price with the PPL	none	6.0%	5.7%	0.6%	5.0%

Table 5: Effect of heterogeneity with the private (PPL) pay-lane

a: This uniform distribution has the same variance and mean as the base case triangular distribution of NT users.

8.5 Concluding the sensitivity analysis

NT queuing costs are lower with a more heterogeneous relative size of the value of time, because congestion externalities are lower. Hence, there is less to gain from tolling, and the FB welfare gain is lower. Heterogeneity in β does not affect NT prices. Nevertheless, the welfare gain of FB tolling increases with heterogeneity in β; because, with heterogeneity, tolling makes the arrival ordering more efficient. With increased heterogeneity in β, the relative efficiency of the pay-lane is higher, while it is lower with a more heterogeneous relative size of the value of time.

The relative efficiencies of the pay-lanes are higher in the heterogeneous user base case than with homogeneous users. If the base case had a more heterogeneous μ or a less heterogeneous β, then the relative efficiencies could be higher with homogeneity than with heterogeneity. The FB welfare gain is lower with the base case than with homogeneity. Nevertheless, also in this case, if the calibration of the model would be different (e.g. a more heterogeneous β), this result might change. Whether a policy is more or less beneficial with heterogeneity than with homogeneity depends on the empirical question what distribution heterogeneity has: how much more heterogeneous is the value of time than the value of schedule delay?

9. Conclusion

This paper analysed how, in the bottleneck model, heterogeneity affects the effects of congestion tolling. The heterogeneity is in the value of schedule delay early (β) and relative size of the value of time to the value of schedule delay (μ). The value of time of a type *i* user is $\mu_i \cdot \beta_i$.

We focused on time-variant tolling. In reality, tolls are uniform over the peak or are step tolls. For future research, it seems attractive extend the research on how heterogeneity affects these less flexible tolls. One might also look at what effects other types of heterogeneity (e.g. in the value of uncertainty or between the value of schedule delay early and late) have on second-best tolling.

A more heterogeneous relative size of the value of time lowers the average no-toll equilibrium congestion externality. Hence, there is less to gain from tolling, and the welfare gains of tolling are lower. With a more heterogeneous value of schedule delay, the welfare gain of tolling is higher. With more heterogeneity, the gain from the more efficient arrival ordering tolling causes is higher, meaning that tolling lowers mean schedule delays costs more.

With a pay-lane, only a part of the road is tolled, the remainder (the free-lane) remains toll free. The pay-lane toll consists of a time-invariant and a time-variant part. The public pay-lane has a negative time-invariant toll (i.e. subsidy) that maximises welfare. The public pay-lane is used by highest values of time and schedule delay *and* by all the lowest values of schedule delay early (β). These *low-β-users* arrive on the outside of the pay-lane peak. They enjoy the negative tolls at the outside of the peak, whereas the large schedule delays then are not that costly. Hence, the negative toll at the outside of the peak causes the surprising result that the low values of schedule delay use the pay-lane. The private pay-lane has a positive time-invariant toll that maximises revenue. It is used by the highest values of time and schedule delay. The relative efficiency of a pay-lane increases with heterogeneity in the value of schedule delay, it decreases with heterogeneity in the relative size of the value of time.

If there is not too much more heterogeneity in the value of time than in the value of schedule delay, then the mean travel price can be lower with first-best public tolling than without tolling. The share of NT users that would face lower prices with first-best tolling increases with the heterogeneity in the value of schedule delay, while it decreases with the heterogeneity in the relative size of the value of time. For the private pay-lane, the share that would gain reduces with both the heterogeneity in the value of schedule delay and time. The first-best toll and private pay-lane are never a Pareto improvement. For most distributions of the heterogeneity, the public pay-lane is a Pareto improvement. Only if the value of schedule delay is (almost) homogeneous are there users who lose.

The distributional effects of tolling can be surprising. First-best tolling is most harmful for users with an average value of schedule delay and a slightly larger value of time, while it raises the price less for the lowest values of time and schedule delay. The lowest values of time and schedule delay are among those who gain most from a public pay-lane. Further, these users lose relatively little with the private pay-lane, while higher value of time and schedule delay free-lane users and even some pay-lane users lose more. Hence, in our model, tolling is not most harmful for the lowest values of time and schedule delay.

Acknowledgements
This research was financially supported by Transumo. Transumo (TRANsition SUstainable MObility) is a Dutch platform for companies, governments and knowledge institutes that cooperate in the development of knowledge with regard to sustainable mobility. The authors are grateful for the comments of Eva Gutiérrez-i-Puigarnau.

Literature

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

Arnott, R., de Palma, A. and Lindsey, R. (1991). A temporal and spatial equilibrium analysis of commuter parking. Journal of public economics 45, 301-335.

Arnott, R., de Palma, A. and Lindsey, R. (1992). Route choice with heterogeneous drivers and group-specific congestion costs. *Regional Science and Urban Economics* 22, 71–102.

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

de Palma, A. and Lindsey, R. (2000) Private toll roads: Competition under various ownership regimes. The Annals of Regional Science 34(1), 13-35.

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.

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

van den Berg, V. and Verhoef, E.T. (2009). Second-best congestion tolling in the bottleneck model with continuous heterogeneity in value of time. VU University working paper; available on request.

Verhoef, E.T. and Small, K.A. (2004). Product differentiation on roads. *Journal of Transport Economics and Policy* 38, 127-156.

Appendix: Numerical solution method for a pay-lane equilibrium

This appendix discusses the numerical solution procedure for the pay-lanes. Three iterative procedures are used: the second is around the first and the third procedure around the second. We begin with some starting time-invariant toll, critical $\mu^*[\beta_i]^0$ curve and distribution of users. The first iterative procedure searches for the distributions for which the prices and inverse demands are equal. The starting distribution is used to calculate prices and the demands implied by these prices. The next iteration uses this iteration's demand, approximated by a cubic spline, as the starting distribution. The convergence criterion is a maximum absolute percentage change in the number of pay-lane and free-lane drivers of 10^{-12}% from one iteration to the next.

For now, however, prices on the pay-lane and free-lane for $\mu^*[\beta_i]$ curve users are not equal. Hence, a new curve is sought for which these price differences are smaller. This switches some of pay-lane drivers that face lower prices on the free-lane to the free-lane; and vice versa for the free-lane. After this, the first iterative procedure is repeated again to find the new equilibrium user distribution. The convergence criteria for the second procedure is a maximum absolute percentage difference between pay-lane and free-lane prices for $\mu^*[\beta_i]$ curve users of 0.075%.

The third procedure searches for the optimal time-invariant toll. For the public pay-lane, we start with calculating welfare for three time-invariant tolls (−€1.50, −€2.00 and −€2.50). Next, we fit a second-order polynomial to the tolls and their welfare. This polynomial is maximised to find the predicted toll ($\overline{\tau}^p$). Next iteration's toll is $\overline{\tau}^{t+1} = (1 - ss) \cdot \overline{\tau}^t + ss \cdot \overline{\tau}^p$, where ss is the step size. Then, welfare resulting from this new toll is calculated. We fit a polynomial on the last two tolls (i.e. −€2.00 and −€2.50) and this iteration's toll and the welfares corresponding to these tolls, to find the next iteration's toll. If there is no increase in welfare from one iteration to the next, the step size is halved. Both consumer surplus and toll revenue seem globally concave in the time-invariant toll. Therefore, this procedure should converge to the welfare maximizing toll.

We repeat this third procedure until the absolute in welfare between iterations is below 0.25, on a PL welfare of around €260,000. After convergence, we use the time-invariant toll with the highest welfare in the series of calculations as the optimal toll. The procedure for the PPL is basically the same as for the PL, only now toll revenues are maximised instead of welfare.