Comparing the Efficiency of Alternative Policies for Reducing Traffic Congestion

Ian W.H. Parry

June 2000 • Discussion Paper 00–28

Resources for the Future
1616 P Street, NW
Washington, D.C. 20036
Telephone: 202–328–5000
Fax: 202–939–3460
Internet: http://www.rff.org

Comparing the Efficiency of Alternative Policies for Reducing Traffic Congestion

Ian W.H. Parry

Abstract

This paper compares the efficiency of a single lane toll, a congestion tax applied uniformly across freeway lanes, a gasoline tax, and a transit fare subsidy at reducing traffic congestion. The model incorporates a variety of conditions required to reach an efficient outcome. These include conditions for the efficient allocation of travel among competing modes, travel at peak versus off-peak periods, and drivers with high and low time costs sorted onto faster and slower freeway lanes.

Each policy violates some or all of the efficiency conditions. Under wide parameter scenarios, the single lane toll, gasoline tax, and transit subsidy forgo at least two thirds of the efficiency gains under an "ideal" congestion tax that varies across lanes. In contrast, the uniform congestion tax can achieve more than 90 percent of the efficiency gains, despite failing to separate out drivers with high and low time costs onto different freeway lanes.

Key Words: externalities, efficiency effects, congestion taxes, single lane tolls, rail subsidies, gasoline taxes

JEL Classification Numbers: R41, R48, D62

Contents

Comparing the Efficiency of Alternative Policies for Reducing Traffic Congestion

Ian W.H. Parry*

1. Introduction

Increasing gridlock on roads in metropolitan areas has heightened interest in policies to reduce traffic congestion.[1] The traditional response to congestion was to invest in more road capacity but, despite considerable investments, highway capacity has not kept pace with the growth in vehicle miles traveled. As a result congestion has grown steadily worse.[2] New roads tend to fill up with "latent" traffic demand until the (combined monetary and time) costs of an extra trip equals that on existing, competing routes (Downs, 1992, Arnott and Small, 1994). Similarly, expanding public transportation infrastructure is not enough: although new rail capacity may divert some travel off congested roads in the short term, sooner or later the roads tend to fill up with traffic again. Thus it is widely recognized that short run measures to combat congestion are required, in addition to long run capacity investments.

Economists have long advocated time of day pricing of congested roads (e.g., Pigou, 1912, Walters, 1961, and Vickrey, 1963). This policy offers the most cost-effective means of reducing congestion, in the sense that it induces people to efficiently exploit alternatives to peak-hour driving, such as using mass transit, driving on other (non-congested) routes, re-scheduling trips at less busy times, car pooling, and so on. Congestion fees have been slow to catch on, however, probably because of opposition from motorists who have previously had the "right" to travel freely on roads (e.g., Giuliano, 1992). Moreover the efficient fee structure across freeway

* I am grateful to Robin Boadway, Ken Small, and Mike Toman for very helpful comments and suggestions.

[1] In metropolitan Los Angeles and New York City alone, 1.3 billion hours are lost each year because of traffic delays (see *http://mobility.tamu.edu*). Arnott and Small (1994) estimated that the annual costs of time delays from road congestion (for 39 large metropolitan areas in the U.S.) amount to $48 billion per year, or $640 per driver. The total costs of congestion are probably larger than this estimate, since it excludes the extra gasoline consumption, and the inconvenience of people rescheduling or chaining trips to avoid congestion. Note, however, that because completely eliminating congestion is not a practical option, policies for congestion relief cannot be expected to yield benefits of this magnitude.

[2] Vehicle miles traveled in urban areas increased by 289 percent between 1960 and 1991, while total road capacity in urban areas increased by only 75 percent (Department of Transportation, 1993, pp. 18-20).

lanes is complicated by differences in drivers' valuation of time—it is optimal to separate out drivers with high time costs into faster lanes with higher charges and drivers with low time costs into slower lanes with lower charges (Small and Yan, 1999).

Recently in the U.S. there have been some experiments with individual lane tolls that vary with the time of day.[3] There is less opposition from the public to these schemes since they leave open the option of using other lanes on the freeway without having to pay any money. However, the drawback of single lane tolls is that they do not address congestion on unpriced lanes—in fact they exacerbate the problem since drivers substitute away from the priced lane.

Other people continue to advocate higher gasoline taxes or subsidizing transit fares as a short-run measure to mitigate congestion, though it is recognized that these pricing schemes are less efficient than a congestion fee. Gasoline taxes raise the cost of all driving and therefore do not induce the efficient substitution off congested roads onto other transport modes, or travel at off-peak periods. Similarly, transit fare subsidies only affect the price of driving on congested roads relative to public transit, but not the price relative to driving on other roads or at off-peak hours. [4]

To guide policy, it is important to understand the relative economic efficiency of the various short-run policy options for discouraging traffic congestion. For example, the economic case for using single lane tolls to combat congestion rather than higher gasoline taxes and larger transit fare subsidies, or for expanding existing individual lane tolls to cover more lanes, obviously hinges on whether there are large efficiency gains at stake. If not, policy makers may as well stick with the most politically palatable policies.

There have been some recent studies that compare single lane tolls with efficient congestion pricing (e.g., Braid, 1996, Liu and McDonald, 1998, Verhoef and Small, 1999). For example, Liu and McDonald (1998) estimate that single lane tolls on Route 91 in Orange County, California could only generate 10 percent of the economic gains from the most efficient

[3] On route 91 in Orange County, California, two new toll lanes were built in the median of an eight-lane freeway, and the tolls vary according to the time of day. Similar schemes now operate on I-15 in San Diego, and on I-10 in Houston. A variety of other proposals for single lane tolls are currently being studied in a number of different states (e.g., Poole and Orski, 1999, and Small and Gómez-Ibáñez, 1998). Some of these schemes are called high occupancy/tax (HOT) lanes, because single occupancy vehicles pay the toll, while multiple occupancy vehicles pay lower rates or no toll at all.

[4] Another policy that has gained popularity in recent years in the U.S. is high-occupancy vehicle (HOV) lanes. We do not analyze this policy however, as this would require a more complex model with household car pooling decisions.

congestion pricing. More recently, Small and Yan (1999) have found that the efficiency gains from single lane tolls could be three times as large when driver heterogeneity and the potential to sort out drivers with high and low time costs onto fast and slow lanes is taken into account. This result underscores the potential importance of capturing driver heterogeneity in comparative policy analysis.

The purpose of this paper is to compare the economic efficiency of short-run policy measures to reduce congestion, using a unifying analytical and numerical framework. We analyze a congestion tax imposed uniformly across all freeway lanes, a single lane toll, a gasoline tax, and a rail fare subsidy. For each policy, we calculate the efficiency potential expressed relative to that from "ideal" congestion pricing. We employ a generic model rather than one calibrated to a region-specific transport network, and this allows us to illustrate how the relative efficiency effects of policies change as we vary key parameters over a wide range of plausible values.[5]

Our model incorporates three main conditions that would be required in order to induce the most efficient outcome. These relate to the optimum allocation of travel among competing modes at a point in time, and the allocation of freeway travel across peak- and off-peak periods. In addition, efficiency requires separating out freeway drivers with different time costs onto different lanes.

The next section describes our basic model. In Section 3 we solve the model analytically to clarify, qualitatively, the conditions for an efficient outcome and we analyze to what extent these conditions can or cannot be met under alternative policies. For example, both the gasoline tax and transit subsidy fail to satisfy any of the efficiency conditions, while a uniform congestion tax only fails to separate drivers into fast and slow lanes.

In Section 4 we present extensive simulation results that quantify the relative efficiency potential of different policies. The main theme in these results is that—even though the uniform

[5] Other studies tend to look at these policies in isolation (e.g., Dodgson and Topham, 1987, and Glaister and Lewis, 1978, on public transit subsidies). Alternatively, comparisons of a broad range of policy instruments have been conducted with much more complicated models. For example, De Borger et al. (1997) study congestion taxes, transit subsidies and fuel taxes using a more detailed model that captures a variety of other (non-congestion) externalities and policy interventions within the transport system, and welfare effects from changes in tax revenues. Thus, their results are not directly comparable with those reported below (though they are broadly consistent). In short, the relative efficiency of the above four instruments at reducing congestion over a wide range of parameter scenarios has not really been spelled out in the literature. For a good qualitative analysis of alternative policy instruments see Strotz (1965).

congestion tax fails to create fast and slow freeway lanes—it generates efficiency gains that are usually at least three times as large as the efficiency gains under any other policy. The uniform congestion tax typically generates over 90 percent of the maximum efficiency gains under ideal congestion pricing,[6] while the other policies capture at best one third of the maximum gains, and often much less. In other words, in terms of economic efficiency, the most important consideration in instrument choice is to pick a policy that generates the efficient substitution off the congested freeway onto other modes and travel at off-peak hours—creating fast and slow lanes on the highway is less of a concern. These results are robust to alternative assumptions about the costs of congestion, the freeway demand elasticity, the variance in time costs across drivers, and the initial share of travel on different modes.

The final section of the paper discusses some important caveats to the results. We emphasize that certain policies may be justified on other grounds. For example, gasoline taxes can address pollution externalities, and transit fare subsidies are appropriate when the marginal cost of transit service is declining. Our analysis abstracts from these considerations, however, because we focus only on the case for policies based on their impact on reducing congestion. Furthermore, we ignore a number of second-best interactions arising from pre-existing "distortions" within the transport system, and interactions with the tax system, which may significantly affect the relative (second-best) efficiency impacts of alternative policies. In this respect, our analysis is only a building block to a broader policy comparison.

2. Model Assumptions

We use a fairly simplified model in order to focus attention on the ability of policies to induce the efficient allocation of travel across various modes. Consider a model where agents make a number of travel trips from the suburbs to the city center over a given period. Agents can travel in a variety of ways. First, they can drive along a freeway which consists of two lanes, *X* and *Y*. Second, they can use an alternative "backroads" route, which involves, for example, driving along neighborhood and city streets. Third, agents can travel using rail transit. Each of these three options involves travel during peak periods. The final option is to drive at off-peak hours along the freeway. For simplicity, we exclude travel by bus, since this would require modeling its impact on road congestion (we comment on the implications of bus travel below).

[6] This result is consistent with earlier work by Verhoef and Small (1999). They develop a numerical model of freeway congestion, using a distribution of time costs for drivers in the Dutch Randstad area. They find that a toll imposed uniformly across freeway lanes can achieve 92 percent of the efficiency gains from a differentiated toll.

For interpretive purposes we assume that agents gain direct utility from "travel services", hence travel is like a consumption good. There are two types of agents: those with high opportunity costs of time and those with low opportunity costs of time, denoted $i = H$ and L respectively. The number of each type of agent is "large" and is denoted by s^H and s^L. The number of trips per agent over the period is denoted as follows: T_X^i and T_Y^i (peak freeway trips on lanes X and Y); T_B^i (backroads trips); T_R^i (rail trips) and T_{OP}^i (off-peak freeway trips).[7]

There is congestion on the freeway at peak travel time. This means that the presence of an extra vehicle slows down the average speed of other drivers, hence raising their travel times. Agents do not take account of this external cost when deciding whether to use the freeway or not, hence in the absence of policy intervention there is a socially excessive amount of peak freeway traffic. There is no congestion on the backroads, rail, or the freeway at off-peak hours.[8]

Rail, backroads, and off-peak trips require a fixed amount of time, ϕ_j, and a fixed amount of money expenditure (gasoline costs, vehicle wear and tear, rail fares, etc.), θ_j. Total time and money expenses per mode for agent i are therefore

(2.1a) $\phi_j T_j^i, \theta_j T_j^i$ for $j = R, B, OP$

For peak freeway trips z_X and z_Y denote the proportionate increase in travel time due to traffic congestion on lanes X and Y. Total travel time and money expenses for these trips are:

(2.1b) $(1+z_k)\phi_F T_k^i, \theta_F T_k^i$ for $k = X, Y$

Congestion is determined as follows

(2.2) $z_k = z_k(s^H T_k^H + s^L T_k^L)$ for $k = X, Y$

where $z_k(.)$ is convex. Thus, as the number of drivers on the freeway increases, the time delay for each driver increases.

Agents have the following nested CES utility function:

---

[7] For simplicity we do not model carpooling so that all driving is by single-occupancy vehicles. In practice, vehicle occupancy for commuter trips is very low, about 1.1 (e.g., *www-cta.ornl.gov/npts/1995/Doc/trends_report.pdf*, Table 15).

[8] In practice there could be significant congestion associated with other travel options. Greater traffic volumes may also lead to more damages from vehicle accidents, though the evidence on this is mixed (e.g. Small and Gómez-Ibáñez, 1999). But our purpose here is simply to develop a model for quantifying the relative efficiency of alternative policies from reducing one source of congestion externality. The model could easily be extended to capture these other complications (e.g., Parry and Bento, 2000).

$$(2.3) \quad U = U(C^i, l^i, T^i) = \left\{ (C^i)^{\frac{\sigma_U - 1}{\sigma_U}} + (l^i)^{\frac{\sigma_U - 1}{\sigma_U}} + (T^i)^{\frac{\sigma_U - 1}{\sigma_U}} \right\}^{\frac{\sigma_U}{\sigma_U - 1}}$$

where

$$T^i = T^i(T_X^i, T_Y^i, T_B^i, T_{OP}^i, T_R^i) = \left\{ (T_X^i + T_Y^i)^{\frac{\sigma_T - 1}{\sigma_T}} + (T_B^i)^{\frac{\sigma_T - 1}{\sigma_T}} + (T_{OP}^i)^{\frac{\sigma_T - 1}{\sigma_T}} + (T_R^i)^{\frac{\sigma_T - 1}{\sigma_T}} \right\}^{\frac{\sigma_T}{\sigma_T - 1}}$$

C denotes consumption of market goods, l is leisure or non-market time, and T is travel services. The parameter σ_U is the elasticity of substitution between these three "goods;" it determines the price elasticity of total travel demand. In turn, sub-utility from travel services depends on the number of trips made on each of the travel modes. There is imperfect substitution between trips by freeway at rush hour, and trips by backroads, rail, and off-peak freeway travel. In other words, the demand curve for peak freeway travel is downward sloping (the freeway demand elasticity is primarily determined by σ_T). The benefit of peak-period freeway travel for agents is identical, regardless of which lane they travel on (though the costs may differ).

Our specification for utility is restricted in three notable respects. First, although the CES functional form keeps the results more transparent[9], it restricts all travel modes (other than the freeway lanes) to be equally good substitutes. Second, both agents have the same preferences. Third, we do not attach different distributional weights to the utility of low and high wage agents. Thus, our focus is on the pure economic efficiency of policy instruments rather than the broader social welfare effects, which take into account distributional impacts. Each of these restrictions is discussed further in Section 4C.

Agents are subject to the following budget constraint (prior to any congestion policies):

$$(2.4) \quad C^i + \sum_{j=R,B,OP} \theta_j T_j^i + \theta_F \sum_{k=X,Y} T_k^i = \rho^i L^i$$

[9] In particular, we can vary the freeway demand elasticity simply by varying σ_T, that is, changing the degree of substitution between the freeway and other modes in the same proportion. Allowing the degree of substitution between alternative modes to differ would require four different σ's in the $T(.)$ function, thereby complicating the calibration, sensitivity analysis, and interpretation of results.

where ρ^i is an agent's wage rate and L^i is labor supply. This equation equates expenditure on consumption and transportation with money income.[10] Type *H* agents have higher earnings (e.g., because of more skills), thus $\rho^H > \rho^L$. Therefore type *H* agents have a higher opportunity cost of travel time than type *L* agents.[11] Agents are also subject to the following time constraint:

$$(2.5) \quad l^i + \sum_{j=R,B,OP} \phi_j T_j^i + \phi_F \sum_{k=X,Y} (1+z_k) T_k^i = \bar{L} - L^i$$

That is, leisure time and travel time equals the agent's time endowment for the period ($\bar{L}$), minus labor supply. More time spent traveling therefore reduces utility due to reduced time available for leisure and work.

3. Mathematical Solutions: Comparing Policy-Induced Outcomes with the Efficient Outcome

In this section we describe the conditions that would be required for an efficient outcome, and then explore to what extent alternative policies do or do not meet these conditions. We go straight to the key equilibrium conditions here and relegate the (somewhat tedious) derivation of these conditions to Appendix A.

A. Efficient Outcome

This requires that two types of efficiency conditions are met. The first set of conditions, allocating travel on the freeway at peak period vis-à-vis other travel options, can be summarized by the following expression (see Appendix A):

$$(2.6) \quad \frac{U^i_{T_k^i}}{U^i_{T_j^i}} = \frac{\theta_k + \rho^i \phi_k (1+z_k) + MEC_k}{\theta_j + \rho^i \phi_j}; \qquad k = X, Y;\ j = R, B, OP$$

[10] We assume that the output of market goods (consumption and non-labor inputs in transportation) is simply proportional to labor supply. The price of market goods is therefore constant and normalized to unity.

[11] Empirical studies suggest that people value time spent travelling at, very roughly, 50 percent of the market wage (e.g., Small, 1992, pp. 43-45). Allowing for this would have the same effect as reducing the time cost parameters (the ϕ's). Thus, by choosing appropriate values for these parameters in our simulation analysis we can implicitly take into account appropriate values for the opportunity cost of travel time. In addition, the cost of travel time does not necessarily vary in proportion with the net wage across households (Verhoef and Small, 1999). But this does not bring into question our assumption that there are two types of drivers with different time costs.

where $MEC_k = \left(s_L \rho_L T_k^L + s_H \rho_H T_k^H\right)\phi_F z_k'$ denotes marginal external cost on freeway lane *k*. The left hand side of (2.6) is the marginal rate of substitution between peak freeway and other travel modes for agent *i*. The right hand side is the ratio of (social) cost per trip. The cost per trip equals the money cost, plus the time cost weighted by the agent's wage, plus the marginal external cost in the case of freeway drivers. In turn, the marginal external cost in the increase in travel time on the freeway lane caused by one more driver, $\phi_F z_k'$, multiplied by the sum of agents on the freeway lane, where an agent is weighted by her opportunity cost of time.

When equation (2.6) is satisfied for peak freeway travel vis-à-vis rail and backroads we say there is an efficient *inter-modal traffic allocation* and when (2.6) is satisfied for peak freeway travel vis-à-vis off-peak travel we say there is an efficient *inter-temporal traffic allocation*.

The second condition, the efficient *inter-lane traffic allocation*, relates to the distribution of agents across the freeway lanes at peak period. In particular, it is efficient to have a "separating equilibrium" with one faster-moving, less congested, freeway lane that is used intensively by agents with high time costs, and one slower moving lane, used intensively by agents with low time costs (see Small and Yan, 1999, for more discussion). More specifically:

(i) When the fraction of agents with high time costs is relatively large these agents use both lanes, while drivers with low time costs use only the slow lane. In this case the efficiency condition (see Appendix A) is

$$(2.7a) \quad (1+z_X)\phi_F \rho^H + MEC_X = (1+z_Y)\phi_F \rho^H + MEC_Y \qquad \Rightarrow \quad z_X < z_Y$$

This equation equates the (social) cost of one more trip across both freeway lanes for high wage agents. The private time cost to the driver is greater on the slow lane, *Y*, (*X* is the fast lane) since there is more congestion ($z_Y > z_X$). But this is compensated for by a lower (marginal) external cost ($MEC_Y < MEC_X$)—essentially, the costs of adding to congestion in the slow lane are smaller because the value of time (aggregated across drivers) is lower (again, see Appendix A for a proof).

(ii) If instead the fraction of agents with low time costs is relatively large, they use both lanes while drivers with high time costs use only the fast lane. In this case the efficiency condition (see Appendix A) is

$$(2.7b) \quad (1+z_X)\phi_F \rho^L + MEC_X = (1+z_Y)\phi_F \rho^L + MEC_Y \qquad \Rightarrow \quad z_X < z_Y$$

That is, the (social) cost of one more trip for a low time cost driver is the same on both lanes.[12]

The efficient outcome can be induced by "ideal" congestion pricing involving a charge of MEC_X for using the fast lane and MEC_Y for using the slow lane. With no policy intervention the efficiency conditions (2.6) and (2.7) are not met because drivers ignore the *MEC* terms. In equilibrium the private rather than social costs of the marginal trip are equated across modes and across peak/off-peak hours. Similarly, the private cost of a trip for the marginal driver is the same on both freeway lanes, hence $z_X = z_Y$.

B. Policy-Induced Outcomes

(i) Uniform congestion tax. Under this policy, drivers are charged an amount τ_C for using the freeway at peak period, regardless of which lane they use. The equilibrium conditions are (see Appendix A)

$$(2.8) \quad \frac{U^i_{T^i_k}}{U^i_{T^i_j}} = \frac{\theta_k + \rho^i(1+z_k)\phi_F + \tau_C}{\theta_j + \rho^i\phi_j} \qquad k = X, Y; j = R, B, OP$$

$$(2.9) \quad z_X = z_Y$$

Comparing (2.9) with (2.7) we see that this policy fails to induce the efficient inter-lane traffic allocation. In equilibrium the private time and money costs (including the congestion fee) are equal across lanes for each agent. Thus, congestion must be the same on both lanes, and there is a "pooling equilibrium" with both types of agent indifferent between each lane. However, the policy does raise the private cost of peak freeway travel without affecting the relative costs of other travel options. Therefore it can induce the efficient inter-modal and inter-temporal traffic allocations (though these are second best because the efficient inter-lane allocation is not achieved).

(ii) Single lane toll. Under this policy drivers must pay a charge of τ_X for using lane *X* on the freeway at peak period. The equilibrium conditions (see Appendix A) are

[12] The complication of having to deal with two possible equilibria (one with *H* drivers on both lanes and the other with *L* drivers on both lanes) is avoided in Verhoef and Small (1999) by assuming a continuum of driver types.

$$(2.10)\quad \frac{U^i_{T^i_X}}{U^i_{T^i_Y}} = \frac{\theta_F + \rho^i(1+z_X)\phi_F + \tau_X}{\theta_j + \rho^i\phi_j};\ \frac{U^i_{T^i_X}}{U^i_{T^i_Y}} = \frac{\theta_F + \rho^i(1+z_Y)\phi_F}{\theta_j + \rho^i\phi_j}\quad j = R, B, OP$$

$$(2.11)\quad z_X < z_Y$$

For a given amount of peak freeway traffic this policy can potentially induce the efficient inter-lane allocation, since the higher marginal external cost on the fast lane can be reflected in the toll. But, since the congestion externality on the slow lane goes unpriced the policy cannot optimally raise the overall costs of peak freeway travel relative to other travel options. Thus, its ability to achieve inter-modal and inter-temporal traffic efficiency is limited.

In short the single lane toll achieves what the uniform congestion tax does not, and the uniform congestion tax achieves what the single lane toll does not. A combination of these two instruments could yield the most efficient outcome.

(iii) Gasoline tax. This policy involves a tax of τ_G on gasoline expenditures. The equilibrium conditions (see Appendix A) are

$$(2.12)\quad \frac{U^i_{T^i_k}}{U^i_{T^i_j}} = \frac{(1+\tau_G g_k)\theta_k + \rho^i(1+z_k)\phi_k}{(1+\tau_G g_j)\theta_j + \rho^i\phi_j} \qquad k = X, Y;\ j = R, B, OP$$

$$(2.13)\quad z_X = z_Y$$

where the g's denote the fraction of monetary costs that are gasoline costs on a travel mode. The proportionate increase in the private cost of using a travel option is greater, the greater the ratio of gasoline costs ($g\theta$) to the total private costs of using that mode. To an approximation, gasoline taxes have no impact on the cost of electric urban rail systems hence they raise the price of peak-period freeway driving relative to rail travel. But the inter-modal and inter-temporal efficiency conditions are still violated, at least if the gasoline tax drives up the cost of all driving options in the same proportion. In addition, the policy does not induce the efficient inter-lane allocation, because in equilibrium drivers are indifferent between freeway lanes.[13]

[13] In the long run, gasoline taxes encourage the development of more fuel-efficient cars and hence reduce the g's and θ's. Allowing for this would not really affect the flavor of our results, however, since the effect is proportionate across all driving options.

(iv) Transit fare subsidy. Under this policy, agents receive a subsidy (s) for monetary expenditures on rail trips. We assume the subsidy does not vary with the time of day. This policy produces the following equilibrium conditions (see Appendix A):

$$(2.14)\quad \frac{U^i_{T^i_k}}{U^i_{T^i_R}} = \frac{\theta_k + \rho^i(1+z_k)\phi_F}{(1-s)\theta_R + \rho^i\phi_R} \qquad k = X, Y$$

$$(2.15)\quad z_X = z_Y$$

From (2.14) we see that the transit subsidy effectively raises the private cost of peak freeway travel relative to the cost of rail transit. However, the policy still violates the inter-modal efficiency condition because it also reduces the price of rail travel relative to backroads travel. Similarly, since it does not affect the price of peak- versus off-peak driving, it cannot produce the efficient inter-temporal travel allocation. Finally, the policy does not produce inter-lane efficiency either because it does not create differential pricing on freeway lanes.

We summarize the main points from this section in Table 1.

4. Numerical Analysis

We now explore quantitatively how the failure to meet the efficiency conditions affects the relative economic performance of the different policy instruments. To do this requires specifying a functional form for congestion, and details on this are provided in Appendix B. Subsection A below describes the parameter values used to calibrate the model; subsections B and C present the simulation results and sensitivity analysis.

Table 1. Potential for Meeting Efficiency Conditions under Alternative Policies

	Uniform congestion tax	Single lane toll	Gasoline tax	Transit subsidy
Inter-modal allocation	yes	very limited	no	no
Inter-temporal allocation	yes	very limited	no	no
Inter-lane allocation	no	yes	no	no

A. Parameter values

We are not concerned with the absolute efficiency gains from policy intervention per se—these obviously vary with the size and other characteristics of specific transportation systems. Instead we calculate the fraction of the maximum efficiency gain (under ideal congestion pricing) that can be achieved under alternative policies and over wide ranges of values for key parameters. These parameters include the allocation of trips across travel options, the demand elasticity for peak-period freeway use, the relative cost of traffic congestion, and the distribution of time costs across freeway drivers.

We assume that, prior to policy intervention, travel trips by each agent are distributed as follows: 33 percent by freeway driving at peak hours; 33 percent by rail; and 33 percent by driving on non-congested roads (16.5 percent on backroads and 16.5 percent on the freeway at off-peak hours).[14] The relative efficiency effects of some policies, notably the transit subsidy and gasoline tax, are sensitive to these assumptions, and later we report results for alternative traffic allocations.

We choose the transport mode substitution elasticity σ_T to imply that the (magnitude of the) demand elasticity for peak freeway trips is 0.2, 0.4 or 0.8. These values about span the range of estimates from the literature.[15] Note that when the freeway demand elasticity is 0.4, and we use our benchmark assumptions about the initial traffic allocation, then the cross-elasticity of rail travel with respect to the price of peak freeway travel is 0.13.[16] We set $\sigma_U = 0.1$. This implies that 10 percent of the reduction in freeway travel induced by the ideal congestion tax is due to reduced overall demand for travel, and 90 percent is due to substitution into other modes, when the freeway demand elasticity is 0.4 (this is relaxed later).

Our model incorporates an approximately linear (and positive) relation between peak period trip time on the freeway (the inverse of the travel speed) and the traffic flow

---

[14] The share of peak-hour travel in city centers by rail is roughly about 20–30 percent (Pickrell, 1989).

[15] See e.g. the discussion in Small (1992), ch. 2. As a rough rule of thumb, the own price elasticity of demand for peak freeway travel seems to be about 0.33 (in the short run), though there have been a wide range of estimates in the literature. We use a slightly higher medium case value to allow for intertemporal substitution, that is, rescheduling trips to use the freeway at off-peak hours. Still, as demand elasticities go, this is a fairly low value, reflecting people's reluctance to give up the comfort, privacy, and convenience of their cars.

[16] In other words, a 10 percent increase in the cost of peak freeway driving leads to a 4 percent reduction in miles traveled, and a third of this 4 percent reduction is diverted onto rail. This is consistent with evidence from Pickrell's (1989) survey of ten U.S. cities.

($s^H T_k^H + s^L T_k^L$), over the relevant range of traffic reduction (see Appendix B).[17] In addition, we choose the initial traffic flow (relative to the free flow rate) to imply a "low congestion" scenario where the optimal reduction in traffic would be 10 percent when agents are homogeneous and the freeway demand elasticity is 0.4 (see Appendix B). We also consider a "high congestion" scenario when the optimal traffic reduction is 20 percent. Under our alternative assumptions about the freeway demand elasticity, the optimal traffic reduction varies between 6 and 24 percent.[18]

The degree of heterogeneity in time costs among agents determines the relative importance of inter-lane efficiency. We consider a variety of cases where the share of agents with high time costs varies from 0.25 to 0.75, and the higher wage is equal to between 1.5 and 3 times the lower wage. For each case we normalize ρ^L such that the average wage is always unity—hence the total cost of congestion is (approximately) constant across these distributions.

In practice, estimating the relative importance of the time and monetary costs of travel is tricky because of uncertainty over how to value travel time and to what extent vehicle depreciation varies with miles traveled (e.g., Small, 1992, pp. 75-85). However, the relative efficiency effects of the policies are not especially sensitive to these parameters. We assume that time and money costs are (initially) 40 percent and 60 percent respectively of the total costs on all modes.[19] Gasoline costs are assumed to account for 40 percent of total money costs for all driving options and zero for (electric) rail transit.[20]

B. Results for Benchmark Parameter Values

(i) Optimal congestion tax. We begin in Table 2 by illustrating the differential taxation across freeway lanes under ideal congestion pricing. When agents are homogeneous (first row) the

[17] We experimented with a more convex function, but this had little effect on the relative efficiency impacts of policies. See Small (1992), Ch. 3 for a discussion of the relationship between travel speed and flow.

[18] These scenarios for the optimal traffic reduction are roughly consistent with other studies (e.g., Repetto *et al.* (1992), Table 12, top panel).

[19] Small (1992) calculates that time costs are 32 and 48 percent respectively of total travel costs for an expressway and urban arterial respectively (assuming monetary costs consist of running costs and vehicle capital).

[20] Finally, we assume that the value of travel services is 10 percent of the value of output and that leisure is 50 percent of labor supply (our results are not sensitive to these assumptions). Revenues raised by taxes are returned to agents in lump sum transfers proportional to the burden of the tax they bear, while the transit subsidy is financed by a corresponding lump-sum tax. Efficiency changes are calculated by the sum of the proportionate change in utility for each agent, weighted by agent's full income. We solved the model using GAMS with MPSGE.

optimal congestion tax is the same for both freeway lanes. When agents are heterogeneous the optimal tax is greater on the fast lane. This is because the marginal external cost of one more driver is greater on the fast lane, since the resulting increase in trip time is more costly to the high wage agents on this lane. But the optimal tax differential between the lanes is generally fairly modest: even when the number of high and low wage agents is the same and the higher wage is three times the lower wage, the optimal tax on the fast lane is only 11-32 percent greater

Table 2. Optimum Congestion Tax on Fast Lane Relative to that on Slow Lane

Share of agents with high time costs	High wage relative to low wage	Freeway demand elasticity					
		.2		.4		.8	
		low congestion	high congestion	low congestion	high congestion	low congestion	high congestion
0	1	1	1	1	1	1	1
.25	1.5	1.03	1.02	1.04	1.03	1.06	1.06
.25	3	1.12	1.07	1.14	1.11	1.18	1.24
.5	1.5	1.06	1.04	1.07	1.06	1.11	1.09
.5	3	1.18	1.11	1.19	1.16	1.25	1.32
.75	1.5	1.03	1.02	1.05	1.03	1.04	1.04
.75	3	1.09	1.06	1.11	1.07	1.06	1.10

than on the slow lane (see also Small and Yan, 1999). One reason for this is that, although the average time cost of agents is lower on the slow lane, there are more drivers on this lane, and this reduces the difference in the marginal external cost of congestion between the lanes.[21]

(ii) Uniform congestion tax. In Table 3 we show the efficiency potential from the uniform congestion tax, expressed as a fraction of the maximum efficiency gain from ideal congestion pricing. All cell entries in the first row are equal to one, implying that when agents are homogeneous this policy can induce the most efficient outcome.

However the key point from this table is that when we allow for driver heterogeneity, even though the policy fails to sort out drivers with high and low time costs onto fast and slow lanes on the freeway, the resulting efficiency loss is not very large. Typically, this policy achieves more than 90 percent of the maximum efficiency gains. This mirrors our previous result that the difference in marginal external costs between the freeway lanes in the efficient outcome,

[21] In addition, as we increase the fee on the fast lane, some drivers with high time costs tend to displace drivers with low time costs (who move onto other modes) on the slow lane. This displacement effect is weaker when the degree of substitution between the freeway and other modes is stronger, hence the ratio of the optimal tax on the fast lane to the tax on the slow lane increases (slightly) with the freeway demand elasticity (Table 2).

Table 3. Relative Efficiency Gain from Uniform Congestion Tax

Share of agents with high time costs	High wage relative to low wage	Freeway demand elasticity .2 low congestion	.2 high congestion	.4 low congestion	.4 high congestion	.8 low congestion	.8 high congestion
0	1	1	1	1	1	1	1
.25	1.5	.99	1	.99	1	.99	1
.25	3	.93	.97	.95	.97	.96	.97
.5	1.5	.97	.99	.98	.99	.99	.99
.5	3	.82	.94	.89	.96	.93	.97
.75	1.5	.99	1	.99	1	.98	1
.75	3	.94	.98	.97	.99	.98	.99

and hence the efficiency gain from differentiated congestion fees, is mitigated somewhat by the larger number of drivers on the slow lane. Thus, so long as inter-modal and inter-temporal efficiency is satisfied, the additional gains from achieving the efficient inter-lane allocation are limited.

(iii) Single lane toll. Table 4 displays the efficiency potential of the single lane toll relative to the maximum efficiency gain under ideal congestion pricing. There are several points worth noting here.

When agents are homogeneous (first row) the efficiency potential of the single lane toll is minimal—it can capture only 3-10 percent of the maximum efficiency gains (see Liu and McDonald, 1998, Anderson and Mohring, 1996, for more discussion). The key problem here is that this policy does not address congestion on the unpriced lane. Moreover, to the extent that

Table 4. Relative Efficiency Gain from Single Lane Toll

Share of agents with high time costs	High wage relative to low wage	Freeway demand elasticity .2 low congestion	.2 high congestion	.4 low congestion	.4 high congestion	.8 low congestion	.8 high congestion
0	1	.03	.03	.05	.06	.09	.10
.25	1.5	.07	.06	.09	.08	.13	.13
.25	3	.19	.11	.19	.14	.23	.19
.5	1.5	.11	.07	.12	.10	.16	.17
.5	3	.33	.18	.30	.20	.31	.24
.75	1.5	.06	.05	.09	.08	.12	.13
.75	3	.17	.10	.16	.12	.20	.18

drivers substitute away from the priced lane by using the unpriced lane more often, they compound congestion on the slow lane and this works to offset the efficiency gains from reduced congestion on the fast lane. In fact, given that the freeway lanes are perfect substitutes in demand while freeway travel and other modes are imperfect substitutes, about 80 percent of the displaced

traffic on the fast lane ends up as additional traffic on the slow lane. Thus, it is not surprising that the efficiency potential from the policy is so small.[22]

Allowing for heterogeneity does enhance the efficiency potential of the single lane toll. The policy now works towards the efficient inter-lane allocation by separating out drivers with high and low time costs onto the priced and unpriced lanes. However, even when there is a substantial amount of heterogeneity, the single lane toll can achieve only a minor fraction of the maximum efficiency gains. For example, when there is the same number of high and low wage drivers and the higher wage is three times the lower wage, the efficiency gains are still only 18-33 percent of the maximum gains.[23]

The results from Tables 3 and 4 suggest therefore that the efficiency gains from extending single lane tolls to cover other lanes on the freeway can swamp the efficiency gains from the initial imposition of the single lane toll. Moreover, the estimates in Table 4 are really upper bound estimates because they assume that half of the freeway lanes are covered by the individual lane toll. In practice, tolls may only cover one out of three or four lanes.

(iv) Transit fare subsidy. Table 5 shows that the failure of the rail subsidy to achieve the efficiency conditions can dramatically limit its overall economic potential. The top row shows the relative efficiency potential from the rail subsidy in the model with homogeneous agents. We see that the rail subsidy only captures 11-24 percent of the maximum efficiency gains. The basic problem here is that the rail subsidy does not induce any substitution away from peak-period freeway travel onto non-transit travel options (i.e., travel on backroads or on the freeway at off-peak hours). We discuss these problems in more detail below.

---

[22] The inter-lane substitution is somewhat sensitive to the curvature of the travel time/traffic flow curve, that is the rate at which congestion increases on the slow lane.

[23] See Small and Yan (1999) for more discussion. Note that the increased efficiency potential from incorporating heterogeneity is somewhat larger than would be suggested by a comparison of Tables 2 and 3. However, allowing for heterogeneity also increases, slightly, the ability of the single lane toll to induce inter-modal and inter-temporal substitution. This is because freeway lanes are no longer viewed by the different agents as perfect substitutes.

Increasing the freeway demand elasticity usually, though not always, improves the relative performance of the single lane toll. A higher elasticity facilitates the creation of a faster lane, since low wage agents are more willing to move onto other modes. But on the other hand, a higher elasticity also raises the relative efficiency gains from inter-modal and inter-temporal substitution, hence compounding the inefficiency of the single lane toll.

Table 5. Relative Efficiency Gain from Rail Subsidy

Share of agents with high time costs	High wage relative to low wage	Freeway demand elasticity					
		.2		.4		.8	
		low congestion	high congestion	low congestion	high congestion	low congestion	high congestion
0	1	.11	.12	.17	.19	.21	.24
.5	1.5	.10	.12	.17	.19	.21	.28
.5	3	.09	.12	.16	.19	.20	.24

In the second and third rows in Table 5 we see that allowing for heterogeneity among drivers tends to further reduce the efficiency potential of the rail subsidy, but only by a slight amount.[24] Thus, the failure to satisfy the inter-modal and inter-temporal efficiency conditions is much more important than the failure to satisfy inter-lane efficiency in explaining the very limited economic potential of the rail subsidy.

Finally, note that the relative performance of the transit subsidy worsens as the freeway demand elasticity is reduced. In the efficient outcome a small part of the reduction in peak period travel is due to agents reducing their overall demand for travel services, rather than substituting between travel modes. An additional drawback of the transit subsidy is that it *increases* the overall demand for travel services. This source of inefficiency becomes more significant as we reduce the willingness of agents to substitute between travel modes relative to the overall demand for travel services.

(v) Gasoline tax. Table 6 reports the efficiency results for the gasoline tax. This policy is operationally similar to the rail subsidy in our model. It raises the cost of peak-period freeway travel relative to rail travel, but since the cost of all driving options increases by the same proportion, it does not induce any substitution away from peak-freeway travel onto backroads or off-peak travel. Thus, the policy only induces a minor fraction of the maximum efficiency gains—between 25 and 38 percent.

Note that the efficiency potential of the gasoline tax is somewhat better than that of the rail subsidy. This is because the gasoline tax reduces the overall demand for travel services, hence avoiding the additional source of inefficiency under the transit subsidy. The relative efficiency discrepancy between the gasoline tax and rail subsidy is therefore noticeably more

[24] We do not report the results when the share of high wage agents is 0.25 and 0.75 since the reduction in efficiency is even smaller than when this share is 0.5.

Table 6. Relative Efficiency Gain from Gasoline Tax

Share of agents with high time costs	High wage relative to low wage	Freeway demand elasticity .2 low congestion	.2 high congestion	.4 low congestion	.4 high congestion	.8 low congestion	.8 high congestion
0	1	.35	.38	.30	.34	.28	.34
.5	1.5	.29	.36	.28	.34	.27	.33
.5	3	.33	.36	.28	.32	.25	.29

pronounced when the freeway demand elasticity is very small, since the overall substitution out of travel services is more significant relative to substitution within travel modes. As noted below, as we reduce the overall travel demand elasticity to zero, the efficiency impacts of the gasoline tax and transit subsidy converge.[25]

In short we have seen that, under our benchmark parameter scenarios, achieving the inter-modal and inter-temporal efficiency condition is much more important than achieving the efficient inter-lane allocation. The uniform congestion tax, which satisfies the first two conditions, can usually produce more than 90 percent of the maximum efficiency gains, while the single lane toll, rail subsidy and gasoline tax, which fail to satisfy these two conditions, sacrifice around two thirds or more of the maximum efficiency gains.[26]

(vi) Optimal traffic reductions. In Table 7 we compare the optimal (percentage) reduction in peak-period freeway travel under the different policies. Under ideal congestion pricing this varies from 5.9 percent (low congestion, low freeway demand elasticity) to 24.5 percent (high congestion, high freeway demand elasticity). Optimal traffic reductions under the uniform congestion tax are very similar, but they are dramatically smaller under the other three policies. For example, they vary between 0.6 percent and 7.7 percent under the rail subsidy, and between 0.4 percent and 3.7 percent under the single lane toll.

[25] Our model overstates the efficiency gain from a gasoline tax because it excludes car travel that does not compete with the congested freeway, but which would be distorted by a gasoline tax. But, since our main theme is that huge efficiency gains are forgone by using gasoline taxes (and other instruments) instead of a uniform congestion tax, this omission makes our results conservative.

[26] Another policy we might have examined is a time-invariant freeway toll applied uniformly across both lanes (this type of toll is common on highways in New Jersey). The efficiency potential of this policy would be between that for the uniform congestion tax and the gasoline tax. This policy is worse than the uniform congestion tax because it covers off-peak travel, but is better than the gasoline tax because it does not cover driving on backroads.

Table 7. Optimal Reduction in Peak Freeway Traffic

	Freeway demand elasticity					
	.2		.4		.8	
	low congestion	high congestion	low congestion	high congestion	low congestion	high congestion
Homogeneous agents						
Ideal congestion pricing	5.9	10.2	10.1	16.2	16.9	23.5
Uniform congestion tax	5.9	10.2	10.1	16.2	16.9	23.5
Single lane toll	0.4	0.4	1.1	1.1	2.8	2.4
Rail subsidy	0.6	1.2	1.7	3.2	3.8	5.6
Gasoline tax	2.2	4.0	3.0	5.6	4.9	9.0
Heterogeneous agents						
$s_H = .5$, $w_H = 1.5\ w_L$						
Ideal congestion pricing	6.1	10.4	10.2	16.9	16.5	20.7
Uniform congestion tax	6.0	10.2	10.5	16.5	16.9	21.3
Single lane toll	0.4	0.5	0.9	0.7	1.9	3.1
Rail subsidy	0.7	1.3	1.8	3.2	3.8	7.7
Gasoline tax	2.0	4.1	3.2	6.0	5.2	8.3
$s_H = .5$, $w_H = 3\ w_L$						
Ideal congestion pricing	6.2	10.8	10.6	17.3	17.6	24.5
Uniform congestion tax	6.0	10.3	10.6	17.5	17.0	26.0
Single lane toll	0.7	0.1	1.4	2.0	3.2	3.7
Rail subsidy	0.7	1.4	1.9	3.5	3.0	6.7
Gasoline tax	2.0	4.1	3.1	5.9	5.1	6.7

C. Further Sensitivity Analysis and Discussion

We finish up this section by exploring the sensitivity of the above results to some additional parameter variations and specifications for utility. For simplicity, we focus on the case of low congestion costs, a freeway demand elasticity of 0.4, equal shares of high and low wage agents, and a high wage equal to three times the low wage.[27] The results are summarized in Table 8.

(i) Share of travel by peak freeway. In the second row in Table 8 we vary the share of trips by peak freeway travel between 0.1 and 0.6 (the shares on other modes are scaled up and down in the same proportion). The greater the share of trips by peak freeway the smaller the possibilities for substituting into other travel options and hence the smaller the relative efficiency loss from

[27] The flavor of our results is similar under alternative assumptions for these parameters.

policies that fail to achieve inter-modal and inter-temporal efficiency. This has some impact on raising the efficiency potential of the rail subsidy and gasoline tax. However, even when the freeway share is 0.6, these policies still only capture 20 percent and 40 percent respectively of the maximum efficiency gains (given our other parameter assumptions).

Table 8. Sensitivity of Relative Efficiency Gain to Additional Parameter Variation

	Uniform congestion tax	Single lane toll	Rail subsidy	Gasoline tax
1. Benchmark case	.89	.30	.16	.28
2. Peak freeway share =.1–.6	.91–.84	.29–.32	.06–.20	.09–.40
3. Rail traffic share =.1–.67	.89–.89	.30–.30	.03–.51	.09–.89
4. Overall traffic demand: σ_U =.05–.2	.89–.89	.30–.30	.19–.10	.24–.30
5. Freeway demand elasticity higher for low-wage agents	.91	.28	.18	.29
6. Dist. weight for low and high wage agents = 1.15, .85	.98	.14	.17	.29

Note: These results assume an (aggregate) freeway demand elasticity of 0.4, the low congestion cost scenario, the share of high cost drivers is 0.5 and the high wage is 3 times the low wage.

(ii) Share of travel by rail. In the third row we hold the peak freeway share of trips constant at 0.33, and vary the rail share of trips between 0.1 and 0.67 (the combined travel share on backroads and off-peak freeway varies between 0 and 0.57). When the transit share is 0.67, all travel is either on the freeway at peak hours or by rail. In this highly extreme case, there is no intertemporal allocation condition, and both the gasoline tax and the rail subsidy achieve the efficient inter-modal allocation. In fact the gasoline tax becomes equivalent to the uniform congestion tax—both policies sacrifice 11 percent of the maximum efficiency gains because they do not induce the efficient inter-lane allocation. But the rail subsidy still loses 49 percent of the maximum efficiency gains. As already noted, the rail subsidy increases rather than reduces the overall demand for travel services, and the efficiency loss from this effect is much more important because the subsidy applies to 67 percent rather than 33 percent of travel trips in this case.

As we allow for traffic on backroads/off-peak freeway, and reduce the relative share of traffic by rail, the potential efficiency gain from the rail fare subsidy rapidly declines. When the transit share is 0.33 and 0.1, the efficiency potential of the fare subsidy is 16 and 6 percent of the maximum efficiency gains (see Tables 5 and 8). In short these results underscore the basic point that, on efficiency grounds, the case for using fare subsidies to reduce congestion is generally very weak, and particularly when rail travel accounts for a relatively small share of non-freeway travel.[28] Similarly, the efficiency potential of the gasoline tax declines rapidly as we increase the share of driving on alternative non-congested roads. In contrast, the efficiency potential of the uniform congestion tax and single lane toll are independent of how non-peak-freeway trips are split between rail and driving on non-congested roads.

(iii) Overall travel demand elasticity. In the fourth row of Table 8 we vary σ_U between 0.05 and 0.2. This implies that under ideal congestion pricing the fraction of the reduction in peak freeway travel due to reduced overall demand for travel services—rather than substitution between travel modes—varies between .06 and .17. This has some modest impact on the relative efficiency effect of the rail subsidy and gasoline tax. Reducing the overall elasticity of demand for travel services reduces (and in the limit would eliminate) the relative efficiency difference between these two policies. This is because these policies have the same effect on inter-modal substitution in our model, but have opposite effects on the overall demand for travel services.

(iv) Heterogeneous preferences. A key parameter in the utility function is the transport mode substitution elasticity, σ_T. In practice low wage agents might be more willing to substitute between modes in response to pricing policies than high wage agents. We explore this possibility by increasing σ_T for low wage agents by 50 percent, and reducing it for high wage agents by 50 percent.

Comparing rows 1 and 5 in Table 8 we see that this generalization does not really affect the results. Under ideal congestion pricing, there is now even less substitution off the peak freeway by high wage agents and more by low wage agents. But the same qualitative response occurs under all the policies since, to varying degrees, they raise the cost of peak freeway driving

[28] Adding travel by bus to the analysis (hence reducing the modal share of rail) would further dilute the relative efficiency gains from a rail-only subsidy. But if the combined share of road and bus were higher, and a general public transit subsidy was applied, the efficiency gain would be larger.

and the freeway demand elasticity for low wage agents is increased relative to that for high wage agents.

(v) Incorporating distributional weights. We now compare policies using a broader notion of social welfare than pure economic efficiency. In the above simulations the proportionate change in the utility of each agent is weighted by their full income. In particular, we multiply the (full income-weighted) proportionate change in utility by 0.85 for high wage agents and 1.15 for low agents. Comparing rows 1 and 6, again this has little effect on the relative performance of most policies. This is not surprising, since we have assumed that modal shares are the same across agents, and therefore most policies do not bear more heavily on one type of agent than another.[29]

More generally though, modal shares may vary across agents. Suppose for example that peak freeway is used relatively intensively by high wage agents and transit is used relatively intensively by low wage agents. In this case high wage agents would bear a disproportionate burden of congestion taxes and single lane tolls, while low wage agents would benefit disproportionately from a transit subsidy. Thus, the relative performance of the transit subsidy would improve.

(vi) Relaxing the CES assumption. In practice, backroads driving is probably a closer substitute for peak-freeway driving than transit or off-peak driving. Suppose we were to use a more general utility function that allowed for this possibility, while holding the overall freeway demand elasticity constant. Most likely, the results for the uniform congestion tax and the single lane toll would not be affected—the efficiency of these policies depends on the overall freeway demand elasticity not the relative substitution between different modes. However, the relative efficiency of the rail subsidy would probably fall further. This is because the cross-substitution between rail and peak freeway is reduced, implying less impact on congestion from the transit subsidy. Similarly, the efficiency of the gasoline tax would fall, as it would induce less overall substitution out of driving.

[29] The exception is the single lane toll, whose efficiency is reduced further. This is because this policy disproportionately benefits the high wage agents by creating a faster lane for them.

5. Conclusion

This paper compares the efficiency potential of a variety of alternative (short-run) policies for reducing a congestion externality associated with peak period freeway travel. Our model incorporates three main conditions that would be required to generate an efficient outcome. These include equating the marginal social cost of trips across different travel modes at a point in time and between peak and off-peak travel. In addition, efficiency requires sorting out drivers with high and low time costs onto fast and slow lanes on the congested freeway.

We find that a congestion tax imposed uniformly across freeway lanes can generally achieve more than 90 percent of the maximum efficiency gains under ideal congestion pricing, even though it does not induce a separating equilibrium with fast and slow lanes on the freeway. Thus, inducing the efficient substitution by peak-period freeway users onto other travel modes and off-peak travel is much more important than creating lanes with differential speeds within the freeway. In contrast the efficiency gains from policies that do not optimally exploit all alternative travel modes to the congested freeway are severely limited. Transit fare subsidies, gasoline taxes and single lane tolls at best achieve only one third of the maximum efficiency gains, and often much less than this.

There are a number of caveats to bear in mind, however. First, certain policies might be justified for other reasons. For example, gasoline taxes can partly address externalities associated with mobile air pollutants. The marginal cost of transit travel may be declining, justifying some level of subsidy, because an increase in service frequency in response to higher demand can reduce a passenger's expected wait time on the platform (Mohring, 1972). Our analysis abstracts from these other sources of efficiency gain, since our focus is on the case that can be made for policies based on their congestion impacts alone.

Second, on the other hand, since we highlight the strong efficiency case for broad congestion taxes over other policies, we have deliberately been conservative in estimating the efficiency drawbacks of other policies. For example, if we allow for more than two lanes on the freeway, individual lane tolls will perform worse than in the above analysis if they are applied to fewer than half of the freeway lanes. If we incorporate car travel that does not compete with the congested freeway, such as driving associated with leisure activities, this would reduce the relative efficiency potential of the gasoline tax because the gasoline tax distorts the amount of this other travel.

Third, the only pre-existing source of distortion in our model is the congestion externality. In practice there are a variety of other distortions within the transportation sector that may significantly affect the relative efficiency effects of anti-congestion policies. These include externalities associated with accidents, pollution, and congestion on other competing routes, and pre-existing policies such as sub-optimal pricing of mass transit, parking subsidies or fees, and gasoline taxes (e.g., De Borger et al., 1997, Newbery, 1990, Parry and Bento, 2000). More generally, it is important to take into account how policies interact with the tax system. Parry and Bento (1999) have recently shown that the net effect of a revenue-neutral tax on congestion that is caused by people commuting to work could stimulate labor force participation at the margin. This can lead to an important source of efficiency gain because taxes drive a large wedge between the marginal social benefit and marginal social cost of labor. In contrast, a congestion tax with revenues returned as lump sum transfers rather than income tax cuts, reduces the return to work effort (net of commuting costs) and produces a relatively large efficiency loss in the labor market.

Fourth, for the most part, our analysis does not capture the distributional effects of policies. A comprehensive (general equilibrium) analysis would be tricky because distributional effects depend on how congestion policies affect equilibrium prices in labor, housing, and land markets. They also depend crucially on how revenues from the policies are recycled (or how polices are financed in the case of transit subsidies). For some discussion of these issues see e.g. Small (1981). Fifth, we have used a static model that ignores long run considerations such as the efficiency impacts of building more road and rail infrastructure. We also ignore potential efficiency effects arising from the impact of policies on housing location.

Finally, our analysis abstracts from the political feasibility of different policies. Clearly, there is a lot of hostility towards the idea of broad-based congestion taxes from motorists, even if they expect to recoup some of the revenues raised in the form of other tax cuts (e.g., Harrington et al., 1998). In this respect single lane tolls, if they make people aware of the virtues of road pricing, could eventually turn out to be a useful “Trojan horse” for broader-based congestion taxes.

References

Anderson, David L., and Herbert D. Mohring. 1996. Distributional Consequences of Congestion Pricing: Analysis of a Network with Heterogeneous Commuters. Discussion paper, Department of Economics, University of Minnesota.

Arnott, Richard, and Kenneth A. Small. 1994. The Economics of Traffic Congestion. *American Scientist* 82: 446-455.

Braid, Ralph M. 1996. Peak-load Pricing of a Transportation Route with an Unpriced Substitute. *Journal of Urban Economics* 40: 179-197.

De Borger, Bruno, S. Ochelen, Steff Proost and Didier Swysen. 1997. Alternative Transport Pricing and Regulation Policies: An Efficiency Analysis for Belgium in 2005. *Transportation Research* 2D: 177-198.

Department of Transportation. 1993. *National Transportation Statistics*, Annual Report, Bureau of Transportation Statistics, Washington, DC.

Dodgson, John S., and N. Topman. 1987. Benefit-Cost Rules for Urban Transit Subsidies: An Integration of Allocation, Distibutional and Public Finance Issues. *Journal of Transport Economics and Policy* 21: 57-71.

Downs, Anthony. 1992. *Stuck in Traffic: Coping with Peak-Hour Traffic Congestion*. Brooking Institution, Washington, DC.

Giuliano, Genieve. 1992. An Assessment of the Political Acceptability of Congestion Pricing. *Transportation* 19: 335-358.

Glaister, Stephen, and David Lewis. 1978. An Integrated Fares Policy for Transport in London. *Journal of Public Economics* 9: 341-355.

Harrington, Winston, Alan J. Krupnick and Anna Alberini. 1998. Overcoming Public Aversion to Congestion Pricing. Discussion paper 98-27, Resources for the Future, Washington, DC.

Liu, Louie Nan, and John F. MacDonald. 1998. Efficient Congestion Tolls in the Presence of Unpriced Congestion: A Peak and Off-Peak Simulation Model. *Journal of Urban Economics* 44: 352-366.

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

Newbery, David M. 1990. Pricing and Congestion: Economic Principles Relevant to Pricing Roads. *Oxford Review of Economic Policy* 6: 22-38.

Parry, Ian W.H., and Antonio M. Bento. 1999. Revenue Recycling and the Welfare Effects of Road Pricing. Discussion paper No. 99-45, Resources for the Future, Washington, DC.

Parry, Ian W.H. and Antonio M. Bento. 2000. Estimating the Welfare Effect of Congestion Taxes: The Critical Significance of other Distortions within the Transport System. Discussion paper, Resources for the Future, Washington, DC.

Pickrell, Donald H. 1989. *Urban Rail Transit Projects: Forecast Versus Actual Ridership and Costs*. Cambridge, MA: US Department of Transportation.

Pigou, Arthur C. 1912. *Wealth and Welfare*. Macmillan, London.

Poole, Robert W., Jr. and C. Kenneth Orski. 1999. *Building a case for HOT Lanes: A New Approach to Reducing Urban Highway Congestion*. Policy Study No. 257, Reason Foundation, Los Angeles, CA.

Repetto, Robert, Roger C. Dower, Robin Jenkins, and Jacqueline Geoghegan. 1992. *Green Fees: How a Tax Shift can work for the Environment and the Economy*. World Resources Institute, Washington DC.

Small, Kenneth A. 1981. The Incidence of Congestion Tolls on Urban Highways. *Journal of Urban Economics* 13: 90-111.

Small, Kenneth A. 1992. *Urban Transport Economics*. Fundamentals of Pure and Applied Economics 51, Harwood Academic Press.

Small, Kenneth A., and Jia Yan. 1999. The Value of "Value Pricing" of Roads: Second-Best Pricing and Product Differentiation. Discussion paper, University of California at Irvine.

Small, Kenneth A., and Jose A. Gómez-Ibáñez. 1998. Road Pricing for Congestion Management: The Transition from Theory to Policy. In K.J. Button and E.T. Verhoef (eds.), *Road pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility*. Cheltenham, UK: Edward Elgar.

Small, Kenneth A., and Jose A. Gómez-Ibáñez. 1999. Urban Transportation. Forthcoming in P. Cheshire and E.S. Mills, *Handbook of Regional and Urban Economics, Volume 3: Applied Urban Economics*. North-Holland: Amsterdam.

Strotz, Robert H. 1965. Urban Transport Parables. In J. Margolis (ed.), *The Public Economy of Urban Communities*. Washington, DC: Resources for the Future.

Varian, Hal. 1984. *Microeconomic Analysis*. (Second edition) Norton, New York.

Verhoef, Erik T., and Kenneth A. Small. 1999. Product Differentiation on Roads: Second-Best Congestion Pricing with Heterogeneity under Public and Private Ownership. Discussion paper, University of California at Irvine.

Vickrey, William S. 1963. Pricing in Urban and Suburban Transport. *American Economic Review Papers and Proceedings* 53: 452-465.

Walters, Alan A. 1961. The Theory and Measurement of Private and Social Cost of Highway Congestion. *Econometrica* 29: 676-699.

Appendix A: Deriving the First-Order Conditions Described in Section 3

(i) Efficient Solution. The efficient solution involves a separate tax on each freeway lane, denoted τ_k. To solve for these taxes, we obtain the agent's first order conditions, and then maximize utility subject to these constraints.

Using the equations of Section 2, and combining the time and budget constraints into a full income constraint, agent *i*'s optimization problem can be defined

$$V(\tau_X,\tau_Y,z_X,z_Y,G^i)=Max \quad U\left\{C^i,l^i,T^i(T_X^i,T_Y^i,T_B^i,T_{OP}^i,T_R^i)\right\}$$

$$+\lambda^i\left\{\rho^i\overline{L}+G^i-\rho^i l^i-\rho^i\sum_{j=R,B,OP}\phi_j T_j^i-\rho^i\phi_F\sum_{k=X,Y}(1+z_k)T_k^i-C^i-\sum_{j=R,B,OP}\theta_j T_j^i-\sum_{k=X,Y}(\tau_k+\theta_F)T_k^i\right\}$$

where the Lagrange multiplier λ^i is the marginal utility of income, G^i is a lump-sum transfer to agent *i* (see below) and *V*(.) is indirect utility. Differentiating utility with respect to C^i, l^i, T_j^i and T_k^i yields the following conditions:

(A1) $\quad U_{C^i}^i=\lambda^i; \quad U_{l^i}^i=\rho^i\lambda^i$

(A2) $\quad U_{T_k^i}^i=\lambda^i\left\{\rho^i\phi_F(1+z_k)+\tau_K+\theta_F\right\}; \quad U_{T_j^i}^i=\lambda^i\left\{\rho^i\phi_j+\theta_j\right\}$ for $k = X, Y$ and $j = B, OP, R$

From differentiating the indirect utility function

(A3) $\quad \dfrac{\partial V^i}{\partial\tau_k}=-\lambda^i T_k^i; \quad \dfrac{\partial V^i}{\partial z_k}=-\lambda^i\rho^i\phi_F T_k^i; \quad \dfrac{\partial V^i}{\partial G^i}=\lambda^i$

As in the numerical model, the efficient solution is found by maximizing the sum of utilities, where the proportionate change in utility of an agent is weighted by that agent's share of full income in aggregate full income (this is the Negishi procedure). That is, we maximize with respect to τ_X and τ_Y

$$\sum_{i=L,H}s^i\eta^i V(\tau_X,\tau_Y,z_X,z_Y,G^i)$$

where

(A4) $\quad \eta^i=\dfrac{\rho^i\overline{L}+G^i}{V^i\sum_{i=L,H}s^i(\rho^i\overline{L}+G^i)}$

and V^i is the value of utility for agent *i*. For τ_X this maximization yields

$$\text{(A5)} \quad \frac{dV}{d\tau_X} = 0 = s^L \eta^L \left\{ \frac{\partial V^L}{\partial \tau_X} + \frac{\partial V^L}{\partial z_X}\frac{dz_X}{d\tau_X} + \frac{\partial V^L}{\partial z_Y}\frac{dz_Y}{d\tau_X} + \frac{\partial V^L}{\partial G^L}\frac{dG^L}{d\tau_L} \right\}$$

$$+ s^H \eta^H \left\{ \frac{\partial V^H}{\partial \tau_X} + \frac{\partial V^H}{\partial z_X}\frac{dz_X}{d\tau_X} + \frac{\partial V^H}{\partial z_Y}\frac{dz_Y}{d\tau_X} + \frac{\partial V^H}{\partial G^H}\frac{dG^H}{d\tau_L} \right\}$$

Tax revenues are returned lump sum to agents in proportion to the burden of the congestion tax burden they bear. That is, $G^i = \tau_X T_X^i + \tau_Y T_Y^i$, and

$$\text{(A6)} \quad \frac{dG^i}{d\tau_X} = T_X^i + \tau_X \frac{dT_X^i}{d\tau_X} + \tau_Y \frac{dT_Y^i}{d\tau_X}$$

Note that for CES preferences, an agent's indirect utility is proportional to full income (Varian, 1984, pp. 129-130), therefore $\lambda^i / V^i = (\rho^i \overline{L} + G^i)^{-1}$. Using this, and substituting (A3) and (A6) into (A5), we can obtain after some manipulation

$$\text{(A7)} \quad (\tau_X - MEC_X)\frac{dT_X}{d\tau_X} + (\tau_Y - MED_Y)\frac{dT_Y}{d\tau_X} = 0$$

where

$$MEC_X = (s_L \rho_L T_X^L + s_H \rho_H T_X^H)\phi_F z_X' \qquad MEC_Y = (s_L \rho_L T_Y^L + s_H \rho_H T_Y^H)\phi_F z_Y'$$

$$\frac{dT_X}{d\tau_X} = s^L \frac{dT_X^L}{d\tau_X} + s^H \frac{dT_X^H}{d\tau_X} \qquad \frac{dT_Y}{d\tau_X} = s^L \frac{dT_Y^L}{d\tau_X} + s^H \frac{dT_Y^H}{d\tau_X}$$

Similarly, we can obtain a symmetrical condition to that in (A7) for τ_Y. These two equations are satisfied when $\tau_X = MEC_X$ and $\tau_Y = MEC_Y$. Using these solutions and (A2) it is straightforward to obtain equation (2.6).

To derive the efficient inter-lane allocation, we define some critical number of low and high wage agents, $\tilde{s}^L$ and $\tilde{s}^H$, for which efficiency requires all low wage agents use the slow lane *Y* and all high wage agents use the fast lane *X*. For any other number of low and high wage agents, $\tilde{s}^L - \Delta$ and $\tilde{s}^H + \Delta$, then we demonstrate that if $\Delta > 0$ equation (2.7a) holds and if $\Delta < 0$ equation (2.7b) holds.

We can define the following expressions:

$$C_X^H = \theta_F + (1+z_X)\phi_F \rho^H + MEC_X$$

(A8) $$C_Y^H = \theta_F + (1+z_Y)\phi_F \rho^H + MEC_Y$$

$$C_X^L = \theta_F + (1+z_X)\phi_F \rho^L + MEC_X$$

$$C_Y^L = \theta_F + (1+z_Y)\phi_F \rho^L + MEC_Y$$

These expressions are the (social) cost of one more trip by agent *i* on freeway lane *k*, which include the private money and time costs, plus the marginal external cost.

Consider the case when the number of low and high wage agents is $\tilde{s}^L$ and $\tilde{s}^H$ where $\tilde{s}^L > \tilde{s}^H$. In addition

(A9) $$C_X^H = C_Y^H$$

(A10) $$C_X^L = C_Y^L$$

that is, the social cost of one extra trip by high wage agents would be the same on either freeway lane, and the same for low-wage agents. Suppose that the fast lane *X* is used exclusively by high wage agents and the slow lane *Y* by low wage agents. Thus, since there are fewer high wage agents, $z^X < z^Y$ and for conditions (8) to hold $MEC_X > MEC_Y$. Suppose a low wage agent were moved to lane *X* and a high wage agent to lane *Y*. On lane *Y*, the cost of the last trip increases by (the MEC terms in (A8) remain constant because the number of drivers on each lane is large)

$$C_Y^H - C_Y^L = (1+z_Y)\phi_F(\rho^H - \rho^L)$$

On lane *X* the cost of the last trip falls by

$$C_X^H - C_X^L = (1+z_X)\phi_F(\rho^H - \rho^L)$$

The net increase in cost is

$$(C_Y^H - C_Y^L) - (C_X^H - C_X^L) = \phi_F(\rho^H - \rho^L)(z_Y - z_X)$$

Since $z^X < z^Y$ this term is always positive. In other words, it is optimal for lanes *X* and *Y* to be used exclusively by high wage and low wage agents respectively.

Now suppose we increase the number of high wage agents on lane *X* by an arbitrarily small amount to $\tilde{s}^H + \Delta$, and reduce the number of low wage agents on lane *Y* to $\tilde{s}^L - \Delta$. In this case $C_X^H > C_Y^H$ since z_X and MEC_X have increased, while C_Y^L is less than C_X^L since z_Y and MEC_Y have fallen. Thus, it is optimal to shift high cost drivers off lane *X* and onto lane *Y* until

$C_X^H = C_Y^H$ and we obtain condition (2.7a). By a symmetrical logic, we obtain condition (2.7b) when the number of low wage agents is increased by Δ and the number of high wage agents reduced by Δ.

(ii) Policy-Induced Outcomes

If all the congestion policies were imposed together the individual maximization problem for agent *i* would be

$$Max \quad U\{C^i, l^i, T^i(T_X^i, T_Y^i, T_B^i, T_{OP}^i, T_R^i)\}$$

$$+\lambda^i\Big\{\rho^i\bar{L} + G^i - \rho^i l^i - \rho^i \sum_{j=R,B,OP}\phi_j T_j^i - \rho^i\phi_F \sum_{k=X,Y}(1+z_k)T_k^i - C^i - \sum_{j=R,B,OP}\theta_j T_j^i - \sum_{k=X,Y}\theta_F T_k^i$$

$$-\sum_{k=X,Y}\tau_k T_k^i + s\theta_R T_R^i - \tau_G \sum_{j=R,B,OP} g_j\theta_j T_j^i - \tau_G\phi_F \sum_{k=X,Y} g_k T_k^i\Big\}$$

The first order conditions for agent *i* are

$$U_{C^i}^i = \lambda^i$$

$$U_{l^i}^i = \rho^i\lambda^i$$

$$U_{T_k^i}^i = \lambda^i\{\theta_F + \rho^i\phi_F(1+z_k) + \tau_K + \tau_G\theta_F g_k\} \qquad k = X, Y$$

$$U_{T_j^i}^i = \lambda^i\{\theta_j + \rho^i\phi_j + \tau_G g_j\theta_j\} \qquad j = B, OP$$

$$U_{T_R^i}^i = \lambda^i\{\theta_R + \rho^i\phi_R + \tau_G g_R\theta_R - s\theta_R\}$$

Using these equations it is straightforward to derive the conditions in (2.8) to (2.14) by setting the appropriate policy parameters equal to zero (e.g. $\tau_G = s = 0$ and $\tau_X = \tau_Y$ for the uniform congestion tax). Note that, unless $\tau_X \neq \tau_Y$, then z_X must equal z_Y for agents to be indifferent between freeway lanes. When $\tau_X > 0$ and $\tau_Y = 0$, then equilibrium requires $z_X < z_Y$.

Appendix B. Specification of Congestion in the Numerical Model

In order to program the model in GAMS with MPSGE, we use a formulation for congestion that is similar to that in a model of traffic congestion developed by Thomas Rutherford.[30] A peak freeway trip on lane *k* by agent *i* requires a unit of a CES composite $\tilde{D}$ where

$$\text{(B1)} \quad \tilde{D}_k^i = \left\{ ((1+z_k)\phi_k T_k^i)^{\frac{\sigma_F - 1}{\sigma_F}} + (RA_k)^{\frac{\sigma_F - 1}{\sigma_F}} \right\}^{\frac{\sigma_F}{\sigma_F - 1}} \qquad k = X, Y$$

$\tilde{D}$ is "produced" by travel time and a variable *RA*, which denotes "road availability" per car per unit of time. Implicitly, the smaller *RA* is, the less space there is between individual cars and the slower the average driving speed. Thus, more time is required to produce a unit of $\tilde{D}$ and make a freeway trip, i.e. congestion z_k is greater.

Road availability is determined as follows

$$\text{(B2)} \quad RA_k = \frac{CAP}{(s^L T_k^L + s^H T_k^H) - \overline{T}}$$

where *CAP* is the (fixed) road capacity (e.g ten lane miles) and $\overline{T}$ is the free flow traffic level. This equation just states that road availability per car per unit of time equals road capacity divided by the flow of traffic on the freeway lane—i.e., the total number of trips per unit of time, $s^L T_k^L + s^H T_k^H$—net of the free flow level.[31]

Equations (B1) and (B2) essentially determine a standard positive relation between traffic flow and trip time (the inverse of the travel speed). Lower values of σ_F imply a more convex (or less concave) relation between travel time and traffic flow. We choose σ_F to imply that this relation is roughly linear over the relevant range of traffic reductions (our results are not especially sensitive to alternative specifications). The denominator in (B2) determines the initial extent of congestion. We choose $\overline{T}$ to imply that the optimal traffic reduction is either 10 percent or 20 percent under our benchmark values for other parameters.

[30] A detailed discussion of this model can be found at *http://nash.colorado.edu/tomruth/congest/Index.html*.

[31] The relevant range for traffic flows under alternative policies is always well above $\overline{T}$, hence RA_k is always finite.