ROAD CONGESTION PRICING: WHEN IS IT A GOOD POLICY?

By Andrew W. Evans*

1. INTRODUCTION

Road congestion pricing is back on the transport policy agenda. As the irresistible force of increasing traffic demand meets the immovable object of fixed road space in urban areas, the benefits to be gained from intervening with congestion pricing grow greater. In addition, the technical costs of operating a pricing system grow smaller. Most transport economists have long favoured road congestion pricing, and their case seems to be growing stronger. Morrison (1986) and other papers in the same special issue of Transportation Research provide a useful survey. The current professional view in Britain is epitomised by the report Paying for Progress (Chartered Institute of Transport, 1990) written by nine transport economists and planners, who state that "the theoretical case for road pricing is irrefutable" (p. v). A similar American view is the Brookings Institution report Road Work (Small et al., 1989). Nevertheless, the only city where it has yet been adopted is Singapore; Hong Kong drew back after detailed technical and economic study (reviewed by Hau, 1990).

The aim of this paper is to consider some objections to road congestion pricing, irrefutable though the case is said to be. We do not contend that the objections make an irrefutable case against road pricing, but they do need consideration in principle and in practice. We accept of course that the economic benefit to be obtained from a congested road is higher with congestion pricing than without it. We also accept for the purpose of this paper that pricing systems are now practicable, and that their costs are small. If their costs are significant, the arguments against road pricing are strengthened. Two objections are:

(1) Congestion pricing may redistribute welfare from road users to the rest of the community, and introduce horizontal inequity between users who are subject to it and those who are not; moreover, the size of the net benefit obtained may be small compared with the scale of this redistribution.

---

*The Flinders University of South Australia, University College London, and Imperial College London. Earlier versions of this paper were presented at the Second International Conference on Privatisation and Deregulation in Passenger Transportation in Tampere, Finland, on 16-20 June 1991, and at university seminars. The author is grateful to participants and referees for comments.
Although congestion pricing provides correct incentives for road users to make efficient decisions about their use of the road system, it does so at the cost of setting up perverse incentives for governments.

We investigate these objections by means of a simple economic model of a congested road network in a stationary state. We adopt an exponential demand function and couple this with three different forms of congestion cost function. Our arguments are entirely theoretical, and entirely concerned with the use of the road network. We do not consider the costs or the practicalities of a pricing system. Nor do we consider environmental or safety arguments about road pricing, because these are different from congestion arguments.

The paper continues as follows. Section 2 presents the economic model for the case where the traffic is economically homogeneous. This has been called the "standard" case of road pricing (Layard, 1977; Starkie, 1986). In doing so, we restate the usual argument in favour of congestion pricing. Section 3 discusses the first objection above for the standard case. Section 4 discusses government incentives. Section 5 considers whether the arguments are altered if the traffic consists of a mixture of economic types, such as cars and buses. We conclude that the first objection may then be weaker, and indeed that a principal argument in favour of congestion pricing is the help it may bring to buses. Section 6 summarises the conclusions.

2. AN ECONOMIC MODEL OF A CONGESTED ROAD NETWORK

We consider a congested urban road network of fixed capacity, serving a variable number of trips with a fixed origin-destination pattern. We analyse the economics of this network by means of a pair of equations, a demand function and a congestion cost function. The demand function gives the amount of traffic per unit time, \( q \), measured in passenger-car-unit-kilometres (pcu-km), which would use the road network for any given average generalised cost (that is, variable user cost) per pcu-km, \( g \). In this paper, generalised cost is defined as the variable user cost net of the cost in uncongested conditions and with no road user charge: it thus comprises the opportunity cost of the delay due to congestion, \( v \), and any price or tax imposed for the use of the road, \( p \). We suppose that for \( q < c \) the average congestion delay is a function of the degree of congestion on the network, \( q/c \), where \( c \) is the fixed capacity of the network, measured in pcu-km per unit time. In the homogeneous-traffic or "standard" case, which we discuss initially, we also suppose that the value per minute of delay is the same for all pcu-km, from which it follows that \( v \), and therefore \( g \), is the same for all pcu-km. Therefore our pair of equations is:

\[
q = q(g)
\]  
and

\[
g = p + v(q/c) \quad (q < c)
\]

For any choice of \( p \), we may solve this pair of simultaneous equations for \( q \) and \( g \), and also deduce the consumer surplus, government revenue, and economic benefit per unit time.
2.1 Demand Function

We now consider the specific functional forms chosen for \( q \) and \( v \). For demand, \( q \), we have chosen the exponential form:

\[
q = \alpha \exp(-g/\mu) \quad (\alpha > 0, \mu > 0)
\]

where \( \alpha \) and \( \mu \) are parameters.

Parameter \( \alpha \) has the dimensions of pcu-km per unit time, and characterises the general level of traffic demand. More specifically, it is the number of pcu-km per unit time that would materialise if \( g \) were zero, or in other words if there were no congestion and no taxes or prices for road use; we therefore label it potential demand. \( \alpha \) may have any positive value, including a value that may greatly exceed the capacity of the road, \( c \); in that case, a large proportion of potential demand is deterred by a high value of \( g \), generated by high prices or high congestion delays or both. We treat \( \alpha \) as exogenous, beyond the control of policy-makers, and we are interested in exploring how the economics of road congestion pricing change as \( \alpha \) increases while \( c \) remains fixed. That is our interpretation of the irresistible force meeting the immovable object.

Parameter \( \mu \) has the same dimensions as \( g, p \) and \( v \), that is price ($) per pcu-km. Its simplest interpretation stems from the property of the exponential demand function that, whatever the level of realised demand, \( q \), may be, consumer surplus, \( s \), is proportional to \( q \), with constant of proportionality \( \mu \). That is,

\[
s = \mu q
\]

\( \mu \) is therefore \( s/q \), the average consumer surplus (or net user benefit) per pcu-km. Clearly, the more users value their journeys on average, the higher will be \( \mu \). Another interpretation of \( \mu \) stems from the property of the exponential demand function that the elasticity of demand with respect to generalised cost is \( g/\mu \). For any given value of \( g \), the elasticity is therefore inversely proportional to \( \mu \), and the higher \( \mu \) is, the less responsive is realised demand to changes in \( g \). In other words, the more users value their journeys, the less they will be deterred by marginal changes in \( g \).

We have chosen the exponential demand function because it is a widely used transport demand function. Its shape (illustrated in Figures 1 and 3 below) implies a skewed distribution of journey valuations, which accords with common sense; the majority of journeys are valued at less than the average valuation, but a few are valued very highly, and there is no upper limit to the values. In other words, there would always be some trips, no matter how much they cost. In addition, the exponential demand function has some convenient mathematical properties, including those mentioned above. The exponential demand function is widely used in urban transportation models, and in theoretical work on road economics, of which a recent example is Williams and Moore (1990), but we are not aware of any specific empirical support for or against it in the present context. It is also used to model public transport demand, where there is perhaps more empirical support (Glaister, 1984; Evans, 1985).

In order to test the sensitivity of our results to the choice of demand function, we reworked some of the results presented in the following section with the comparable linear demand function in place of the exponential. The main effects of congestion pricing are
similar over most of the demand range. The differences which exist work in the direction of weakening the case for congestion pricing, so our adoption of the exponential in preference to the linear demand function is conservative.

2.2 Congestion Cost Functions
We now turn to the congestion cost function, \( v \). It became clear during the work for this paper that there is no consensus about the appropriate form of a broad-brush congestion function for urban policy assessment, in spite of the long history of the analysis of road congestion pricing. Moreover, its form makes a big difference to the effects of congestion pricing and to the strength of the objections discussed here. It follows that further research work on congestion cost functions is needed.

Functions considered
In the absence of a consensus we consider three different congestion cost functions. They are:

1. The \textit{inverse linear} congestion cost function is derived from the assumption that there is a downward-sloping linear relationship between the average speed on an urban road network and the traffic throughput on the network (in pcu-kms per unit time); we assume that this relationship applies for all speeds from the uncongested speed down to a speed approaching zero. We define the capacity of the network as the limit of the throughput as the average speed tends to zero.

2. The \textit{truncated inverse linear} congestion cost function is likewise derived from an assumption of a linear relationship between average speed and average throughput, but we now assume that traffic reaches capacity at some non-zero speed. As demand increases speed may then fall below this level, and congestion delays increase, but there is no increase in throughput.

3. The \textit{bottleneck} congestion cost function is derived on the assumption that the urban road network performs like a pure bottleneck with a fixed capacity, in which there are no delays at all so long as demand is at or less than capacity, but on which delays rise, and speeds fall, indefinitely as potential demand increases.

Figure 1 shows these three congestion cost functions as the heavy curves, together with demand functions for two values of \( \alpha \). Figure 2 shows the corresponding relationships between the average speed on the urban network (as a percentage of the uncongested average speed) and the average traffic throughput (as a percentage of capacity). Figures 1 and 2 show the particular case of the truncated linear function in which the speed at which capacity is reached is half the uncongested speed, but obviously it would be possible to have other cut-off speeds.

The inverse linear congestion cost function has the property that average delay due to congestion increases throughout the range of traffic throughput, and increases without limit as throughput approaches capacity, as illustrated in Figure 1. This type of congestion function is the one most commonly used in expositions of the theory of congestion pricing.
FIGURE 1
Congestion Cost Functions

FIGURE 2
Speed-Throughput Relationships
This type may be derived in one form or another from a variety of probabilistic queueing models, including models of simple queues (Cox and Smith, 1961), and Webster's model for the delay at a fixed cycle traffic signal (see, for example, Kimber et al., 1986, p. 12).

The truncated inverse linear congestion cost function is similar to the linear one for some of the upper range of average speed, but it has the property that traffic throughput reaches capacity at a non-zero speed (point A in Figure 2), which in turn implies that throughput reaches capacity with a finite congestion delay (point A in Figure 1). If the demand function cuts the congestion function to the left of point A in Figure 2, the properties of the equilibrium are similar to those of the non-truncated function discussed above; an example is the demand function with $\alpha/c = 0.9$. However, for higher values of potential demand $\alpha$, the demand function may reach capacity above point A, say at point B in Figure 1. We then assume that equilibrium between demand and capacity is attained through additional non-probabilistic queueing delay; the amount of delay is what is needed to choke off enough potential demand to reduce realised demand to the capacity of the network. As potential demand increases, delays increase with no change in throughput.

The bottleneck congestion cost function may be thought of as the limiting case of the truncated linear function, in which the slope of the linear part tends to zero. Its main interest here is as the limit of the truncated linear, though situations with the characteristics of bottlenecks do exist. The time-dependent pure bottleneck has been extensively studied (Vickrey, 1969; Arnott et al., 1990).

**Parabolic speed-throughput relationship**

The parabolic relationship between speed and throughput, illustrated as the dotted curve in Figure 2, generates yet another set of congestion cost functions. We do not consider these in detail in this paper, but they do need discussion. The parabolic relationship is commonly derived by considering the speed-density relationship of traffic on an unimpeded highway such as a freeway (see, for example, Morrison, 1986). The parabola may be applied in two ways. First, the upper branch only may be combined with a vertical relationship for speeds below half the congested speed, making the whole relationship rather like the truncated linear. Alternatively, both branches of the parabola may be applied, in which case throughput is presumed to fall as speed falls below half the uncongested speed, eventually tending to zero as speed tends to zero.

It is sometimes argued that logically all speed-throughput relationships must bend back in this way to go through the origin, because when the average speed tends to zero, traffic flow must tend to zero. This is incorrect. The average speed of traffic on a network can be written as

\[
\frac{\text{pcu-kms per hour}}{\text{pcu-hours per hour}}.
\]

The numerator of this fraction is traffic throughput. If the fraction is to tend to zero, it is not necessary for the numerator to tend to zero; an alternative is that the numerator remains non-zero and finite (say at network capacity) while the denominator tends to infinity. That
is what is presumed to happen in the three congestion cost functions discussed above. Therefore, although a backward-bending speed-throughput relationship is an empirical possibility, it is not a logical necessity.

**Reasons for our choice of congestion cost functions**
Previous strategic-level analyses of congestion on urban road networks have commonly used linear or truncated linear speed-throughput relationships. Newbery (1988, p. 165) used them in his study of road user charges in Britain, citing empirical work by Duncan et al. (1980) based on the British Urban Congestion Studies of the 1970s, and by Harrison et al. (1986) on area speeds and flows in Hong Kong. The London Area Model, which is a broad-brush policy model used by the London Planning Advisory Committee for exploring alternative transport policies for London used them (May and Gardner, 1990). In a slightly different form, they are used in the official British road cost-benefit analysis program COBA (Department of Transport, 1980). None of these sources provides strong empirical evidence in favour of linear or truncated linear relationships, but these relationships have been the ones most commonly used in practice. We have followed this practice. We have added the bottleneck because it is the limit of the truncated linear relationship, and because it is of interest in its own right.

None of the sources above makes a distinction between the full linear and truncated linear relationships. Implicitly, they appear to work with truncated linear functions, but they assume that their demand functions cut the congestion cost function below capacity (to the left of point A in Figure 1), in which case the truncation makes no difference. However, in our context the truncation (if present) does make a difference.

We now explain why we do not pursue the congestion cost functions based on the parabolic speed-throughput function. The function which combines the upper branch of the parabola with a vertical relationship for low speeds at capacity is a serious contender, and we have carried out basic calculations with this function. However, we have omitted the results from the paper, because they are similar (though not identical) to the results from the truncated linear function, and we wish to avoid presenting too many results. It is not surprising that the results are similar, because both speed-throughput relationships combine downward slopes at higher speeds with vertical relationships at lower speeds.

The reason for disregarding the congestion function based on the backward-bending branch of the parabola is deeper. We have argued above that there is no logical requirement for speed-throughput relationships to bend backwards. The question whether they do is therefore an empirical one. Although there is evidence that traffic flows on freeways fall as the average speed falls below a certain value (see, for example, Keeler and Small, 1977, pp.10-13), the author is not aware of corresponding evidence that at a strategic level traffic throughput on urban networks actually falls as average journey speed falls. If it did, we might expect that reductions in pcu-kms would have been observed in large towns where potential demand has multiplied over the last few decades, but capacity has remained fairly static. The author is not aware of such cases; for example, the statistics for London (Department of Transport, 1989d, p.13) do not have that feature. Presumably
other analysts have come to the same conclusion, because the strategic studies cited above do not have backward-bending speed-throughput relationships. We therefore feel that the evidence is not strong enough to consider this case. However, it must be said that if throughput does fall with speed, then the case for congestion pricing becomes stronger.

**Algebraic formulation: inverse linear congestion cost function**

If average network speed is a linear function of throughput, \( q \), we may write it as proportional to \( 1 - (q/c) \), where \( c \) is the network capacity, which is by definition the traffic throughput at which the average speed falls to zero. Average journey time per pcu-km is the reciprocal of speed, and is proportional to \( 1/(1 - (q/c)) \). The value of travel time per pcu-km is therefore also proportional to \( 1/(1 - (q/c)) \). Let the constant of proportionality be a parameter \( \beta \), with dimensions $ per pcu-km. Then the value of travel time per pcu-km is \( \beta/(1 - (q/c)) \). The interpretation of \( \beta \) is that it is the value of travel time per pcu-km in uncongested conditions, when \( q \) is zero. The value of the delay due to congestion, \( v(q/c) \), is the difference between the value of travel time per pcu-km with \( q > 0 \), and the corresponding value with \( q = 0 \). Therefore

\[
v(q/c) = \beta \left( \frac{1}{1 - (q/c)} - 1 \right)
\]

Therefore

\[
g = p + v = p + \beta \left( \frac{1}{1 - (q/c)} - 1 \right)
\]

The marginal cost of congestion, say \( w \), with respect to an increase in the traffic is \( d(vq)/dq \), or \( (v + qdv/dq) \), which is given by

\[
w = \beta \left( \frac{1}{1 - (q/c)} - 1 \right) + \frac{\beta q}{c} \frac{1}{[1 - (q/c)]^2}
\]

The first term on the right of (7) is the cost of delay borne by the marginal pcu-km itself, and the second term is the cost imposed on all the other traffic. As is well known, the efficient congestion price is equal to this second term. The shapes of \( v \) and \( w \) are shown in Figure 3.

**Truncated inverse linear function**

The general truncated inverse linear function requires an additional parameter, namely the speed at which the throughput first reaches capacity (point A in Figure 2). However, we have considered only the specific case where this speed is half the uncongested speed, which is illustrated in Figures 1 and 2. We can visualise the results of other values of cut-off speed by noting that as the cut-off speed increases, the truncated linear function moves towards the bottleneck function, and as it decreases, it moves towards the full linear function.

For the particular case considered, by the same argument as in the linear case, for \( q < c \), \( v, g, \) and \( w \) are given by
FIGURE 3
Demand and Congestion Cost Functions: Inverse Linear Congestion Costs

\[ v = \beta \left( \frac{1}{1 - (q/2c)} - 1 \right) \]  \hspace{1cm} (8)

\[ g = p + \beta \left( \frac{1}{1 - (q/2c)} - 1 \right) \]  \hspace{1cm} (9)

and

\[ w = \beta \left( \frac{1}{1 - (q/2c)} - 1 \right) + \frac{\beta q}{2c} \left( \frac{1}{1 - (q/2c)} \right)^2 \]  \hspace{1cm} (10)

If \( q = c \), all these quantities are indeterminate. They are determined by the demand function, outlined below.

**Bottleneck function**
For the bottleneck function, for \( q < c \), \( v = 0 \), \( g = p \), and \( w = 0 \). For \( q = c \), all these quantities are undefined, and are determined by the demand function.
2.3 Equilibrium of demand and congestion cost functions
For any choice of \( p \), the resulting traffic throughput \( q \) is given by the intersection of the demand function and generalised cost. With the full inverse linear congestion cost function, the relevant equations are (3) and (6) giving the following equation for \( q \):

\[
q = \alpha \exp \left( -\frac{1}{\mu} \left( p + \frac{\beta}{[1-(q/c)]} - \beta \right) \right)
\]

(11)

This must be solved numerically. The equilibrium congestion delay then follows from (5).

The corresponding equation with the truncated function is

\[
q = \alpha \exp \left( -\frac{1}{\mu} \left( p + \frac{\beta}{[1-(q/2c)]} - \beta \right) \right)
\]

(12)

This applies provided that its solution \( q \) is less than the network capacity \( c \), that is, for values of potential demand \( \alpha \) in the range \( 0 < \alpha < c \exp[(p + \beta)/\mu] \). For values of \( \alpha \) greater than this, the demand function cuts the vertical part of the congestion cost function. In that case, \( q = c \), and \( g \) is obtained by inverting the demand function with \( q = c \) to give

\[
g = -\mu \ln (c/\alpha)
\]

(13)

from which

\[
v = -\mu \ln (c/\alpha) - p
\]

(14)

With the bottleneck function, \( v = 0 \) and \( q = \alpha \exp(-p/\mu) \) for \( 0 < \alpha < c \exp(p/\mu) \); for greater \( \alpha \), \( q = c \), and \( g \) and \( v \) are given by (13) and (14).

2.4 Government revenue, consumer surplus and economic benefit
We now consider the economic outcomes from the use of the road network. Government revenue, \( r \), is given by

\[
r = pq
\]

(15)

Road use imposes marginal costs on the government in the form of maintenance costs, traffic control, policing, and so on. Therefore the net government surplus from road use is less than gross revenue. However, without loss of generality, we may ignore the marginal cost per pcu-km to the government, provided that it is approximately constant, independent of the degree of congestion. This is a reasonable assumption for this analysis, especially as the marginal cost to the government is small compared with the other monetary quantities per pcu-km. In that case, we can measure \( p \), the tax or price per pcu-km, net of the marginal cost to the government. In other words, if \( p = 0 \), road users are paying just their marginal cost to the government. This makes \( r \) net revenue, and therefore pure government surplus.

By the same argument, if there are other constant marginal external costs per pcu-km, such as environmental costs, we can assume that these are being paid without affecting the discussion of congestion.

Consumer surplus or user benefit, \( s \), is given by (4) as \( \mu q \). The economic benefit from the network, \( b \), is the sum of government surplus and consumer surplus:

\[
b = r + s = pq + \mu q
\]

(16)
The three quantities $r$, $s$, and $b$ together measure the economic output of the network.

An important property of the exponential demand function applied to a network with fixed capacity is that consumer surplus has an upper bound. This is because $q \leq c$, $s = \mu q$, and therefore $s \leq \mu c$. This must hold however high the potential demand may be, and for any pricing system. It follows that if the economic benefit of the network substantially exceeds $\mu c$, much of the benefit must be in the form of government surplus. The quantity $\mu c$, the maximum possible consumer surplus, provides a convenient *numeraire* for economic output.

### 2.5 Maximising economic benefit

The optimal congestion price is that which maximises economic benefit. It satisfies different conditions according to whether the corresponding optimal traffic throughput is less than capacity, which can occur with all three of our congestion cost functions, or at capacity, which can occur only with the truncated linear and bottleneck functions.

*Optima with throughput less than capacity*

We consider optima with throughput less than capacity, and we first analyse the case of the full inverse linear congestion cost function, because with this function no other type of optimum is possible. We then consider how to modify the results for the other two congestion cost functions.

Let the optimal $p$ be $p_1$. The first-order condition for this maximum is that $p_1$ must satisfy $db/dp = 0$. This implies $q + (p + \mu) dq/dp = 0$.

If we use (3) and (6), and simplify, this gives

$$p_1 = \frac{\beta q}{c} \left( \frac{1}{1 - (q/c)} \right)^2$$

(17)

This is the well-known result that the optimal congestion price is equal to the value of the delay imposed by the marginal pcu-km on other traffic. If we substitute this value of $p$ in the equilibrium equation for $q$, and simplify, we obtain the following equation for the optimal value of $q$, say $q_1$.

$$q_1 = \alpha \exp \left( \frac{-\beta}{\mu} \left( \frac{1}{1 - (q_1/c)} \right)^2 - 1 \right)$$

(18)

This has to be solved numerically. The solution $q_1$ is what Morrison (1986) labels the "equilibrium short run" level of traffic. It reflects the response of traffic to the optimal congestion price, which makes it an equilibrium, but the road network capacity is treated as fixed, which makes it short-run. In the long-run equilibrium, network capacity would also be optimised. However, if it is impossible to vary the capacity of the road network, the short run is also the long run.

Having calculated $q_1$ and $p_1$, we can calculate the optimal values of $r$, $s$ and $b$ from (15), (4) and (16) respectively. We can likewise calculate $r$, $s$ and $b$ for any other value of $p$, say $p_0$, which may be either zero or positive, and represents the current non-congestion-related tax on road users over and above the marginal cost to the government of road use.
The differences in the values of $r$, $s$ and $b$ represent the economic effect of introducing congestion pricing, and are the main criteria by which we assess congestion pricing.

Figure 3 pulls together the results for the full linear congestion cost function, and is the classic road congestion diagram. The three downward-sloping curves represent three versions of the exponential demand function (3), with values of $\alpha/c$ of 0.5, 1, and 10. These curves are all drawn with the same value of $\mu$, namely $1$ per pcu-km, in some unspecified currency. $\mu$ is thus our numeraire for prices or costs per pcu-km. The two upward-sloping curves are the congestion cost function, $v$, and the marginal congestion cost, $w$, given by (6) and (7). They are drawn with $\beta = 0.40$, or 40 per cent of $\mu$, which we discuss below.

The intersections of the demand curves with the $v$-curve give the values of $q/c$ that would materialise if $p$ were zero for each level of potential demand. The intersections of the demand curves with the $w$-curve give the (lower) values of $q/c$ that would materialise if $p$ were optimal for each level of demand. The arrows between the $v$-curve and $w$-curve give the optimal congestion price for each level of demand. The shaded triangular-shaped areas give the increase in economic benefit which is gained by switching from a zero price to the optimal congestion price at each level of demand. The diagram shows that both the optimal congestion price and the economic benefit due to congestion pricing increase as the level of potential demand increases. If the initial value of $p$ were not zero, the gain in economic benefit would be rather smaller than is shown in the diagram, but it would still increase in the same way. Figure 3 thus restates the standard case for congestion pricing, and shows that the case becomes stronger as potential demand increases.

We now consider how this analysis applies to the truncated inverse linear congestion cost function. The equation corresponding to (18) to be satisfied by the optimal $q_1$ is

$$q_1 = \alpha \exp\left(-\frac{\beta}{\mu} \left(\frac{1}{1-(q_2/2c)} - 1\right)\right)$$  \hspace{1cm} (19)

but this applies only to that range of potential demand, $\alpha$, for which the solution $q_1 < c$. We can find the boundary of this range by setting $q_1 = c$ in (19), which gives $\alpha = c \exp(3\beta/\mu)$. Thus for $0 < \alpha < c \exp(3\beta/\mu)$, the optimal throughput is given by (19), and the optimal congestion price is then determined by the equation corresponding to (17). For $c \exp(3\beta/\mu) \leq \alpha$, the optimal $q_1 = c$, and the optimal congestion price is determined by the argument in the following section.

In the case of the bottleneck function, congestion delay is zero for $\alpha < c$; therefore the optimal price is also zero, and $q_1 = \alpha$. For $c \leq \alpha$, $q_1 = c$, and the optimal price is determined by the argument in the following section.

**Optima with throughput at capacity**

We now turn to optima where $q_1 = c$ for the truncated and bottleneck congestion cost functions. We consider the bottleneck case first. In this case, generalised cost $g$ is again given by (13), the inversion of the demand curve, $g = p + v = -\mu \ln (c/\alpha)$. The optimal value of $p$ is that which makes the value of congestion delay, $v$, exactly equal to zero, so

$$p_1 = -\mu \ln (c/\alpha)$$  \hspace{1cm} (20)
ROAD CONGESTION PRICING: WHEN IS IT A GOOD POLICY? A. W. Evans

The argument in the case of the truncated linear function is similar, but the optimal \( v \) is no longer zero, but \( \beta \), obtained by putting \( q = c \) in (8). Therefore the optimal congestion price is given by

\[
p_1 = -\mu \ln \left( \frac{c}{\alpha} \right) - \beta
\]

(21)

2.6 Maximising revenue

In addition to considering the maximisation of economic benefit from the road network, we also consider the effects of prices set so as to maximise the road operator’s revenue. This requires a set of equations broadly similar to those for maximising benefit; we omit the details.

2.7 Parameter values

The parameters in the model are \( \alpha \) (potential demand in pcu-km per unit time), \( c \) (road network capacity in pcu-km per unit time), \( \mu \) (average consumer surplus in $ per pcu-km), \( \beta \) (value of un congested travel time in $ per pcu-km), and \( p_0 \) (existing road use tax, net of marginal cost, in $ per pcu-km). As in Figures 1 to 3, we may conveniently take the value of \( c \) as the numeraire for traffic, giving it the value 1 without loss of generality and measuring \( \alpha \) and \( q \) in terms of \( c \). We may likewise take the value of \( \mu \) as the numeraire for costs and prices per pcu-km, giving it the value 1 without loss of generality, and measuring \( \beta \) and \( p \) in terms of \( \mu \). This makes the product \( \mu c \) the automatic numeraire for the economic outcomes \( r, s, \) and \( b \), and, as noted above, this numeraire is equal to the maximum possible consumer surplus per unit time. We do not specify any particular value of \( \alpha \), but consider below the whole range from zero to a large multiple of \( c \), because we are interested in how the effects of congestion pricing change with potential demand. In practice, instead of using \( \alpha \), we can usually express the underlying level of demand in other ways, such as the average journey speed before the introduction of congestion pricing or the initial traffic throughput. These have the advantages of being more easily interpreted, and, unlike \( \alpha \), of having a finite range. The value of \( \alpha \) is then implicitly deduced. However, the relationships between \( \alpha \) and the other variables do depend on the congestion cost functions and on the values of the other parameters in the model, and if these are changing we must be careful how we interpret comparisons.

That leaves the values of \( \beta \) and \( p_0 \) to be set in terms of the numeraires. We set \( \beta \) to be 40 per cent of \( \mu \), or $0.40. We set \( p_0 \) to be 5 per cent of \( \mu \), or $0.05. We now explain this choice of values.

\( \beta/\mu \) is the ratio of the average value of travel time per pcu-km in uncongested conditions to the average consumer surplus per pcu-km. We use data for Great Britain to estimate this ratio, mainly because of its familiarity to the author, though limited parallel data for Australia give similar results. Table 1 gives the value of time per vehicle-km in uncongested conditions for the main vehicle classes, derived from the Department of Transport’s standard values of time at 1988 prices (1989b, Table 3) by assuming an uncongested speed over urban networks of 40 km per hour. This is converted to a value of time per pcu-km in uncongested conditions, using the pcu-values of non-car vehicles.
given in Table 1. The value of time averaged over all vehicle classes is 13.2p per pcu-km, and the averages for the vehicle classes range from 8.9p for “other” (that is, non-light) goods vehicles to 40.2p for buses and coaches. We next need an estimate of the average consumer surplus per pcu-km, $\mu$, in the same currency. This is more difficult. Our approach is to make use of the previously-noted property of the exponential demand function that at any level of demand $\mu$ is the ratio of generalised cost, or any component of it, to (the negative of) the elasticity of demand with respect to that component. One such component is fuel prices, which for cars averaged about 3.5p per car-km in 1988 (Department of Transport, 1989b, Table 6). The Department of Transport (1989c, p. 20) has estimated the general elasticity of demand for car use with respect to fuel prices as -0.15, which with the exponential demand function implies an average consumer surplus of 3.5/0.15 or 23.3p per car-km. This is about twice the average value of uncongested urban travel time for cars of 11.7p per km, implying a value for $\beta$ of 50 per cent of $\mu$, or $0.50. However, the average consumer surplus per pcu-km for freight vehicles in urban areas is probably higher relative to the value of travel time than it is for cars, because there are fewer substitutes for freight movement, implying a lower value of $\beta/\mu$. We have therefore taken our overall average value of $\beta$ to be $0.40. This in turn implies an overall average consumer surplus in Britain of 33p per pcu-km.

Lastly, we must set $p_0$. Because the existing road tax per pcu-km in most countries is small compared with the efficient price in congested urban areas, our results are not very
sensitive to the choice of $p_0$. Nevertheless, most developed countries make large government surpluses from road user taxes, so that it would be unrealistic to set $p_0$ to zero. (The United States is an exception (Small et al., 1989, Table 1.1).) In 1988 the British government's surplus from road taxes, excluding general taxes on road users and net of all road expenditure including police, was about 1.8p per vehicle-km (Department of Transport, 1989a, Tables 1.17 and 2.2). This is about 5 per cent of consumer surplus; we have therefore taken $p_0$ to be $0.05$ per pcu-km. We should note that at low traffic levels the efficient congestion price is less than the existing road user tax, so that a move to efficient pricing would imply a reduction in the existing tax. This may be unrealistic, but we have allowed the reduction to stand in the following analysis.

2.8 Homogeneous traffic
As mentioned in Section 1, we initially make the assumption that traffic is homogeneous. This means that, first, all units of traffic (pcu-kms) are presumed to have the same demand function with the same value of $\mu$, which means that they all have the same distribution of user valuations. All units of traffic also have the same congestion function, implying that they have the same value of $\beta$, and the same value of travel time. We do not assume that all vehicles are identical, but we do assume that the $\mu$'s and $\beta$'s for different types of vehicle are proportional to their pcu-values. In Section 5 we explore the consequences of relaxing this assumption.

3. ECONOMIC EFFECTS OF OPTIMAL CONGESTION PRICING WITH HOMOGENEOUS TRAFFIC
We now present results of calculations with the model discussed above. As mentioned previously, the main economic outputs of the urban road network are government surplus, $r$, consumer surplus, $s$, and the sum of these two surpluses, $b$. The main effects of introducing optimal road congestion pricing are measured by the changes in these quantities when the existing tax per pcu-km is replaced by the benefit-maximising congestion price.

The main results concerning the introduction of optimal pricing are presented in Tables 2 and 3, and in Figure 4. Table 2 and Figure 4 give results for the inverse linear congestion cost function, and Table 3 compares this with the other two functions. For the inverse linear function, Table 2 presents the initial values of $r$, $s$, and $b$, and the changes in $r$, $s$, and $b$ induced by the optimal congestion price, for a selection of initial average speeds. This table also gives the corresponding initial traffic level and the implied level of potential demand, $\alpha$. Figure 4 plots the changes in $r$, $s$, and $b$ against the initial average speed. Traffic levels are measured as percentages of the network capacity, $c$. The economic outputs are measured as percentages of $\mu c$, which is the upper limit to the consumer surplus, or, equivalently, as $\$ per 100 pcu-kms of capacity.
### TABLE 2
Effects of Optimal Congestion Pricing: Homogeneous Traffic and Inverse Linear Congestion Costs

| Initial Speed as % of Uncongested Speed | Initial Traffic % of capacity | Implied Potential Demand | Initial Gov't Surplus r | Initial Consumer Surplus s | Initial Economic Benefit b | Effect of Shift to Optimal Congestion Pricing on r | on s | on b |
|----------------------------------------|-------------------------------|--------------------------|-------------------------|----------------------------|-----------------------------|---------------------------------|------|------|------|
| 95                                     | 5                             | 5.4                      | 0.25                    | 5.0                        | 5.25                        | -0.1                            | +0.1 | +0.0 |      |
| 90                                     | 10                            | 11.0                     | 0.5                     | 10.0                       | 10.5                        | -0.0                            | +0.0 | +0.0 |      |
| 80                                     | 20                            | 23.2                     | 1.0                     | 20.0                       | 21.0                        | +1.2                            | -1.1 | +0.0 |      |
| 70                                     | 30                            | 37.4                     | 1.5                     | 30.0                       | 31.5                        | +3.7                            | -3.4 | +0.3 |      |
| 60                                     | 40                            | 54.9                     | 2.0                     | 40.0                       | 42.0                        | +8.0                            | -6.7 | +1.3 |      |
| 50                                     | 50                            | 78.4                     | 2.5                     | 50.0                       | 52.5                        | +14.4                           | -10.6 | +3.8 |      |
| 40                                     | 60                            | 114.9                    | 3.0                     | 60.0                       | 63.0                        | +24.3                           | -14.8 | +9.5 |      |
| 30                                     | 70                            | 187.3                    | 3.5                     | 70.0                       | 73.5                        | +41.2                           | -18.6 | +22.6 |      |
| 20                                     | 80                            | 416.6                    | 4.0                     | 80.0                       | 84.0                        | +77.5                           | -21.2 | +56.4 |      |
| 10                                     | 90                            | 3,463                    | 4.5                     | 90.0                       | 94.5                        | +203.8                          | -20.5 | +183.3 |      |

Source: based on economic model in text.

### TABLE 3
Effects of Optimal Congestion Pricing with Three Congestion Cost Functions

<table>
<thead>
<tr>
<th>Initial Speed as % of Uncongested Speed</th>
<th>Inverse Linear</th>
<th>Truncated Inverse Linear</th>
<th>Bottleneck Inverse Linear</th>
<th>Change in Consumer Surplus due to Congestion Pricing as a percentage of Initial Consumer Surplus</th>
<th>Change in Economic Benefit due to Congestion Pricing as a percentage of Initial Economic Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>+5.1</td>
<td>+5.1</td>
<td>0.0</td>
<td>+0.1</td>
<td>+0.1</td>
</tr>
<tr>
<td>90</td>
<td>+0.1</td>
<td>+0.1</td>
<td>0.0</td>
<td>+0.0</td>
<td>+0.0</td>
</tr>
<tr>
<td>80</td>
<td>-5.6</td>
<td>-5.6</td>
<td>0.0</td>
<td>+0.2</td>
<td>+0.2</td>
</tr>
<tr>
<td>70</td>
<td>-11.4</td>
<td>-11.4</td>
<td>0.0</td>
<td>+1.1</td>
<td>+1.1</td>
</tr>
<tr>
<td>60</td>
<td>-16.7</td>
<td>-16.7</td>
<td>0.0</td>
<td>+3.1</td>
<td>+3.1</td>
</tr>
<tr>
<td>50</td>
<td>-21.2</td>
<td>-21.2</td>
<td>0.0</td>
<td>+7.2</td>
<td>+7.2</td>
</tr>
<tr>
<td>40</td>
<td>-24.6</td>
<td>-14.9</td>
<td>0.0</td>
<td>+15.0</td>
<td>+22.8</td>
</tr>
<tr>
<td>30</td>
<td>-26.6</td>
<td>-5.4</td>
<td>0.0</td>
<td>+30.7</td>
<td>+51.3</td>
</tr>
<tr>
<td>20</td>
<td>-26.5</td>
<td>0.0</td>
<td>0.0</td>
<td>+67.1</td>
<td>+114.3</td>
</tr>
<tr>
<td>10</td>
<td>-22.7</td>
<td>0.0</td>
<td>0.0</td>
<td>+194.0</td>
<td>+304.8</td>
</tr>
</tbody>
</table>

Source: based on economic model in text.
3.1 Benefits of congestion pricing

Inverse linear congestion cost function
The final column of Table 2 and the central curve of Figure 4 show the changes in economic benefit induced by congestion pricing for the inverse linear congestion cost function. The results show that the increases in $b$ are small compared with all the other economic quantities and changes in economic quantities in the system, for initial speeds down to about one-third of the uncongested speed. For lower initial speeds, the change in benefit increases rapidly. The range of initial speeds for which the change in benefit is small includes much of the serious congestion likely to be encountered in practice, except that caused by untoward events such as accidents. Even daytime and peak hour traffic in central London falls within this range; it moves at 11 miles per hour (Department of Transport, 1989d, Table 5).

Other congestion cost functions
The last three columns of Table 4 compare the increase in benefit induced by optimal pricing for the inverse linear with that for the other two congestion cost functions. The increase in benefit is now expressed as a percentage of the initial benefit. The functions
are compared on the basis of the same initial average speed. This seems the most natural
and revealing comparison, because average speed is the most observable characteristic of
congested traffic, but we must note that this implies different initial traffic levels, and
different values of potential demand $\alpha$. Other comparisons would give different results.
For example, we could compare the congestion functions on the basis of the same initial
traffic level (corresponding to the vertical lines in Figure 1), in which case the inverse
linear function would give the greatest benefit and the bottleneck the smallest, rather than
the reverse as shown in Table 4.

Table 4 shows that, starting with the same average speed, optimal congestion pricing
gives the same percentage benefit for the inverse linear and the truncated inverse linear
functions, for all initial speeds down to half the free speed, which is the cut-off speed in
the truncated case. It is clear that, given the same other parameters, all truncated inverse
linear functions will give the same results when normalised in this way, down to the cut-
off speed. This confirms the intuition that the truncation is irrelevant, provided that the
demand curve initially cuts the congestion function to the left of the cut-off speed.
However, the benefit in the truncated case rises rapidly once the cut-off speed is passed.
This is because the system then begins to act like a bottleneck. The bottleneck gives the
greatest benefit from congestion pricing, whenever the initial speed is less than the free
speed. This is because in a bottleneck with homogeneous traffic optimal congestion
pricing induces no reduction in throughput and no change whatever in road use. It simply
replaces wasted time with an equivalent charge. (This also applies in the time-dependent
case: see Arnott et al., 1990.)

3.2 Changes in consumer surplus

Inverse linear congestion cost function
The penultimate column of Table 2 and the bottom curve of Figure 2 show that for the
inverse linear congestion cost function the change in consumer surplus, $s$, is negative for
all initial speeds below about 90 per cent of the uncongested speed. (Above that speed the
efficient congestion price is lower than the initial tax rate.) This implies that, with
homogeneous traffic, congestion pricing makes road users worse off, because some are
priced off the road, and, for those who stay, the extra congestion price is only partially
offset by the value of the reduction in travel time. This is a well-established property of
homogeneous traffic with this type of congestion cost function (Glazer, 1981). Table 3
shows that the loss in consumer surplus is in the range 15-30 per cent over a wide range
of initial speeds.

Other congestion cost functions
The first three columns of Table 4 compare the reduction in consumer surplus induced by
optimal pricing for the inverse linear with that for the other two congestion cost functions.
This table shows that the truncated linear again gives the same results as the full linear for
all initial speeds down to the cut-off speed; below that speed the reduction rapidly falls to
zero, as the system begins to behave like a bottleneck. In a bottleneck optimal congestion

230
ROAD CONGESTION PRICING: WHEN IS IT A GOOD POLICY?  A. W. Evans

pricing produces no losses for road users, whatever the initial speed. This is again because it simply substitutes a charge for wasted time, between which road users are indifferent.

It is clear from these results that the economics of bottlenecks are different from the economics of congestion cost functions based on downward-sloping speed-throughput relationships. If a congestion cost function consists of a combination of these, such as the truncated inverse linear, it is important to know in which segment the demand curve cuts the congestion function.

3.3 Horizontal equity

The first objection to congestion pricing was that it may redistribute welfare from road users to the rest of the community, and introduce horizontal inequity between users who are subject to it and those who are not. It is clear immediately that this objection does not apply to bottlenecks, at least for homogeneous traffic, because road users lose nothing. However, it does apply to the inverse linear and to the linear part of the truncated inverse linear congestion cost functions. We argue as follows.

Without road pricing, road users in congested (or more congested) conditions already get a worse service from the road system than those in uncongested (or less congested) conditions for the same tax price, because their journeys are slower. This could be regarded as horizontally inequitable, but it is accepted because it is obvious that little can be done about it. However, congestion pricing gratuitously compounds the inequity by making road users in congested conditions even worse off, to the benefit only of other people. This is what makes congestion pricing horizontally inequitable. Who the beneficiaries are depends on how the revenue is used, but it is clear the beneficiaries would not be the same as the payers; if they were, there would be no point in pricing.

The argument would be different if the revenue were to be used to expand the road system, as assumed in the long-run pricing studies of freeways in the United States (Keeler and Small, 1977), and underlying part of the argument in the Brookings Report (Small et al., 1989). Using the revenue to expand the road system would ensure, first, that the revenue benefited those who paid, and, secondly, that congestion was reduced, and therefore so was the congestion charge. It is for this reason that peak period charging gives rise to smaller equity problems if it is applied optimally to networks that are expandable as well as congestible, such as telephone networks. However, we are assuming in this paper that the road network capacity is fixed, which is realistic for older, high-density, cities.

Whether this horizontal inequity is acceptable depends partly on the scale of the benefits. If the net benefits were large in relation to the loss in road users’ consumer surplus, the inequity to users might be regarded pragmatically as an acceptable price to pay for these benefits. However, we saw in Table 2 and Figure 4 that for initial speeds down to about one-third of the uncongested speed, the benefits are not large relative to the loss in consumer surplus. It follows that at these traffic levels the major effect of congestion pricing would be to redistribute welfare from the affected road users to the rest of the community, with only a relatively small additional net benefit. It follows that in this
situation the case for congestion pricing depends on a judgement about the balance to be struck between redistribution and net benefit.

3.4 Data for London

How do data for London compare with the theoretical calculations above? The London Planning Advisory Committee's estimates of the effects of road congestion pricing in London, published in *Paying for Progress* (Chartered Institute of Transport, 1990, p.31), are that the benefit would be at least £400 million per year and the revenue at least £600 million per year, suggesting a net loss to road users of about £200 million per year.

First, is the increase in benefit of £400 million larger or smaller than is suggested by the calculations? The obvious comparison is with initial total benefit of the London road system. We may estimate from Department of Transport statistics (1989a, Table 2.2, and 1989d, p.3) that there were a little over 30 billion pcu-kms in 1988. If the estimate above of 0.33p for the average consumer surplus per pcu-km is of the correct order of magnitude, the total consumer surplus in London in 1988 was a little over £10 billion, to which we should add a small sum for government surplus. The £400 million increase is about 4 per cent of the initial total benefit. This is broadly in line with the figures for the mid-speed ranges of the linear and truncated linear functions in Table 3.

Secondly, how does the net benefit of £400 million compare with the net loss to road users of £200 million? This seems to be a better outcome than is suggested by the figures for the inverse linear congestion cost function in the last two columns of Table 2. This may be because a bottleneck model may be more appropriate in some circumstances, and this gives a much better outcome than the inverse linear model. In addition, our calculations so far have assumed homogeneous traffic, and we shall see in Section 5 that mixed traffic also gives a better bargain. Nevertheless, the net loss of £200 million to the London road users is an important matter. It is sometimes suggested that equity in congestion pricing can be achieved by spending the revenue in such a way that no income group is disadvantaged. This is not a solution, because horizontal equity is an individual, not a group, matter. It is no answer to an individual loser to know that other people in the same income group are gainers: so the loss should be avoided if possible.

4. GOVERNMENT INCENTIVES

Governments are monopoly suppliers of roads. The absence of a direct charging mechanism currently limits the scope for exploiting this monopoly, even if governments wished to do so. However, congestion pricing would remove this limitation, and give governments the power to raise a substantial amount of revenue. The traditional arguments for road congestion pricing assume that governments would not exploit this monopoly power, or in other words that they would not respond to those same incentives presumed to motivate road users and other economic agents. This assumption has been seriously questioned, particularly by the "public choice" school in economics. This
section uses the model to explore these incentives and their effects. Governments would be faced with the two incentives common to monopolists: to overcharge and to undersupply.

4.1 Pricing

Table 4 presents the results of assuming, pessimistically, that the government sets prices to maximise revenue for each of the congestion cost functions. It shows the percentage changes in the consumer surplus and the economic benefit compared with the initial fixed-tax position. The purpose is to give an indication of the nature and direction of the effects if a government responded to its incentive to raise prices. The most important conclusion is that the government incentive problem affects not only the linear and truncated linear congestion cost functions, but also the bottleneck. We might therefore infer that it affects all congestion cost functions, or at least all that do not bend back. Not surprisingly, Table 4 shows that, if the government exploited its monopoly to the full, road users would be much worse off than under efficient pricing, and that the total benefits might be less than with no pricing at all.

However, it is not reasonable to suppose that the government would act as a thorough-going monopolist, so this outcome is unduly pessimistic. In practice, the danger is that the congestion price would be set neither optimally nor monopolistically; it would be set without reference to the economics of the roads at all. Instead, it would be set according to revenue needs of the recipients. The revenue would either be absorbed in the government’s general tax revenue, or it would be hypothecated to specific purpose(s), such as public transport support. In the former case, the government’s budget would determine the price, and in the latter case the public transport budget would. The incentive to monopoly exploitation would arise from the fact that it would often be easier to put up the price than cut the budget of these organisations. To find an illustration of this process, we need look no further than the history of the Road Fund in Britain (Plowden, 1973).

It is sometimes suggested (for example in Paying for Progress, paragraph 11.6) that an advantage of road pricing is that it raises revenue, which could properly be used for road or other transport purposes. However, in Britain and many other countries, there is already no shortage of such revenue. The British government’s surplus on transport taxes, excluding general taxes on the sector and net of all transport public expenditure, is about £5,000 million per year (Department of Transport, 1989a). This could properly be used for transport purposes, but governments have chosen to use it for other purposes. Adding to this revenue would therefore relax a non-binding constraint, and unless the institutional arrangements were changed, there is no obvious reason why congestion pricing would lead to greater transport public expenditure.

4.2 Provision of capacity

If the congestion price were optimal, the government would have an incentive to undersupply road capacity for a large range of demand. The obvious measure of this is the elasticity of government revenue, \( r \), with respect to capacity, \( c \), that is, the percentage
change in government surplus for a 1 per cent change in network capacity. We have calculated this for fixed and optimal prices, and for the different congestion cost functions. With fixed prices the elasticity is always positive, implying that increasing capacity always generates revenue (though of course not necessarily enough to pay for the capacity), whereas with optimal prices the elasticity is negative for initial speeds down to about one-third of the uncongested speed. In this range increasing capacity reduces revenue, because it leads to a reduced optimal congestion price, which outweighs the increase in traffic. This perversely gives the government a financial incentive to promote congestion.

Again, it would be unduly pessimistic to suppose that the government would close roads for the purpose of increasing revenue; it would also be pessimistic to suppose that the government would refrain from increasing capacity purely because that would threaten revenue. On the other hand, if the government did increase capacity, say by installing a new and expensive traffic control system, it seems unlikely that it would consequently reduce the congestion price. On the contrary, it might well argue that since the road users are the beneficiaries of the control system, the price they pay should be raised to help meet its cost. This is another reason why we could expect the congestion price to be arbitrary.
4.3 Regulation
The usual defence against exploitation by monopolies is a regulatory body to control prices and services. There would be a strong case for a road pricing regulator if congestion pricing were introduced.

5. MIXED TRAFFIC

5.1 Characteristics of Mixed Traffic
In this section we consider how the results are affected if the traffic stream is mixed instead of homogeneous. We have confined our work to the case of the inverse linear congestion cost function, but there is no reason to expect that the conclusions about relative effects of congestion pricing on mixed traffic would be different for the other functions.

The obvious manner in which the traffic stream may be mixed is that different components may have different values of time, or, in the terms of our model, different values of $\beta$. Table 1 shows that there are large systematic variations in the average value of time per pcu-km between different classes of vehicle, and individual vehicles will have even larger variations. It has long been recognised (for example, Richardson, 1974; Layard, 1977) that congestion pricing systematically discriminates in favour of pcu-kms...
with high values of time, and that this may lead to different results from the homogeneous case. What is also important, but less explicitly discussed, is whether or not the variations in the value of time are correlated with variations in the user valuations of pcu-kms.

To explore these matters we suppose in this section that the traffic stream is composed of two parts, which may differ in either or both of (1) their values of time, and (2) the average valuation of their pcu-kms. The values of time are represented in an extended version of our model by two values of the parameter $\beta$, $\beta_1$ and $\beta_2$. To represent the different average valuation of pcu-kms, we suppose that each part of the traffic stream has a separate exponential demand function, each having a different value of $\mu$, $\mu_1$ and $\mu_2$. The total demand function is then the sum of these two exponential demand functions, as illustrated in Figure 5. The total demand is no longer exponential, unless $\mu_1 = \mu_2$. In general, the new demand function has more high-value, more low-value, and fewer medium-value pcu-kms than does the exponential, and therefore a greater variance. The mathematics of the extended model are more complicated, but still sufficiently tractable to allow us to explore the main effects of congestion pricing with mixed traffic.

In the illustrative calculations presented below, where the parameters $\beta$ or $\mu$ differ, we have taken the higher value to be five times the lower, and we have supposed that 75 per cent of the potential pcu-kms have the lower value(s) and 25 per cent have the higher. These figures were inspired by the variations in the average values of time between the vehicle classes in Table 1. Finally, we have taken the average values of the $\beta$'s and $\mu$'s to be the same as previously. This leads to the following specific values for the parameters, which are all measured in $ per pcu-km in the unspecified currency in which the overall average valuation per pcu-km, $\mu$, is $1; \mu = $1.00 (by definition); $\mu_1 = $0.50; $\mu_2 = $2.50; $\beta = $0.40; $\beta_1 = $0.20; and $\beta_2 = $1.00. The demand functions in Figure 5 have these values.

We consider the effect of congestion pricing on four different traffic streams each of which differs in composition while having the same average values of the parameters $\beta$ and $\mu$. These are:

(i) Homogeneous $\beta$; homogeneous $\mu$. This is the homogeneous or "standard" case, discussed above.
(ii) Homogeneous $\beta$; mixed $\mu$.
(iii) Mixed $\beta$; homogeneous $\mu$.
(iv) Mixed $\beta$; mixed $\mu$.

In case (iv) we assume that the pcu-kms with the higher $\beta$ also have the higher $\mu$, thus making the ratio $\beta/\mu$ the same for all pcu-kms. This is a mathematically convenient representation of a situation where there is a positive correlation between the values of time and the user valuations of pcu-kms.

5.2 Effects of congestion pricing with different traffic compositions

Table 5 gives the initial consumer surplus, $s$, the initial economic benefit, $b$, and the changes in these quantities due to congestion pricing for each of these traffic compositions and for a selection of levels of potential demand, $\alpha$, ranging up to twenty times capacity. Because composition (i) is the homogeneous case discussed in the previous section, in
TABLE 5
Results for Mixed Traffic: Inverse Linear Congestion Costs

<table>
<thead>
<tr>
<th>Potential Demand (as % of Capacity)</th>
<th>Characteristics of the Traffic</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Homog. Demand:</td>
<td>Mixed Demand:</td>
<td>Homog. Demand:</td>
<td>Mixed Demand:</td>
</tr>
<tr>
<td></td>
<td>Single Value of Time</td>
<td>Single Value of Time</td>
<td>Mixed Value of Time</td>
<td>Mixed Value of Time</td>
</tr>
<tr>
<td>10</td>
<td>9.2</td>
<td>9.2</td>
<td>9.1</td>
<td>9.2</td>
</tr>
<tr>
<td>20</td>
<td>17.5</td>
<td>17.7</td>
<td>17.5</td>
<td>17.5</td>
</tr>
<tr>
<td>50</td>
<td>37.4</td>
<td>39.5</td>
<td>38.0</td>
<td>37.8</td>
</tr>
<tr>
<td>100</td>
<td>56.5</td>
<td>66.9</td>
<td>59.1</td>
<td>57.5</td>
</tr>
<tr>
<td>200</td>
<td>71.1</td>
<td>105.0</td>
<td>76.4</td>
<td>72.8</td>
</tr>
<tr>
<td>500</td>
<td>81.5</td>
<td>170.3</td>
<td>87.5</td>
<td>83.8</td>
</tr>
<tr>
<td>1000</td>
<td>85.7</td>
<td>211.2</td>
<td>91.1</td>
<td>88.2</td>
</tr>
<tr>
<td>2000</td>
<td>88.5</td>
<td>227.7</td>
<td>93.2</td>
<td>91.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Initial Consumer Surplus, s</th>
<th>Initial Economic Benefit, b</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>9.6</td>
</tr>
<tr>
<td>20</td>
<td>18.3</td>
</tr>
<tr>
<td>50</td>
<td>39.3</td>
</tr>
<tr>
<td>100</td>
<td>59.4</td>
</tr>
<tr>
<td>200</td>
<td>74.7</td>
</tr>
<tr>
<td>500</td>
<td>85.6</td>
</tr>
<tr>
<td>1000</td>
<td>90.0</td>
</tr>
<tr>
<td>2000</td>
<td>92.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Consumer Surplus due to Congestion Pricing</th>
<th>Change in Economic Benefit due to Congestion Pricing</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>+0.0</td>
</tr>
<tr>
<td>20</td>
<td>-0.7</td>
</tr>
<tr>
<td>50</td>
<td>-5.8</td>
</tr>
<tr>
<td>100</td>
<td>-13.3</td>
</tr>
<tr>
<td>200</td>
<td>-19.0</td>
</tr>
<tr>
<td>500</td>
<td>-21.4</td>
</tr>
<tr>
<td>1000</td>
<td>-21.4</td>
</tr>
<tr>
<td>2000</td>
<td>-21.0</td>
</tr>
</tbody>
</table>

Units: the economic figures in this table are measured as percentages of the upper limit to consumer surplus with homogeneous traffic (case (i)), or, equivalently, as $ per 100 pcu-km of capacity.

Source: based on economic model in text.

Table 5 we are particularly interested in results for compositions (ii), (iii) or (iv) that differ substantially from the corresponding results for composition (i). Such differences indicate that the composition of the traffic is having a big effect. Two such differences stand out. The first is that at moderate and high levels of potential demand the initial consumer surplus, $s$, in composition (ii) is much higher than in the homogeneous case (i); this difference is also carried over into economic benefit, $b$. The second difference is that at moderate levels of potential demand the loss in consumer surplus due to congestion pricing in composition (iv) is much less than in composition (i), and at high demand levels there is actually a gain in consumer surplus. We now discuss these results.
Composition (ii) has the compound demand function shown in Figure 5, and a single value of time for all pcu-kms. We saw in Section 2 that the exponential demand function has the property that at any level of demand the average consumer surplus per pcu-km is $\mu$. It follows here that at any level of congestion the average consumer surplus per pcu-km of the lower-valued traffic in the stream is $\mu_1$, or $0.50$, and of the higher-valued traffic is $\mu_2$, or $2.50$. The overall average consumer surplus per pcu-km is the average of these two figures, weighted by the proportion of each type in the traffic stream. It is clear from Figure 5 that as congestion, and therefore generalised cost, increases, the proportion of higher-valued traffic in the stream increases, eventually dominating. At higher levels of congestion, the demand function therefore behaves like an exponential with parameter $\mu_2$. It follows that the previous upper limit of $\mu c$ on the total consumer surplus from the road is now replaced by $\mu c_c$, which in our figures is two and a half times greater. All this happens without the aid of congestion pricing. The introduction of congestion pricing in this case has no notable effect over and above the effect we would expect from the homogeneous case. (It is true that the absolute reduction in consumer surplus due to congestion pricing is higher in composition (ii) than the others, but the proportional reduction is not.) Is traffic composition (ii) a realistic possibility? In other words, is there a class of traffic with highly-valued pcu-kms but with fairly low values of time? Table 1 suggests that there is: “other goods vehicles”. These have low values of time per pcu-km, but we may guess that they do highly valued work, because there are few substitution possibilities, especially in urban areas. We may therefore think of traffic composition (ii) as comprising a mixture of ordinary cars and “other” (that is, the heavier) goods vehicles. The analysis above suggests that the effect of congestion pricing in this case would be similar to its effect on a homogeneous traffic stream with the same average properties.

Composition (iv) has the same compound demand function as composition (ii), but now the higher-valued traffic stream also has a proportionately higher value of time. Table 5 shows that without congestion pricing the consumer surplus in this case is close to that of the homogeneous case (i). This shows that, in contrast to composition (ii), the higher-valued traffic does not come to dominate the traffic stream at higher levels of congestion. Why not? The reason is that, because of its higher value of time, congestion affects this traffic more seriously than the other traffic. Therefore, the tendency for the high-valued traffic to dominate is countered. In this particular example, the counterbalancing is exact, and the proportion of high-valued traffic remains constant at all levels of congestion. This is a consequence of the specific assumption that both components of the traffic have the same values of $\beta/\mu$.

However, congestion pricing now changes things. It discriminates in favour of traffic with the higher value of time, because it saves time, which is valued more highly by this traffic, in return for a price which is the same for all. It may therefore actually reduce the generalised cost per pcu-km for the traffic with the higher value of time. If this traffic also has a higher consumer surplus, as in (iv), congestion pricing promotes the substitution of traffic with a high consumer surplus for traffic with a lower one. This substitution process, which happens of its own accord in (ii), happens only with the help of congestion pricing in (iv). This gives congestion pricing a role that is absent in the homogeneous case, and
it makes congestion pricing a better buy. Its benefits are greater and the adverse effect on consumer surplus is less than in the homogeneous case. This conclusion is not new: for example, Glazer (1981) gives an example where congestion pricing leads to an increase in consumer surplus by this mechanism. We return to composition (iv) after a brief discussion of (iii).

Composition (iii) has a homogeneous demand function, but two different values of time. Table 5 shows that this performs like the homogeneous case (i); it does not produce extra benefits like (iv). Why not? Congestion pricing still discriminates in favour of the traffic with the higher value of time, but in this case there is no point in doing so, and in any case it is a lost cause. There is no point because the favoured traffic is not valued any more highly than the unfavoured: therefore there is no gain in consumer surplus from the substitution. It is a lost cause because, with or without congestion pricing, the traffic with the high value of time cannot compete with the other traffic: it has higher costs but is not specially valued. Therefore as congestion increases this traffic forms a decreasing minority of the stream. Congestion pricing does not overcome this.

The conclusion is that mixed traffic makes a big difference to the effect of congestion pricing only in case (iv), where there is a component of traffic which is valued more highly than average, and also has a higher value of time. Is there a class of traffic which fits this bill? Clearly, yes: high-occupancy vehicles, particularly public transport vehicles, where the large number of people per pcu leads both to highly valued pcu-kms and high values of time per pcu-km. “Working cars” and light goods vehicles may also be in this category. In contrast to the usual situation where road users are harmed by congestion pricing, this class of road users might gain. We now use the results of the model to consider how our previous objections to congestion pricing are modified in this case.

5.3 Congestion pricing when some traffic is highly valued and has a high value of time

The first group of columns in Table 6 shows the percentage changes induced by optimal congestion pricing in economic benefit, and in the consumer surplus for each of two components of the traffic stream and all traffic.

The first objection was that congestion pricing is horizontally inequitable, because it makes the affected road users worse off, to the benefit only of other people. The indicator of this in Tables 2 and 3 was the fall in consumer surplus due to congestion pricing, typically in the range 15-30 per cent. Table 6 shows that with mixed traffic this objection is weaker, but also more complex. If road users are treated as a single group, their loss in consumer surplus is much smaller, at most about 13 per cent; at high levels of congestion road users actually gain in welfare. However, the pattern is different for the two groups of users. The users of the lower-valued traffic are much worse off than in the homogeneous case: their loss in consumer surplus increases without limit until it approaches 100 per cent at high levels of congestion, when they are priced off the road altogether. However, the users of the higher-valued traffic are better off than in the homogeneous case: at lower levels of congestion, their welfare is little changed by congestion pricing, as the price is roughly balanced by the value of time they save; at higher levels of congestion the price is less than the value of time saved, and they gain in welfare.
TABLE 6

<table>
<thead>
<tr>
<th>Initial Speed as % of Uncongested Speed</th>
<th>Percentage Changes due to Optimal Congestion Pricing in</th>
<th>Percentage Changes due to Traffic Segregation in</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Consumer Surplus</td>
<td>Economic Benefit</td>
</tr>
<tr>
<td></td>
<td>Lower Valued Traffic</td>
<td>Higher Valued Traffic</td>
</tr>
<tr>
<td></td>
<td>All Traffic</td>
<td>All Traffic</td>
</tr>
<tr>
<td>90</td>
<td>-0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>80</td>
<td>-11.6</td>
<td>-5.1</td>
</tr>
<tr>
<td>70</td>
<td>-22.7</td>
<td>-9.4</td>
</tr>
<tr>
<td>60</td>
<td>-33.1</td>
<td>-12.1</td>
</tr>
<tr>
<td>50</td>
<td>-42.8</td>
<td>-12.8</td>
</tr>
<tr>
<td>40</td>
<td>-52.7</td>
<td>-10.1</td>
</tr>
<tr>
<td>30</td>
<td>-64.6</td>
<td>-0.9</td>
</tr>
<tr>
<td>20</td>
<td>-83.7</td>
<td>+24.6</td>
</tr>
<tr>
<td>10</td>
<td>-99.9</td>
<td>+166.6</td>
</tr>
</tbody>
</table>

Source: based on economic model in text.

The adverse government incentives are still present with mixed traffic, so the main additional effect of congestion pricing in the mixed over the homogeneous case is that it promotes the substitution of high-valued/high time value traffic for lower-valued/low time value traffic, which does not otherwise happen of its own accord. This works to the additional disadvantage of the lower-valued traffic, but it does produce welfare gains through the more efficient use of road space. Among the gainers are bus operators and users, and this in turn could lead to second-round benefits, as the bus system improves, and some car users could start using public transport. Although this additional effect is not without losers, it seems reasonable to judge its effect as favourable. The conclusion is that the equity objection to congestion pricing is weaker if the traffic stream has a sizeable bus component, or more generally a sizeable high-occupancy vehicle component.

5.4 Traffic segregation

However, if we desire the substitution of high-occupancy for low-occupancy vehicles, we must ask whether it could be achieved in ways other than by congestion pricing. Obviously it can: by bus priority measures, along the lines recommended by the Bus and
ROAD CONGESTION PRICING: WHEN IS IT A GOOD POLICY? A. W. Evans

Coach Council (1990), or by reserving part of the road capacity for buses or other high-occupancy vehicles. This too is not new. A detailed study by Small (1983), concerned with peak-hour bus priority on expressways and using a time-dependent bottleneck model, concluded that:

"under idealised conditions in which lane capacities can be allocated ... without waste, bus priority ... achieves roughly half ... the benefits of marginal cost pricing. The distribution of benefits is quite different, however" (p. 68).

We may use the different model of this paper to reach surprisingly similar conclusions. In both Small’s case and our own, any actual segregation scheme will of course fall short of idealised conditions. Suppose we start with the same mixed traffic situation as above, that is, with 25 per cent of potential demand of the higher-value type, and 75 per cent of the lower-value. Instead of introducing congestion pricing, we segregate the two types of traffic, and allocate a part of the road capacity to each. In order to achieve the desired substitution of high-valued traffic for low-valued traffic, we should allocate capacity disproportionately to the former, which is just what bus priority measures do. Suppose we arbitrarily allocate half the capacity to the higher-valued traffic and half to the lower-valued traffic, and suppose that each type then operates independently within its own road space.

The right hand group of columns of Table 6 shows the economic effects in the same form as those of congestion pricing. The increase in economic benefit due to segregation is about half that due to congestion pricing over much of the range of initial speeds. However, both types of road user are better off with segregation than with congestion pricing over most of the demand range. This is simply because with segregation neither group has to pay congestion prices to a third party. The lower-valued traffic retains much economic welfare even at high congestion levels, instead of being priced off altogether. It is true that the lower-valued traffic is worse off with segregation than in the initial position, but that is inevitable if it is desired to reap the benefits of promoting high-occupancy vehicles. Finally, because the government does not make any money out of congestion, it has no perverse incentives.

5.5 Conclusion on Mixed Traffic

The conclusion of this section is that the most important difference between homogeneous and mixed traffic arises where there is highly-valued traffic that also has a high value of time. Public transport and other high-occupancy vehicles are the prime examples. The fact that this traffic is highly-valued means that it is economically efficient for it to use scarce road capacity at the expense of other traffic, but its high value of time prevents it doing so in congested conditions without help. Congestion pricing is one means of providing this help, and for this reason congestion pricing is a "better buy" in this situation than with homogeneous or other mixes of traffic. However, congestion pricing is not the only means of providing the required help. Comprehensive bus priority measures provide some of the benefits without the objections to congestion pricing.
6. CONCLUSIONS

(1) The effects of road congestion pricing are sensitive to the congestion cost function. Congestion pricing is a better buy for bottlenecks than for linear speed-throughput relationships.

(2) The scale of unwanted redistribution from road users to the rest of the community induced by optimal pricing depends on the congestion cost function. It is zero for bottlenecks with homogeneous traffic, but may be large relative to the net gain in benefit for linear speed-throughput relationships.

(3) In mixed traffic, if there is a correlation between the value of time of traffic units and the value of the journeys they are making, congestion pricing is a better buy than otherwise. This is because it discriminates in favour of the traffic with high values of time, and through the correlation promotes the substitution of high-value journeys for low-value journeys. Public transport and other high-occupancy vehicles would benefit.

(4) Nevertheless, it is possible to help high-occupancy vehicles in other ways, such as by the selective allocation of road space. It is possible that both high-occupancy and ordinary vehicle users would be better off with this than with pricing, because neither would have to pay charges to a third party.

(5) A pervasive problem for road congestion pricing is that it creates perverse incentives for governments. Governments would presumably not exploit their monopoly in road provision to the full, but they would have a strong incentive to depart from economic efficiency in setting prices, and substitute a revenue target. There would be a strong case for an independent body to regulate prices.

Date of receipt of final typescript: May 1992

REFERENCES


Chartered Institute of Transport (1990): *Paying for Progress*. Chartered Institute of Transport, 80 Portland Place, London W1 4DP.


ROAD CONGESTION PRICING: WHEN IS IT A GOOD POLICY? A. W. Evans