Commercial road supply with incentive regulation

Mark O. Harvey

Bureau of Infrastructure, Transport and Regional Economics, Canberra, Australia

Paper accepted for publication in the *International Journal of Sustainable Transportation* on 3 January 2013. First submitted 19 February 2012

Address for correspondence

Dr Mark Harvey

Bureau of Infrastructure, Transport and Regional Economics

GPO Box 594, Canberra, ACT, 2601, Australia

Tel: +61 2 6274 6720, +61 417 203 503

Mark.Harvey@infrastructure.gov.au

Abstract

Price regulation to limit monopoly charging by a commercial road supplier prevents full transfer from users to the supplier of marginal benefits from investing and maintaining to improve service quality. Marginal revenue from quality improvements falls short of marginal social benefit leading to under-provision. A form of incentive regulation is proposed that ensures full benefit transfer. Out of revenues raised from users, the regulator pays the supplier a shadow toll that varies with the level of service quality as measured by users' average generalised costs. Under the assumptions of the model, profit-maximising and economically efficient investment and maintenance outcomes align.

Keywords: incentive regulation, road commercialisation, road privatisation, service quality

1. Introduction

1.1 Commercial roads

The incentive the profit motive creates for technical efficiency and innovation is reason enough to consider commercial approaches to road supply, but the more pressing motivation is to overcome impediments to efficient supply and pricing of roads. Pressures to restrict government spending, borrowing and taxation in many countries have led to underfunding of investment and maintenance for roads. For example, the US federal fuel tax, which is hypothecated to the Highway Trust Fund, has not been increased in line with inflation since 1993, while vehicles have grown more fuel efficient and road construction and maintenance costs have risen faster than inflation (Geddes 2010). Roads compete for funds with other government-provided services. Revenues collected from road users are often diverted to other uses (Semmens 2006). Maintenance tends to be underfunded relative to construction. Deferring maintenance in the short term can be very expensive in the long term (Roth 2006; Semmens 2006; Zietlow 2006). While cost–benefit analysis plays a significant role in government decision making about roads, other considerations intervene. Spending scarce funds on uneconomic projects leaves less for economically-warranted projects, compounding the underfunding problem.

The case for pricing congested urban roads is well-established and the necessary technology has been available for some time. Surveys have shown that public acceptability of the idea of congestion pricing is low, though support increases after implementation. The theoretical literature has identified this as the main barrier to implementation (Schade 2003; Ubbels and Verhoef 2006, Zmud and Arce 2008). The public may be more ready to accept road pricing from a private than a public entity (Small and Verhoef 2007, p. 201).

In other network industries, such as electricity, gas, water and telecommunications, provision by public utilities or regulated private firms separates the activity from government

services funded by taxation. Investment and maintenance decisions are made on commercial grounds. Beneficiaries pay for the services they receive at prices related to costs of provision. Governments can still influence investment and pricing outcomes to promote social or equity objectives, but it is more transparent, for example, requiring explicit directives and subsidies.

To date, road commercialisation has been limited to newly constructed toll roads that are either major urban arterials or interurban highways — roads at the highest level in the hierarchical structure of road networks. There are no examples of widespread commercialisation of existing roads or of lower-level roads. Newbery (1994, pp. 238-9) suggests that the complexity of the legal and administrative arrangements that govern road supply could make private ownership subject to independent regulation impractical, but a public utility with commercial objectives might be possible if 'good incentives for efficient management and investment' can be put in place. We leave the issue the legal and administrative arrangements for elsewhere and focus on the incentives for efficient management and investment. Here, another important difference between roads and other network industries matters — the service quality dimension. For other network industries, regulation is primarily concerned with limiting monopoly pricing. Regulating service quality is of secondary importance because there is not a large range of different quality levels that are economically efficient and commercially viable under different circumstances. For individual roads, a very wide range of service qualities can be warranted under different circumstances, from an earth track to a multi-lane freeway. The economic problem for road commercialisation is to secure a good outcome in terms of both price and quality.

1.2 Approaches to commercialisation

The approaches to road commercialisation investigated in the economics literature to date concern toll roads with various mechanisms to move the price and capacity away from the profit-maximising monopoly outcome towards the welfare maximising outcome.

Competition between operators of different links in a network is one such mechanism. In some models, the suppliers choose price only, for example, De Palma and Lindsey (2000), and Small and Verhoef (2007, ch. 6). In others, capacity is a choice variable as well, for example, Xiao et al. (2007), Verhoef (2008), and Zhang et al. (2008). The models find that the greater the number of competitors supplying parallel links, the closer price and capacity approach perfectly competitive levels. However, as Verhoef (2008) notes, such models seem theoretical because the lumpiness in road infrastructure limits the number of competitors.

Auction of a toll franchise can ensure the winning bidder earns no more than a normal profit, maximising the surplus available to users and the government. The government can set the toll and award the franchise to the bidder offering the greatest capacity, or set capacity and select the bidder offering the lowest toll. Authors have examined other rules for selecting the winning bid that allow bidders to choose both the toll and capacity. Examples of rules found to perform well are to select the bidder offering to produce the maximum patronage (Verhoef 2007) or the least sum of generalised price (including the toll) faced by users plus the subsidy per user (Ubbels and Verhoef 2008).

Build-operate-transfer schemes offer further possibilities for regulating tolls and capacity supplied. For a new tolled link in a network of existing untolled links, Yang and Meng (2000 and 2002) examined the possible outcomes, including those with maximum welfare, maximum profit and zero profit, in toll–capacity space. Guo and Yang (2009) consider bilateral negotiation between the government and road supplier as well as auctions. Tan et al. (2010) identify the contract curve in capacity–quantity space between the social and monopoly optimums, which is Pareto-efficient in the sense that, for each point on the curve, neither social welfare nor profit can be increased without reducing the other. They demonstrate that price-cap, rate-of-return and minimum capacity regulation lead to outcomes off the contract curve. Demand regulation (determining price and capacity to achieve a

minimum demand level) and markup regulation (specifying a maximum allowable profit on each unit of output) are efficient.

Our incentive regulation scheme is a novel approach to commercial road supply suited to a large public utility or private firm that controls a substantial portion of a road system. It may be relevant to toll roads where, as Ubbels and Verhoef (2008) observe, penalties are needed to ensure the winning bidder delivers the toll and capacity promised. The scheme's applicability extends to uncongested roads, a major component of a road system largely neglected in the economics literature. The scheme could be seen as an extension of the performance-based contracts now being used for maintenance of roads and other infrastructure. Such contracts stipulate minimum *conditions* for assets instead of required works or services, and payment depends on how well the standards are met (Zeitlow 2006). In our scheme, price is regulated, while performance in providing service quality is measured in monetary, not physical terms. The rewards and penalties reflect the social values of over- or under-performance.

The idea of internalising part of road user costs in order to induce a supplier to make economically optimal decisions has been raised previously for road safety. Roth (1996, pp. 50-56) discusses transferring legal liability for the cost of crashes to a commercial supplier. This would create a financial incentive to provide safer roads, and, if the supplier has the necessary power, to set and enforce regulations and to exclude accident-prone users.

Section 2 of the paper sets out a typical model of the economics of a single road. Section 3 discusses some of the economic impediments to road commercialisation. Regulation of road user charges to prevent monopoly pricing is shown to lead to serious under-provision of service quality. Our incentive regulation scheme, explained in section 4, targets the underlying problem of failing to internalise gains to users from improvements in

service quality. Section 5 considers the scheme further, briefly addressing cost recovery and imperfect information, then showing how it can apply to maintenance and networks.

2. The welfare maximising road standard

It is assumed there is perfect information, road standard is perfectly divisible, and capital is malleable, which implies instantaneous adjustment. Such assumptions are common in theoretical models of road economics despite the existence of indivisibilities or lumpiness in investment in individual roads. Above the minimum width necessary for vehicles to pass, lane capacity can be varied by altering width and alignment. For full networks, the effects of indivisibilities or lumpiness for individual roads would be lessened by pooling costs, revenues and benefits across road segments. Pooling would also occur over time as demand grows (Verhoef and Mohring 2009).

Road users incur the 'generalised cost' of travel comprised of costs of vehicle operation, travel time, trip time variability (unreliability) and crash risk. We do not consider environmental externalities to avoid complicating the model. The average generalised cost for a vehicle travelling the length of a road segment depends on the volume of traffic, q, and the standard of infrastructure provided, x, that is, $c = c(q, x)$. The main aspect of infrastructure standard here is capacity, but it can include safety and factors that affect travel time in the absence of congestion such as alignment. Hence, the term 'road standard' is used throughout this paper rather than 'capacity'. Road standard is synonymous with 'service quality' until congestion occurs, which causes service quality to deteriorate given the physical infrastructure.

Average generalised cost rises with traffic volume above some minimum level at which road users start to slow each other down and congestion occurs ($\partial c/\partial q \geq 0$), and falls with capital invested as the road becomes wider and straighter, until the point is reached

where further improvements have no effect ($\partial c/\partial x \leq 0$). The inverse demand curve for road use is $p = p(q)$ where p is the generalised price equal to the sum of average generalised cost and a distance-related road user charge, τ, that is, $p = c + \tau$.

Generalised cost provides a measure of users' valuation of the service quality offered by the road. In the quality economics literature, the approach would be to express the demand curve as $\tau = \tau(q, a_1, a_2 \cdots, a_n)$ where the a's are quality attributes including vehicle operating cost, time, unreliability and crash risk. Being negative attributes, a fall in any of them shifts the demand curve upward, increasing users' willingness-to-pay (WTP) valuation of a given quantity. Each attribute varies with road standard and, when there is congestion, with quantity as well, which is one reason why the generalised cost specification is so much more convenient. The change in WTP from a unit change in attribute a_i for a given quantity level, q', is $-\partial WTP/\partial a_i = -\int_0^{q'} \partial\tau/\partial a_i \, dq = -q'v_i$ where v_i is the average value over all users of the marginal unit of attribute i. Under the generalised cost approach, $c = \sum a_i \, v_i$. The welfare gain from a unit change in a_i is estimated as $-q' \frac{\partial c}{\partial a_i} = -q' a_i \frac{dv_i}{da_i} - q' v_i$. For the generalised cost approach to correctly measure the welfare change, we require $dv_i/da_i = 0$, that is, the average value of the quality attribute does not change. In most cases, this is an acceptable assumption.

Road standard is a function of the amount of capital invested in the road. The annualised capital cost, $K = K(x)$, is comprised of the investment cost of the assets annuitised over the life of the assets at a given rate of return on capital.[1] The rate of return is

1. As road standard is multi-dimensional, it might be convenient to express x in dollars of expenditure. For any given level of expenditure, the particular combination of different road standard dimensions would be that which minimises users' average generalised cost. With x expressed in dollars, $dK/dx = 1$.

the normal rate that would be earned in a competitive industry with similar risk characteristics and is assumed to equal the social discount rate.

The social welfare function to be maximised is welfare (W) = road users' WTP for the traffic volume q' minus road users' costs (cq) minus the road supplier's annualised investment cost (K).

$$W = \int_0^{q'} p(q)dp - cq - K \tag{1}$$

The first-best optimal pricing and road standard conditions are found by partially differentiating, equation (1) with respect to q and x. To derive the optimal charge, $\hat{\tau}$,

$$\frac{\partial W}{\partial q} = p - c - q\frac{\partial c}{\partial q} = 0$$

Since $p = c + \tau$, the well-known result for the optimal congestion charge is obtained.

$$\hat{\tau} = q\frac{\partial c}{\partial q} \tag{2}$$

For the optimal road standard

$$\frac{\partial W}{\partial x} = -q\frac{\partial c}{\partial x} - \frac{dK}{dx} = 0 \text{ which implies } -q\frac{\partial c}{\partial x} = \frac{dK}{dx} \qquad \text{(3a and 3b)}$$

We return to equation (3b) after considering the case where a fixed user charge, τ, is set exogenously at a level that may not be optimal. The traffic volume, q, is then endogenous to the model, determined by the level of c via the demand curve.

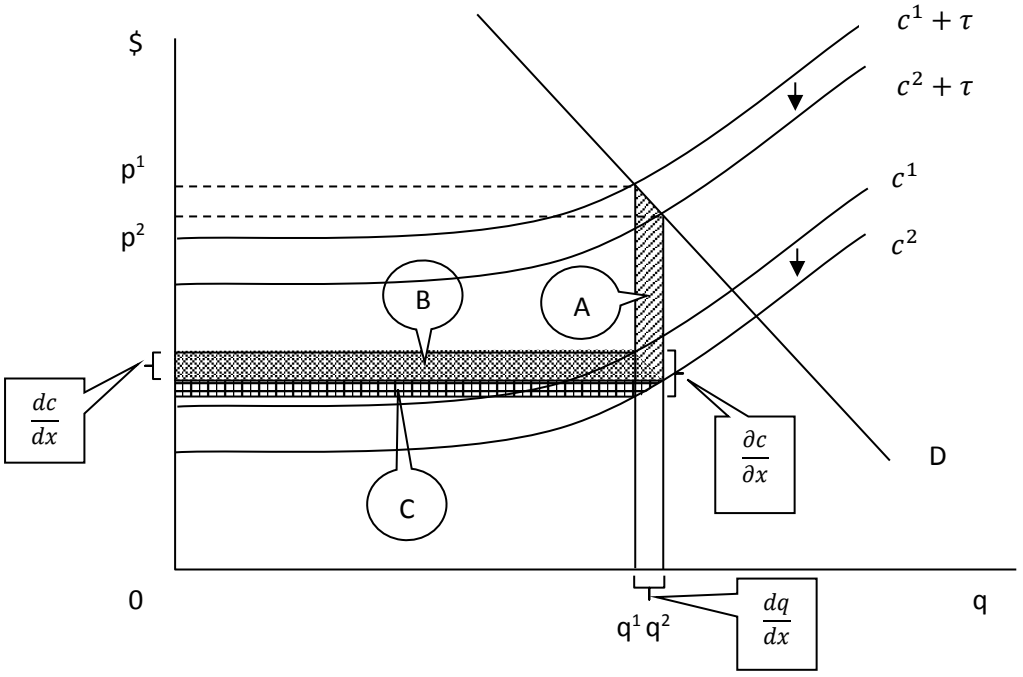
The welfare maximising road standard is found where,

$$\frac{dW}{dx} = p\frac{dq}{dx} - c\frac{dq}{dx} - q\frac{dc}{dx} - \frac{dK}{dx} = \tau\frac{dq}{dx} - q\frac{dc}{dx} - \frac{dK}{dx} = 0 \tag{4a}$$

$$\tau\frac{dq}{dx} - q\frac{dc}{dx} = \frac{dK}{dx} \tag{4b}$$

Figure 1 explains equation (4b) by showing the welfare effect of a small downward shift of the cost curve from c^1 to c^2 due to a unit increase in road standard, where there is a fixed user charge. The welfare gain from marginal generated traffic is area $A \approx \tau \frac{dq}{dx}$, the difference between price and average generalised cost for the increase in quantity. This benefit is passed on to whoever levies the charge. Area $B \approx -q \frac{dc}{dx}$ is the benefit to existing traffic from the fall in generalised costs. The marginal benefit to society is the sum of areas *A* and *B*.

Figure 1 Welfare changes from a small downward shift in the average generalised cost curve



With perfect divisibility, small increments in investment will be warranted as long as the marginal benefit exceeds the marginal cost of improving the road, *dK/dx*. With diminishing returns from additional investment, the optimal level of investment is found where the marginal benefit equals the marginal cost of improvement, at which point the

marginal benefit–cost ratio (MBCR) equals one, $\left(\tau \frac{dq}{dx} - q \frac{dc}{dx}\right) / \frac{dK}{dx} = 1$. The MBCR is the annual gain to society from investing an additional one dollar per annum in the road.

To show that equation (3b) is a special case of equation (4b) when the charge is optimal, we take the total differential of the user cost function, $dc = \frac{\partial c}{\partial q} dq + \frac{\partial c}{\partial x} dx$, multiply it by *q/dx,* and substitute the optimal price, equation (2).

$$q \frac{dc}{dx} = \hat{\tau} \frac{dq}{dx} + q \frac{\partial c}{\partial x} \tag{5}$$

When the user charge is set at the optimal level, that is, $\tau = \hat{\tau}$, the marginal benefit in equation (4b) is equal to the marginal benefit in equation (3b). The latter is the downward shift of the average cost curve from a unit increase in road standard times the (fixed) traffic volume. In figure 1, it is approximately the sum of areas *B* and *C*. Recalling that *dc/dx* and *∂c/∂x* are negative, when the user charge is set at the optimal level, we can write equation (5) as: –area *B* = area A – (areas *B* + *C*), which implies area *A* = area *C*. This conclusion can be deduced from the envelope theorem.

3. Some economic impediments to widespread road commercialisation

3.1 Limits on use of commercial investment criteria

On congested urban roads, the potential exists to use financial analysis rather than cost–benefit analysis to determine the optimal road standard because of approximate constant returns to scale. Individual roads are subject to economies of scale but for major urban roads, diseconomies of scale from intersections and high costs of expansion due to high urban density are offsetting. (See Small and Verhoef 2007, ch 3 for a literature survey.)

Constant returns to scale implies that optimal charges will cover capital costs. Using the same model as in the present paper, Mohring and Harwitz (1962, pp. 81-7) showed that,

for a single road with constant returns to scale, optimal pricing and investment in capacity lead to exact cost recovery.[2] Subsequent authors have shown that the 'Mohring–Harwitz theorem' holds more generally — growing traffic, heterogeneous users, time-varying demand and networks. (See Small and Verhoef (2007) and Verhoef and Mohring (2009) for literature surveys.)

If a regulator required a commercial supplier of congested urban roads to charge optimal prices, the optimal level of investment could, in principle, be found at the point where costs are exactly recovered. As Newbery (1994, p. 239) states, 'If revenue exceeds interest and maintenance costs, roads should be expanded. If revenue falls short of the interest and maintenance costs, traffic should be allowed to increase until congestion costs and hence the price charged rises to cover the costs'.

Most of the road system by length consists of uncongested or low-volume roads including many inter-city highways, and within urban areas, minor arterials and suburban streets. The optimal congestion price is zero or close to it most of the time. Walters (1968) argued that such roads approximate to being pure public goods. In transport, this argument dates back to Dupuit (1844). Walters attributes the public good nature of low-volume roads to a mix of economies of scale and indivisibilities. The basic two lane road is the principal indivisibility. He further argues that low-volume roads are subject to jointness in the supply

2. For constant returns to scale, average users' generalised cost as a function of traffic volume and road capacity must be homogeneous of degree zero (a proportional increase in traffic volume and capacity leaves average user cost unchanged), which would be the case if user cost was a function of volume–capacity ratio, *q/x*. Further, capital cost must be proportional to capacity. By Euler's theorem, homogeneity of degree zero in the user cost function implies $q\frac{\partial c}{\partial q} + x\frac{\partial c}{\partial x} = 0$. Substituting in the optimal price and investment conditions, equations (2) and (3b) respectively, and letting $K = kx$, total revenue equals total capital cost, $\hat{\tau}q = kx$. Verhoef (2008) notes that the Mohring–Harwitz result is more general than this. The ratio of revenue to capacity cost, $\hat{\tau}q/K$, equals the elasticity of the capital cost function with respect to capacity.

of capacity and service quality. Supply of one automatically implies the other is available. Investment to improve road quality by building a wider, smoother, straighter road with more passing opportunities is often found to be economically warranted based on savings in time, vehicle operating costs, and crash costs to road users. However, these improvements also add to capacity, keeping the optimal congestion price to practically zero.

For uncongested roads, the pure economic approach is to charge zero variable prices (other than to trucks for pavement damage) and to recover the deficit from general taxation or a land tax that does not affect resource allocation. Local roads, funded by rates levied on properties by local governments, are funded by land taxes. Recovery of costs from road users will result in welfare losses, but these can be minimised through a combination of access and variable charges that differ between user classes reflecting different abilities to pay (Ramsey pricing). Fuel taxes are a form of variable user charge. As well as being related to distance travelled, they bear some relationship to ability to pay because larger vehicles with higher operating and capital costs pay more in absolute terms. However, fuel consumption is also affected by other factors unconnected with ability to pay.

Given the demand curve and the user cost and capital cost relationships for a road, there is a range of price–standard combinations for which revenue from variable charges equals capital and maintenance costs. Within the feasible range, one can choose an arbitrary amount for the variable user charge and then provide the best possible road out of the available revenue. For a congested road with constant costs, the welfare maximising price–standard combination is the one with the optimal congestion charge for the standard. For an uncongested road, the only way to tell is to use cost–benefit analysis.

3.2 Monopoly price and investment outcomes

Once the optimal standard is known, whether through comparing revenues with costs in the

case of optimal congestion pricing with constant costs or through cost–benefit analysis, there is the problem of inducing the supplier to invest to that standard.

Other authors have derived the profit maximising conditions for commercial supply of a congested road where both the user charge and capacity are unregulated (for example, Verhoef 2007, Small and Verhoef 2007). In terms of our model, the monopolist's profit function is

$$\pi = \tau q - K = p(q)q - cq - K \tag{6}$$

Differentiating with respect to q to find the profit maximising user charge, τ_m, and substituting $\tau_m = p - c$ and equation (2)

$$\frac{\partial \pi}{\partial q} = q\frac{dp}{dq} + p - q\frac{\partial c}{\partial q} - \mathrm{c} = 0$$

$$\tau_m = -q\frac{dp}{dq} + q\frac{\partial c}{\partial q} = -q\frac{dp}{dq} + \hat{\tau} \tag{7}$$

The user charge is set above the socially optimal charge given the road standard. The profit maximising road standard is found where

$$\frac{\partial \pi}{\partial x} = -q\frac{\partial c}{\partial x} - \frac{dK}{dx} = 0 \tag{8}$$

which is identical to the welfare maximising investment condition, equation (3a).

Equations (3a) and (8) represent the cost minimising trade-off between road users' costs and capital costs for a given traffic volume. They minimise $cq + K$ with q fixed. However, in the monopoly case, road standard is optimised for a lower traffic volume because of the higher, monopoly charge.[34] The reason the monopolist's optimal road standard given traffic volume

3. Oum et al. (2004) found the same for the level of capacity supplied by an unregulated airport monopoly.

is the same as the social optimum is that the monopolist is able to appropriate the entire user benefit from an improvement in standard.[5]

As noted previously, provided user valuations of quality attributes stay constant, generalised cost can be used to measure *welfare changes* from service quality improvements. To use it for measuring *profit changes*, as equation (8) does, requires an additional assumption. Specifying the inverse demand curve as $\tau = \tau(q, x)$ and holding traffic volume fixed, the gain to the monopolist from a unit improvement in road standard is $q\frac{\partial \tau}{\partial x}$, which is quantity times the valuation of the improvement by the *marginal* user. Equation (8) correctly represents profit maximising behaviour only if the *average* increase in WTP for all users is the same as for the *marginal* user. As Spence (1975, p. 419) states, 'The average [valuation of a quality change] is the relevant quantity for welfare, but the firm responds to the marginal individual'. Xiao et al. (2007) recognise this by noting the need to assume homogeneous users with identical values of time. For a given quantity, road standard will be undersupplied

---

4. Tan et al. (2010) show that, under the assumptions of footnote 2, the profit and welfare maximising service qualities are the same. The volume–capacity ratios are identical because the profit maximising capacity is below the welfare maximising capacity in the same proportion as for volume.

5. The monopolist could increase the charge by the full amount of *∂c/∂x* leaving generalised price unchanged. But in keeping with equation (7), it allows generalised price to fall, generating some additional traffic, which lessens the average generalised cost reduction to *dc/dx*. By allowing generalised price to fall, the monopolist is not sharing the benefit from the road standard improvement with road users. Using equation (5), we can express the monopolist's marginal revenue gain in equation (8) as $-q\frac{\partial c}{\partial x} = \hat{\tau}\frac{dq}{dx} - q\frac{dc}{dx}$. Part of the cost saving gain, $\hat{\tau}\frac{dq}{dx}$, is converted into surplus earned from generated traffic. The monopolist earns an additional surplus from generated traffic of $(\tau_m - \hat{\tau})\frac{dq}{dx}$, which is a welfare gain due to price exceeding marginal social cost. The fall in generalised price is a transfer of this surplus from generated traffic to existing traffic, which can be seen by multiplying equation (7) by *dq/dx* to obtain $(\tau_m - \hat{\tau})\frac{dq}{dx} = -q\frac{dp}{dx}$.

if intra-marginal users value an improvement in road standard more highly than the marginal user and conversely (Spence 1975, Sheshinski 1976, Sappington 2005). It depends on whether the quality improvement makes the demand curve steeper or flatter. Whether the average valuation over all users of a road standard improvement can reasonably be assumed to be the same as the marginal user's valuation is an empirical question. The answer could vary between individual roads depending on users' trip purposes and the availability of substitutes offering higher or lower price–quality combinations.[6] Non-constant returns to scale will also affect the profit-maximising road standard because they cause the value of dK/dx at the lower traffic volume in equation (8) to differ from the value at the social optimum in equation (3b) (Tan et al. 2010). While it is certain an unregulated monopoly road supplier will charge an above-optimum price given the road standard, it not possible to generalise about how the profit-maximising road standard compares with the optimum standard either at the monopoly price or at the welfare maximising price.

There is no ambiguity about the road standard supplied when, in order prevent monopoly charging, a regulator fixes the charge. Spence, Sheshinski and Sappington all demonstrate that '… a monopoly supplier of a single product will always supply less than the welfare maximising quality when the firm is required to sell its product at a fixed price … because a price ceiling prevents the firm from capturing any of the incremental consumers' surplus that a higher service quality would engender' (Sappington 2005, pp. 130-1). Say a regulator directs a road monopolist to charge a fixed amount, τ, which is below the profit maximising price. The supplier's profit function becomes

6. For example, say a privately operated toll road competed with a congested untolled road. At high tolls, only drivers with high values of time would use the toll road. As the toll fell, drivers with progressively lower values of time would be attracted to the toll road. Hence, the value of time would fall as quantity rose along the demand curve. The marginal user of the toll road would then place a lower value a time-saving quality improvement than the average user.

$$\pi = \tau q - K$$

With the charge set exogenously, the monopolist has just the road standard to optimise.

$$\frac{d\pi}{dx} = \tau \frac{dq}{dx} - \frac{dK}{dx} = 0 \qquad (9)$$

From a unit increase in road standard, the supplier gains only the marginal benefit from generated traffic, area $A \approx \tau \frac{dq}{dx}$ in figure 1. The supplier is unable to appropriate the marginal benefit to existing users, area *B*. With marginal revenue from investment, area *A*, below marginal social benefit, areas *A* + *B*, the supplier will under-invest.[7]

In the absence of regulation, roads will be priced at monopoly levels. Fixing the monopoly pricing problem by having a regulator set the user charge creates another problem, under-provision of service quality. For a congested road, if the regulator were to set the charge at the optimal level *given road capacity*, the incentive to under-invest would be magnified because at lower capacities, the optimal congestion charge would be higher. The charge would fall as the supplier improves road standard, which is the opposite of the incentive required to induce the supplier to provide optimal service quality. The amount per vehicle received by the supplier needs to *increase* with road standard, which is the basis of our incentive regulation scheme.

4. Incentive regulation solution

4.1 Incentive regulation applied to road service quality

Incentive regulation uses financial rewards and penalties, instead of commands, to encourage good performance by a public utility or private supplier. Sappington (1994, p. 246) defines

7. Tan et al. (2010) show that price-cap regulation for a road under a BOT contract leads to under-supply of capacity.

incentive regulation as 'the implementation of rules that encourage a regulated firm to achieve desired goals by granting some, but not complete, discretion to the firm'. Price cap regulation is the best-known form of incentive regulation. With a fixed ceiling on the price a regulated firm can charge, the firm has an incentive to minimise costs just as a pricing-taking competitive firm (Lyon 1994, p. 12).

Sappington (2005) discusses use of bonuses and penalties to induce a regulated firm to achieve desirable service quality levels at minimum cost. The regulator can specify service quality targets at levels it estimates to be welfare maximising, and then set bonuses for exceeding the target and penalties for under-achievement.

> 'If the bonuses and penalties presented to the firm closely approximate the marginal benefits and costs to consumers of increases and decreases in quality, the profit-maximising regulated firm will expand quality to the point where the marginal benefit of additional quality to consumers (and thus the firm's marginal reward) equals the firm's marginal cost of increasing quality' (p. 134).

Practical difficulties with such schemes include availability to the regulator of information about consumers' valuations of quality, the multi-dimensional nature of service quality, and non-linear customer valuations of quality (Sappington 2005).

These difficulties are far less pronounced for roads. A substantial body of knowledge and computer modelling capability has been developed to estimate generalised costs for the purposes of undertaking economic and financial appraisals of investment projects and maintenance decisions. Generalised cost provides a single monetary measure of road users' valuations of service quality. Models that estimate generalised costs use data on the physical characteristics of roads (number of lanes, lane widths, shoulder widths, surface type, legal speed limit, gradient, curvature, roughness) and on the characteristics of the traffic (average annual daily traffic level, vehicle type proportions, hourly volume distribution, directional split) and include speed–flow relationships. The ability to make objective estimates of road

users' valuations of service quality makes service quality regulation with financial rewards and penalties related to impacts on users a practical proposition for roads.

4.2 Simple case of a fixed traffic volume

For ease of exposition, we assume initially that demand is perfectly inelastic over the relevant range. The volume of traffic is therefore fixed. Under the proposed incentive regulation scheme for roads, to avoid monopoly charges, the regulator sets road user charges. To decouple payments to the commercial supplier from the road user charge, revenue from the charges is paid into a road fund. Surpluses earned by the fund accrue to the government and deficits are paid for by the government. From the fund, the regulator pays the supplier a shadow toll per vehicle on each road segment determined by a formula. The regulator estimates the optimal or 'target' standard for the road segment along with the associated levels of annualised capital cost, K^*, traffic volume, q^*, and average generalised cost, c^*. To exactly cover costs at the target road standard, the road supplier has to receive the 'base shadow toll', $\tau^* = K^*/q^*$ per vehicle. The regulator pays the road supplier a shadow toll per vehicle of $\tilde{\tau} = \tau^* - (c - c^*)$. To the extent the road standard is below the target standard, road users incur generalised costs above c^*. The road supplier is then penalised by having the shadow toll reduced by $c - c^*$ per vehicle.

Say the road supplier invests ΔK to reduce generalised costs from c_1 to c_2. The benefit to society is the saving in generalised costs times the volume of traffic, $(c_1 - c_2)q = -q\Delta c$. The investment is economically worthwhile if the benefits exceeds the costs, $-q\Delta c > \Delta K$. The supplier remunerated according to the shadow toll formula would gain $\{[\tau^* - (c_2 - c^*)] - [\tau^* - (c_1 - c^*)]\}q = (c_1 - c_2)q = -q\Delta c$, which is identical to the economic benefit. Thus the shadow toll formula internalises the benefits from investment to improve service quality.

For small changes, the supplier's marginal revenue from investing in an additional unit of road standard is $-q\frac{dc}{dx}$. The supplier compares this with the marginal cost of investment *dK/dx* and invests to the point where they are equal. The supplier's profit maximising level of investment is then the same as the welfare maximising level. If the regulator sets τ^* and c^* correctly, then at the profit maximising investment level, $c = c^*$, so the supplier receives τ^* per vehicle leading to exact cost recovery with a normal profit.

It might not be immediately obvious that the target generalised cost level, c^*, does not guide investment. In the simple case of perfectly inelastic demand, its only role is to help determine the supplier's remuneration. Investment is guided by the mechanism established by the regulator assessing changes in road users' generalised costs and adjusting the shadow toll accordingly. The system gravitates towards the most economically efficient investment outcome.

Accurate measurement and forecasting of traffic levels and estimation of changes in user costs with respect to road standard (q and dc/dx) is important, but no more so than under existing arrangements whereby government road suppliers make decisions based on cost–benefit analysis. The motivation for accurate estimation of traffic levels and users' generalised costs could be stronger under incentive regulation because the data collected and the models used to estimate generalised costs take on enhanced importance due to their role in determining the supplier's remuneration. The data collection methods, the models and the parameters within the models would be part of the legal agreements between the regulator and the supplier.

The shadow toll formula can be modified to induce the supplier to invest at non-optimal levels if required. Where there is serious under-investment, a target set at the economic optimum will cause the supplier to incur losses over a period of years until investment reaches the optimum. The regulator may therefore wish to set less ambitious

targets in the short and medium terms. The regulator might wish to set a target above the optimum if it believes there are agglomeration economies or wishes to engender over-provision of infrastructure for social reasons. The latter is often the case for low-volume roads in rural areas in developed countries.

A non-optimal target road standard can be expressed as a target MBCR, $\mu^* = -q\frac{dc}{dx}/\frac{dK}{dx}$, not equal to one. A target MBCR above one implies a below-optimal target road standard and conversely. To ensure the supplier earns zero economic profits at a non-optimal target road standard, the regulator must set τ^* and c^* at the levels consistent with the non-optimal target road standard. But this is not by itself sufficient to induce the supplier to invest at the target. To illustrate, at the below-optimal road standard with an MBCR of two, an additional dollar of investment yields two dollars of benefit to society. The shadow toll formula under incentive regulation converts this into two additional dollars of revenue to the supplier. The supplier will therefore be induced to exceed the target. The marginal reward from investing has to be halved. The shadow toll formula can be rewritten as

$$\tilde{\tau} = \tau^* - \psi(c - c^*) \qquad (10)$$

where $\psi = 1/\mu^*$ is termed the 'correction factor'.

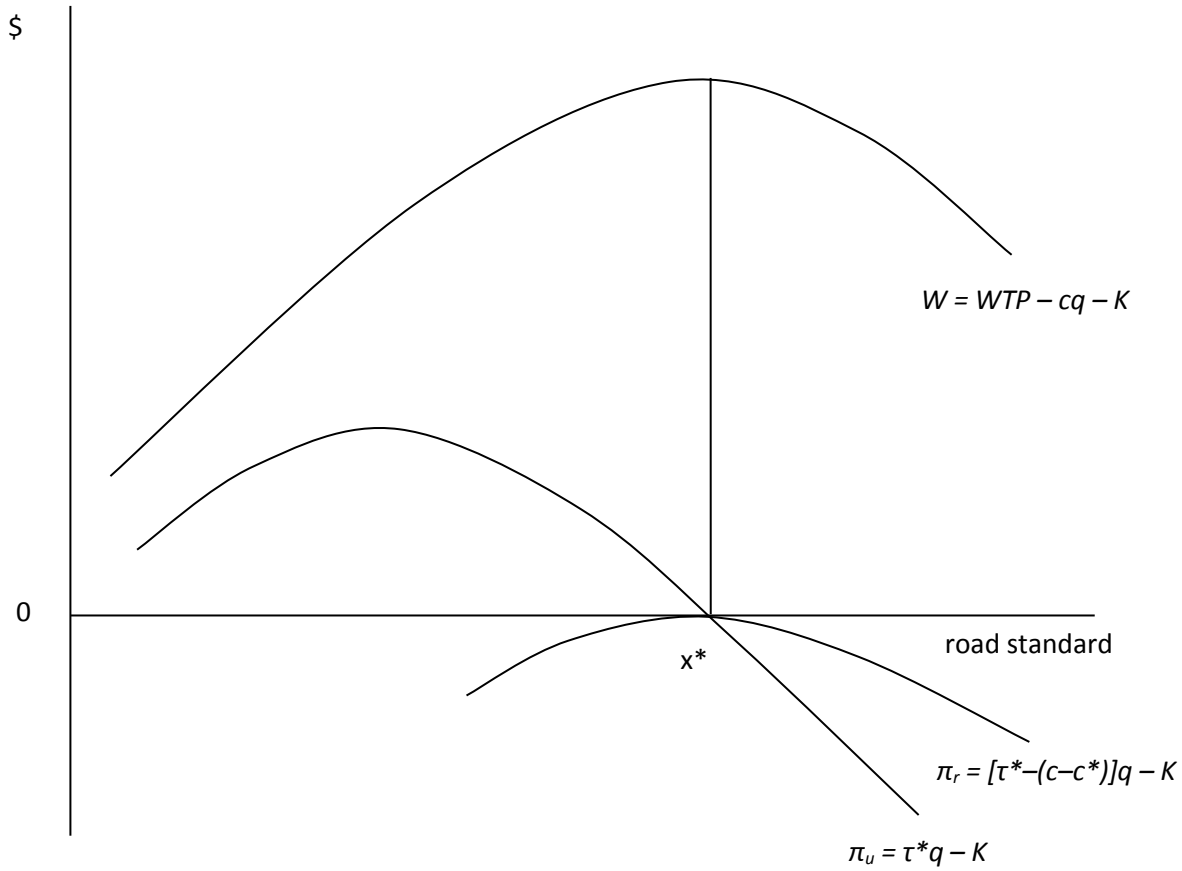
4.3 Variable traffic volume case

When the demand curve is downward-sloping, from a unit increase in road standard, a supplier remunerated according to the shadow toll formula still gains $-q\frac{dc}{dx}$ from existing traffic, which equals the welfare gain, area B, in figure 1. From generated traffic, the supplier gains the shadow toll times the increase in traffic volume, $[\tau^* - (c - c^*)]\frac{dq}{dx}$. For this to equal the welfare gain, area $A \approx \tau\frac{dq}{dx}$ in figure 1, two conditions must hold. First, the base

shadow toll must equal the user charge, $\tau = \tau^*$. Second, the actual road standard must be at the target so that $c = c^*$. The situation where $\tau \neq \tau^*$ is discussed below.

Figure 2 shows the welfare and profit curves as functions of road standard. The welfare curve, $W = WTP - cq - K$, reaches a maximum at the optimal standard, x^*. The 'unadjusted profit curve', $\pi_u = \tau^* q - K$, illustrates the case where a fixed road user charge is paid to a monopoly supplier with no adjustment for road standard. The unadjusted profit curve reaches a maximum below the social optimum, at the road standard where equation (9) holds. The profit curve under incentive regulation, $\pi_r = [\tau^* - (c - c^*)]q - K$, reaches a maximum at x^* in common with the welfare curve.

Figure 2 Welfare and profits as functions of road standard



Equality between marginal benefit to society and marginal revenue to the supplier occurs only at the target standard, which is set to coincide with the optimal standard. The

slope of the π_r curve differs from the slope of the welfare curve on either side of the optimum because $c = c^*$ only at the target standard. The shadow toll, which is the supplier's valuation of marginal generated traffic, rises with road standard (because c falls), while the user charge, $\tau = \tau^*$, which is society's valuation of marginal generated traffic, stays constant. Only for a zero price elasticity of demand do the two curves have the same slope at all road standards, because there is no generated traffic.

4.4 Formal exposition with correction factor

In this section, the assumption that $\tau = \tau^*$ is relaxed and non-optimal target road standards are permitted. It is shown that the correction factor, ψ, in the shadow toll formula can adjust the incentive power of the shadow toll formula to correct for situations where $\tau \neq \tau^*$ as well as for a target MBCR $\neq 1$.

From equation (4b), the welfare maximising condition subject to the constraint that investment be at the target MBCR of μ^* is

$$\left(\tau \frac{dq}{dx} - q\frac{dc}{dx}\right) \Big/ \frac{dK}{dx} = \mu^* \quad \text{or} \quad \frac{\tau}{\mu^*}\frac{dq}{dx} - \frac{q}{\mu^*}\frac{dc}{dx} = \frac{dK}{dx} \qquad \text{(11a and 11b)}$$

The supplier's profit function is

$$\pi = [\tau^* - \psi(c - c^*)]q - K \qquad (12)$$

Profit is zero at the target standard because $c = c^*$, $q = q^*$, and $\tau^* q^* = K^*$.

$$\frac{d\pi}{dx} = [\tau^* - \psi(c - c^*)]\frac{dq}{dx} - \psi q \frac{dc}{dx} - \frac{dK}{dx} = 0 \qquad (13)$$

At the target standard, the profit maximising condition in equation (13) is the same as the welfare maximising condition in equation (11b) provided $\tau = \tau^*$ and $\mu^* = \psi = 1$. In situations where $\tau \neq \tau^*$ or $\mu^* \neq 1$, it is necessary to find the value of ψ that causes equation (13), to equal

zero at the same value of *dK/dx* at which equation (11b) holds.

Letting $c = c^*$, then combining equations (13) and (11b) by eliminating *dK/dx*,

$$\psi = \frac{\left(\tau^* - \frac{\tau}{\mu^*}\right)\frac{dq}{dx}}{q\frac{dc}{dx}} + \frac{1}{\mu^*} \tag{14}$$

In the absence of generated traffic, $dq/dx = 0$ and ψ is the reciprocal of the target MBCR as shown above in section 4.2. With $\mu^* = 1$, the correction factor is still required when $\tau^* \neq \tau$ and $dq/dx \neq 0$ because the private value of marginal generated traffic, τ^*, differs from the social value, τ, at the target. If $\tau^* > \tau$, the supplier over-values each unit of generated traffic by $\tau^* - \tau$, which leads to over-investment. The correction factor causes the excess revenue gain from generated traffic, $(\tau^* - \tau)\frac{dq}{dx}$, to be deducted from the supplier's reward for reducing costs to existing traffic, $q\frac{dc}{dx}$. For non-unitary values of μ^*, the social value of generated traffic has to be adjusted, which explains the τ/μ^* term.

Since $p = c + \tau$ and τ is fixed, $\frac{dp}{dx} = \frac{dc}{dx}$ and

$$\frac{dq}{dx} = \frac{dq}{dp}\frac{dp}{dx} = \frac{q}{p}\eta\frac{dc}{dx} \tag{15}$$

where $\eta < 0$ is the price elasticity of demand. Substituting equation (15) into (14), yields a neater formula for the correction factor

$$\psi = \left(\tau^* - \frac{\tau}{\mu^*}\right)\frac{\eta}{p^*} + \frac{1}{\mu^*} \tag{16}$$

where p^* is the generalised price at the target road standard.

Referring to figure 2 above, with τ^* and c^* set to correspond with the target investment level, the unadjusted profit curve, $\pi_u = \tau^* q - K$, and the incentive regulated profit curve, $\pi_r = [\tau^* - (c - c^*)]q - K$, always intersect at zero profit at the target level of

investment, the point (x^*, 0). However, when either or both $\tau^* \neq \tau$ and $\mu^* \neq 1$, the incentive regulated profit curve does not attain a maximum at that point. The correction factor, ψ, pivots the incentive regulated profit curve around the point (x^*, 0) ensuring it reaches a maximum at that point. As evident in figure 2, the two curves form an upside-down, reversed Greek letter psi.

With optimal congestion charging, the user charge will fall as road standard is increased because congestion is reduced. The derivative *dq/dx* will be different because a unit increase in *x* is associated with a reduction in the user charge, which generates additional traffic. The correction factor formula in equation (14) still holds because it leaves open how *dq/dx* and *dc/dx* are calculated. Equation (16) has to be modified because it relies on equation (15), which assumes a constant user charge.[8]

5. Further consideration of the incentive regulation model

5.1 Difference between the base shadow toll and the user charge

The difference between the base shadow toll and the user charge affects the size of the correction factor and the level of cost recovery from variable user charges. For congested urban roads with constant returns to scale, the base shadow toll and the user charge will be the same with optimal pricing. For uncongested roads, to the extent that fixed user charges, land taxes or general taxes are relied upon to enable variable charges to be brought closer to the optimal level of zero, the user charge will lie below the base shadow toll. Rising costs of capacity expansion, which could occur in dense urban areas, when combined with optimal

8. With $\hat{\tau}$ varying with respect to *x*, $\frac{dp}{dx} = \frac{dc}{dx} + \frac{d\hat{\tau}}{dx}$. Assuming the user charge is set at the optimal level but the target MBCR may differ from one, equation (16) becomes $\psi = \left(\tau^* - \frac{\hat{\tau}}{\mu^*}\right)\left(1 + \frac{d\hat{\tau}}{dc}\right)\frac{\eta}{p^*} + \frac{1}{\mu^*}$. The derivative $\frac{d\hat{\tau}}{dc} = \frac{d\hat{\tau}}{dx}\Big/\frac{dc}{dx}$ is the change in the optimal user charge from the road standard improvement associated with a one dollar change to average generalised cost.

congestion pricing, would lead to a user charge above the base shadow toll and a surplus in the road fund.

5.2 Imperfect information

Information on demand and the existing physical infrastructure could be expected to be equally available to the regulator and the supplier, but there is still uncertainty about the future. The supplier could have better knowledge of construction costs and options to reduce users' generalised costs. It also has an incentive to manipulate information to exaggerate construction costs and hide possible ways to reduce user costs. On the other hand, the regulator would learn from observing demand responses to changes over time and innovations introduced by suppliers.

It was noted above that, in the simple case of a fixed traffic volume, errors in the levels of τ^* and c^* affect the supplier's remuneration only, not investment. In the variable traffic volume case, errors can affect investment via the supplier's valuation of generated traffic. Other things held equal, setting τ^* or c^* too high leads to over-valuation of generated traffic causing over-investment, and conversely. The error in the level of investment depends on the size of dq/dx, which depends largely on the price elasticity of demand. In a network, dq/dx could be quite high if improving the road in question diverts traffic from a substitute road. But the τ^* and c^* values for the substitute road also affect the outcome. If the values for both routes are too high or too low, the two errors could partly cancel each other out. An error in τ^* has no effect on investment if the correction factor is set concurrently because the correction factor makes the profit maximising investment level independent of τ^*.

In conclusion, provided the demand elasticity is not large, the risk of over- or under-remunerating the supplier under incentive regulation is likely to be much greater than the risk of over- or under-investment.

5.3. Maintenance

The incentive regulation scheme can be extended to cover maintenance costs. A formal mathematical treatment is provided in appendix A. The model in appendix A also allows for multiple vehicle types with different demand curves, cost functions and road damage impacts.

In the model, road roughness is included in the average generalised cost function so a rougher road leads to higher user costs. Pavement strength features in the investment cost function. A stronger pavement costs more to build but enables the same average level of roughness to be provided over time for lower maintenance costs. Maintenance costs are split into fixed and variable components. Since fixed maintenance costs are time related and vary with road standard and pavement strength, they can be incorporated into the investment cost function. The average fixed maintenance cost per vehicle at the target road standard would be added to the base shadow toll. Variable maintenance costs depend on traffic level, road standard and pavement strength. Users are assumed to be charged the variable maintenance costs they cause.[9]

Assuming there is just one vehicle type, under incentive regulation, the variable maintenance charge paid to the supplier, m*, is set at the level associated with optimal investment in road standard and pavement strength, together with maintaining optimal average roughness over time. The shadow toll formula is $\tau^* - (c - c^*) + m^*$. Actual variable

9. Maintenance costs comprise a 'routine' component that is fairly constant from year to year and a 'periodic' component that occurs at intervals of many years. Major rehabilitations can occur decades apart. Turvey's (1969) concept of marginal cost can be applied to estimate the marginal cost of road damage done by heavy vehicles. (Cars do negligible damage to paved roads.) The marginal cost of a given increment in demand is the present value of system (road supplier and road user) costs with the increment in demand less the present value of system costs without the increment. Additional heavy vehicle passes bring rehabilitations forward in time increasing the present value of system costs. There remains a large amount of fixed periodic maintenance costs that can be converted to a present value and annuitised over the life of the pavement just as for capital costs.

maintenance cost, *m*, does not appear because the supplier directly incurs maintenance costs, unlike *c*, which is incurred by road users.[10] With the correction factor, the formula becomes $\tau^* - \psi(c - c^*) - \psi(m - m^*) + m$.

The supplier can reduce average variable maintenance costs in two ways. It can allow roads to deteriorate to higher roughness levels before rehabilitiating, which increases the average roughness over time, or it can invest in stronger pavements that deteriorate more slowly. If the supplier saves maintenance costs by allowing roads to deteriorate to higher roughness levels, the additional costs imposed on road users are reflected back on to the supplier through the higher value of *c* in the shadow toll formula. If the supplier saves investment costs by constructing weaker pavements, the consequent additional variable maintenance costs are borne by the supplier without compensation. Thus the supplier faces the same cost trade-offs as for society as a whole.

5.4 Networks

It can be shown that the shadow toll formula aligns profit and welfare maximising investment outcomes in road networks where individual road segments are substitutes or complements for one another. See appendix B for a mathematical treatment.

Say a unit improvement is made to a road segment that diverts traffic from a substitute segment elsewhere in a congested network. Figures 3 and 4 show the demand and cost curves for the substitute segment. The demand curve for the substitute segment shifts leftward from D^1 to D^2. Figure 3 assumes optimal congestion pricing, so users are charged $\hat{\tau} = MC - AC$. The quantity demanded falls from q^1 to q^2. WTP is reduced by areas $A + E$.

10. The value of *m* would rise over time as the pavement deteriorates and fall immediately after a rehabilitation. Pooling of road segments at different stages of their lifecycles would smooth out the fluctuations in payments to the supplier as well as diversify away some of the uncertainty in pavement deterioration.

This loss of value is cancelled out by an equal and offsetting resource cost saving because it is also the area under the marginal social cost curve between q^1 and q^2. It is an axiom of cost–benefit analysis that when the project being appraised causes shifts in demand curves in related markets, if prices equal marginal social costs in those markets, there are no net welfare changes to consider (Mohring 1993).

It is assumed that a single incentive regulated supplier owns both the improved and the substitute road segments, the base shadow toll equals the user charge, $\tau^* = \tau = \hat{\tau}$, and the substitute section is at its target standard. As a result of the reduction in traffic, the supplier loses revenue of $\hat{\tau}\Delta q \approx$ area *A*. As a result of reduced congestion, users' generalised cost falls from c^1 to c^2 causing the shadow toll to rise by $c^1 - c^2 = \Delta c$. The supplier gains $q^2\Delta c$ = area *B* in revenue from users who remain on the road. Since $MC - AC = q\frac{\partial c}{\partial q}$, area *A* equals area *B* and the revenue loss by the suppler from diverted traffic equals the revenue gain from reduced costs for existing traffic. The traffic diversion, which is welfare neutral for society, is also revenue neutral for the supplier under incentive regulation.

Figure 4 illustrates the non-optimal pricing case. The user charge still equals the base shadow toll but is below the optimal charge, $\tau^* = \tau < \hat{\tau}$. The reduction in WTP is areas *C* + *E*. The resource cost saving is areas *A* + *C* + *E*, leaving a net welfare gain of area *A*. Generally, if price is below marginal social cost in a market for a substitute, a leftward shift in the demand curve results in a welfare gain, and a welfare loss occurs if price is above marginal social cost (Mohring 1993). The converse holds in a market for a complement. Under incentive regulation, the supplier makes a net gain of areas *B* – *C*, which equals area *A*, the welfare gain, because area *B* equals areas *A* + *C*. Hence, with non-optimal pricing, revenue changes in related markets under incentive regulation equal welfare changes. Appendix B covers correction factors for networks where base shadow tolls differ from user charges and the regulator sets a non-optimal target MBCR.

Figure 3 Welfare changes for a substitute road segment from a small leftward shift in the demand curve: optimal congestion charge

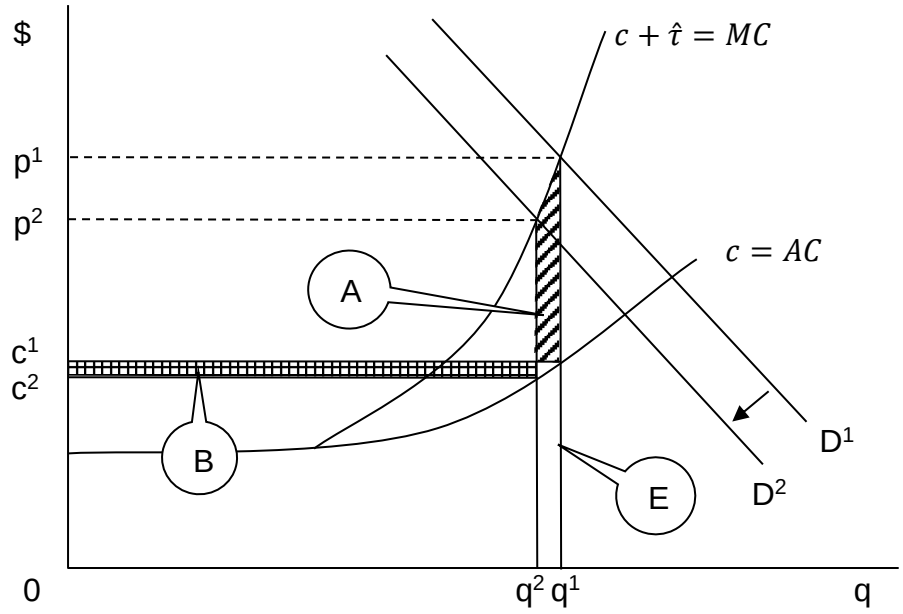
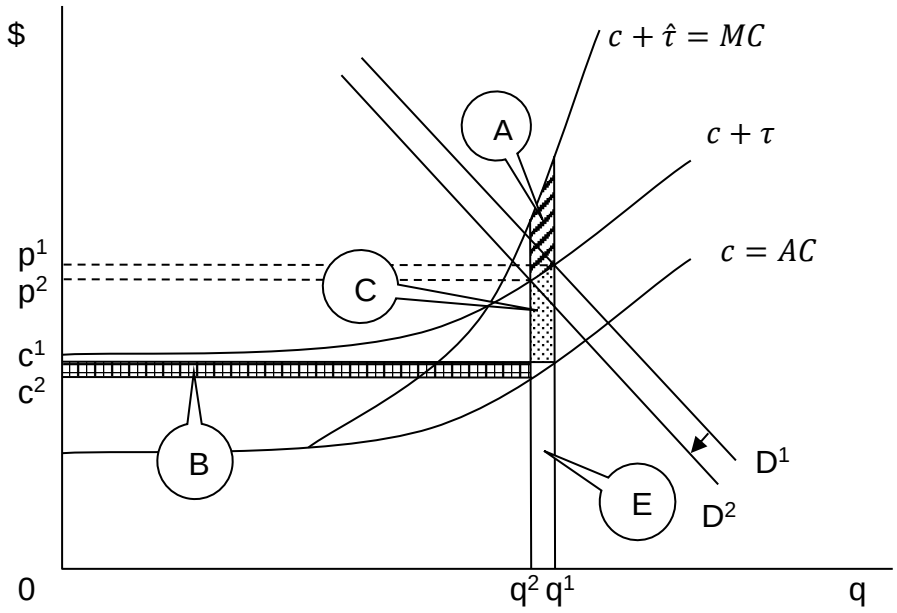


Figure 4 Welfare changes for a substitute road segment from a small leftward shift in the demand curve: below-optimal congestion charge



The foregoing discussion and appendix B assume that all the road segments with related demands are supplied by a single entity. That way, the supplier includes revenue changes on substitute and complement road segments in its financial analyses of investment decisions. This is not essential where there is optimal congestion pricing and a target MBCR of one because, under incentive regulation, there are no net revenue changes on the related segments. If incentive regulation were applied to tolled roads that compete for traffic with unpriced public roads in a congested network, correction factors would have to be employed to offset the distorting effect of the supplier's inability to internalise welfare changes elsewhere in the network.

6. Conclusion

The incentive regulation scheme developed in this paper induces a commercial (profit maximising) road supplier to provide optimal service quality. A regulator sets road user charges to prevent monopoly charging. Revenues raised flow into a road fund out of which shadow toll payments are made to the road supplier. A commercial supplier receiving a fixed price per user will under-invest in service quality because its marginal revenue from improved service quality comprises only the benefit from additional traffic generated by the quality improvement, not the benefit to existing traffic. The solution is to have the shadow toll vary negatively with users' average generalised cost, which serves as a measure of the level of service quality provided. Better service quality is rewarded with a higher shadow toll, and conversely. Thus the scheme converts the marginal social benefit to existing traffic from a service quality improvement into marginal revenue to the supplier.

As for generated traffic, how much of the marginal benefit is converted into marginal revenue depends on the level of the shadow toll. Where the marginal revenue from generated traffic differs from the marginal social benefit, a 'correction factor' in the shadow toll

formula is available to make an offsetting adjustment. With marginal revenue to the supplier from service quality improvements equal to marginal social benefit, the supplier provides the socially optimal service quality level. How the supplier chooses between the different dimensions of service quality and acts to provide them is entirely at its discretion. The supplier is motivated to act in an efficient and innovative manner because it retains any cost savings from achieving a given generalised cost outcome at a lower input cost.

The regulator still has to undertake cost–benefit analyses to estimate settings for parameters in the shadow toll formula. The parameter levels are important to ensure the supplier neither earns excess profits nor incurs losses, but generally are less important for determining the level of investment. As long as the price elasticity of demand is not large, investment is guided primarily by the process of the regulator measuring users' generalised costs and adjusting the shadow toll accordingly.

The scheme has remarkable flexibility. While it could be applied to toll roads, it is suitable for a public utility or private supplier of a large part of an entire road system. It is not dependent upon particular assumptions about returns to scale or the amount of road provision costs recovered from variable charges. Welfare can still be maximised, in a second-best sense, when road user charges are not at optimal levels. The scheme is applicable to congested and uncongested roads. It can work for both the congestion–capacity and the pavement damage–strength dimensions of road supply, and for a network of road segments with inter-related demands. The regulator can engineer above- or below-optimal service quality outcomes if desired. To do so requires explicit alterations to shadow toll formula parameters, which adds to the transparency of government decisions to depart from economically efficient outcomes.

The proposed incentive regulation scheme for road supply has been developed here at a highly conceptual level. It needs to be developed much further and many details worked out

before it could be considered a realistic proposition. How it would perform or be adapted when some of the assumptions are relaxed needs investigation. The paper briefly addressed the effects of imperfect information, but risk and uncertainty and asymmetric information require much deeper consideration, as do dynamics and lumpy investment.

If the scheme could be successfully translated into practice, it would offer a new way to manage a road system. It is widely acknowledged that the existing road supply arrangements have major shortcomings in securing efficient price, investment and maintenance outcomes. So there may be a willingness to consider seriously a radical alternative that has the potential to perform significantly better.

Appendix A Model with maintenance and multiple vehicle types

'Multiple vehicle types' here refers to categories of vehicles with different impacts on the demand for road capacity and wear and tear on pavements. Each vehicle type, distinguished by the subscript i, has its own inverse demand curve, $p_i = p_i(q_i)$, cost curve, which includes quantities of all the n vehicle types, user charge, τ_i, and base shadow toll, τ_i^*. Road roughness, r, is included in the generalised cost function for each vehicle type, $c_i = c_i(q_1, \ldots, q_n, x, r)$ where $\partial c_i/\partial r \geq 0$. Pavement strength, s, is included in the investment cost function $K = K(x, s)$ where $\partial K/\partial s > 0$.

Variable maintenance (pavement damage) cost per vehicle is $m_i = m_i(x, s, r)$ where $\partial m_i/\partial x \geq 0$, $\partial m_i/\partial s \leq 0$, and $\partial m_i/\partial r \leq 0$. Not including q_i in the function implies $\partial m_i/\partial q_i = 0$. Hence, total variable maintenance cost for vehicles of type i, $m_i q_i$, is proportional to the number of vehicles of that type. Fixed maintenance costs depend on x, s, and r. They are added to annualised capital costs so the investment cost function becomes $K = K(x, s, r)$ where $\partial K/\partial r \leq 0$.

Road user charges, τ_i, are set exogenously and may or may not be at the socially optimal level. Road pavement damage (or variable maintenance) costs are charged to users at short-run marginal cost, which equals average variable cost, m_i. The assumption of endogenous optimal charges for vehicle-related pavement damage was chosen over exogenous fixed charges because mass–distance–location charging for heavy vehicles is technically feasible and there are not the public acceptance difficulties faced by congestion pricing. The generalised price faced by road users of type *i* is then $p_i = c_i + \tau_i + m_i$.

The road supplier optimises three variables, *x*, *s* and *r*. With the τ_is set exogenously, and m_is determined by $m_i = m_i(x, s, r)$, the q_is are endogenous to the model, determined by the levels of the c_is and m_is via the demand curves.

The social welfare function to be maximise is

$$W = \sum \left[\int_0^{q_i'} p_i(q_i) dp_i - q_i c_i(q_1, \ldots, q_n, x, r) - q_i m_i(x, s, r) \right] - K(x, s, r) \quad \text{(A1)}$$

The derivative with respect to *x*, after substituting $\tau_i = p_i - c_i - m_i$ is

$$\frac{\partial W}{\partial x} = \sum \left(\tau_i \frac{dq_i}{dx} - q_i G_{ix} \right) - \frac{\partial K}{\partial x} = 0 \quad \text{(A2)}$$

where $G_{ix} = \frac{dc_i}{dx} + \frac{\partial m_i}{\partial x}$. The derivatives with respect to *s* and *r* and the definitions of G_{is} and G_{ir} are not shown because they are identical to those for *x* but with '*x*' replaced with '*s*' or '*r*'.

The derivatives of the user cost function in the $-q_i G_i$ terms are total, not partial. They allow for the indirect impact on users' costs of changes in quantities via the demand curve as well as the direct impacts of changes in *x*, *s* or *r*. While an increase in pavement strength in $\partial W/\partial s$ has no direct effect on users' costs, it reduces variable maintenance costs and hence road wear charges, which increases traffic, q_i, via the demand curve. On congested roads higher traffic would lead to a higher c_i value so that $dc_i/ds > 0$ for all vehicle types. On uncongested roads, there would be no change in c_i so $dc_i/ds = 0$. The equation $\partial W/\partial r = 0$

balances the trade-off between user costs and maintenance costs given road standard and pavement strength.

Target MBCRs with respect to the three decision variables can be defined as $\mu_x^* = \sum \left(\tau_i \frac{dq_i}{dx} - q_i G_{ix}\right) / \frac{\partial K}{\partial x}$, $\mu_s^* = \sum \left(\tau_i \frac{dq_i}{ds} - q_i G_{is}\right) / \frac{\partial K}{\partial s}$ and $\mu_r^* = \sum \left(\tau_i \frac{dq_i}{dr} - q_i G_{ir}\right) / \frac{\partial K}{\partial r}$, which represent the increase in welfare from spending an additional dollar to change *x*, *s* and *r* respectively, holding the other two decision variables constant. In the case of roughness, the numerator and denominator are both negative. All three MBCRs are one at the welfare maximising optimum. For a non-optimal target, we assume $\mu_x^* = \mu_s^* = \mu_r^* = \mu^*$ so the regulator maximises welfare given *K*. Hence,

$$\frac{1}{\mu^*} \sum \left(\tau_i \frac{dq_i}{dx} - q_i G_{ix}\right) = \frac{\partial K}{\partial x} \qquad \text{(A3)}$$

for *x*, and likewise for *s* and *r*.

Under incentive regulation, the shadow toll paid for vehicle type *i* is $\tilde{\tau}_i = \tau_i^* - \psi(c_i - c_i^*) - \psi(m_i - m_i^*) + m_i$. The supplier incurs the variable maintenance costs and the combined investment and fixed maintenance costs, $\sum m_i q_i + K$. Profit is

$$\pi = \sum \tilde{\tau}_i q_i - \sum m_i q_i - K = \sum [\tau_i^* - \psi(c_i - c_i^*) - \psi(m_i - m_i^*)] q_i - K \qquad \text{(A4)}$$

The supplier sets, *x*, *s* and *r* to maximise profits.

$$\frac{\partial \pi}{\partial x} = \sum [\tau_i^* - \psi(c_i - c_i^*) - \psi(m_i - m_i^*)] \frac{dq_i}{dx} - \psi \sum q_i G_{ix} - \frac{\partial K}{\partial x} = 0 \qquad \text{(A5)}$$

The derivatives with respect to *s* and *r* are not shown because they are identical those for *x* but with changed subscripts.

At the target levels of road standard, pavement strength and roughness, profit is zero because $c_i = c_i^*$ and $m_i = m_i^*$ for all *i* and $\sum \tau_i^* q_i^* = K^*$. If $\tau_i^* = \tau_i$ for all *i* and $\mu^* = \psi = 1$, equation (A5) is identical to equation (A2) and likewise for *s* and *r*. For situations where $\tau_i^* \neq$

τ_i and/or $\mu^* \neq 1$, the correction factor is found by combining equation (A5) with equation (A3) by eliminating $\partial K/\partial x$. Since $p_i = c_i + \tau_i + m_i$ and the τ_is are fixed, $dp_i/dx = G_{ix}$ for all *i* and

$$\frac{dq_i}{dx} = \frac{dq_i}{dp_i}\frac{dp_i}{dx} = \frac{q_i}{p_i}\eta_i G_{ix} \tag{A6}$$

where $\eta_i < 0$ is the price elasticity of demand for vehicle type *i*. Substituting (A6) and dividing numerator and denominator by *dK/dx*

$$\psi = \frac{\Sigma\left(\tau_i^* - \frac{\tau_i}{\mu^*}\right)\frac{q_i}{p_i^*}\eta_i G_{iK}}{\Sigma\, q_i G_{iK}} + \frac{1}{\mu^*} \tag{A7}$$

where $p_i^* = c_i^* + \tau_i + m_i^*$ is the generalised price for vehicle type *i* at the target and $G_{iK} = \frac{dc_i}{dK} + \frac{\partial m_i}{\partial K}$ is the increase in user costs and variable maintenance costs from spending an additional dollar to increase road standard. This corresponds with the suggestion in footnote 1, that *x* can be expressed in dollars of expenditure. With multiple vehicle types, cancellation of the $q_i G_{iK}$ terms is not possible to obtain equation (16). Following the same method, an identical result is obtained from the pavement strength and roughness relationships.

Appendix B Incentive regulation in a network

Assume the *n* road segments in a network have related demand curves and are provided by a single supplier. An improvement to the standard of one segment diverts traffic from segments along parallel (substitute) routes causing leftward shifts of their demand curves, and increases traffic on upstream and downstream (complementary) segments causing rightward shifts of their demand curves. The inverse demand curves are represented by $p_i = p_i(q_1, \cdots, q_n)$ for all segments $i = 1$ to *n*.

Multiple-market WTP is the line integral along a path of quantity changes, *C*, from the origin vector to the quantity vector $(q_1', \cdots, q_n')$. Provided the condition for path independence of line integrals (the integrability condition) is met, $\partial q_j/\partial x_i = \partial q_i/\partial x_j$ for all $i \neq j$, total WTP is the same regardless of the path chosen. The welfare function is

$$W = \int_C \sum p_i dq_i - \sum c_i q_i - \sum K_i \qquad \text{(B1)}$$

The vector of charges levied on road users, $(\tau_1, \cdots, \tau_n) = (p_1, \cdots, p_n) - (c_1, \cdots, c_n)$, is exogenously determined and need not be optimal. The condition for optimal investment in segment 1 is

$$\frac{\partial W}{\partial x_1} = \sum p_i \frac{dq_i}{dx_1} - \sum c_i \frac{dq_i}{dx_1} - \sum q_i \frac{dc_i}{dx_1} - \frac{dK_1}{dx_1} = \sum \tau_i \frac{dq_i}{dx_1} - \sum q_i \frac{dc_i}{dx_1} - \frac{dK_1}{dx_1} = 0 \text{ (B2)}$$

Equation (B2), together with the partial derivatives for the other segments from 2 to *n*, constitute a set of simultaneous equations that could be solved to obtain the vector of optimal standards for all segments.

To interpret equation (B2) (but not for comparing with the profit maximising condition under incentive regulation), we make a substitution for $\sum q_i \frac{dc_i}{dx_1}$. In the same way used to derive equation (5) in the main body of this article, with optimal prices represented by $\hat{\tau}_i = q_i \frac{\partial c_i}{\partial q_i}$ for all *i*,

$$q_i \frac{dc_i}{dx_1} = \hat{\tau}_i \frac{dq_i}{dx_1} + q_i \frac{\partial c_i}{\partial x_i} \frac{dx_i}{dx_1} \qquad \text{(B3)}$$

For segment 1, $\frac{dx_i}{dx_1} = 1$. For all other segments, $i = 2$ to *n*, $\frac{dx_i}{dx_1} = 0$ because changes to their standards are not being considered for $\partial W/\partial x_1$. After substituting equation (B3) into equation (B2) for *i* = 1,

$$\frac{\partial W}{\partial x_1} = \sum(\tau_i - \hat{\tau}_i)\frac{dq_i}{dx_1} - q_1\frac{\partial c_1}{\partial x_1} - \frac{dK_1}{dx_1} = 0 \qquad \text{(B4)}$$

Equation (B4) shows that when prices in markets for substitutes or complements are at optimal levels (marginal social costs), $\tau_i = \hat{\tau}_i$ for any $i \neq 1$, changes in these markets due to shifts in demand curves between the base case and project case in a cost–benefit analysis are welfare neutral. Where price is below marginal social cost in a related market, for a leftward shift in the demand curve ($dq_i/dx_1 < 0$), there is a positive benefit equal to the difference between the optimal and the actual user charge for each unit of quantity change. With price above marginal cost, there is a negative benefit. The converse holds for a rightward shift of the demand curve.

Allowing for different correction factors for different segments, the incentive regulated supplier's profit function for the whole network is

$$\pi = \sum[\tau_i^* - \psi_i(c_i - c_i^*)]q_i - \sum K_i \qquad \text{(B5)}$$

Differentiating with respect to x_1, then setting $c_i = c_i^*$ for all segments i, assuming investment is at the target levels in all segments

$$\frac{\partial \pi}{\partial x_1} = \sum \tau_i^* \frac{dq_i}{dx_1} - \sum \psi_i q_i \frac{dc_i}{dx_1} - \frac{dK_1}{dx_1} = 0 \qquad \text{(B6)}$$

Differentiating with respect to the x's for all segments gives rise to n simultaneous equations that can be solved to obtain the profit maximising road standards for all n segments. Equation (B6) is identical to the welfare maximising condition, equation (B2), provided $\tau_i^* = \tau_i$ and $\psi_i = 1$ for all i.

To maximise constrained welfare assuming an identical target MBCR $= \mu^*$ for all segments, from equation (B2)

$$\sum \frac{\tau_i}{\mu^*}\frac{dq_i}{dx_1} - \sum \frac{q_i}{\mu^*}\frac{dc_i}{dx_1} = \frac{dK_1}{dx_1} \qquad \text{(B7)}$$

Combining equations (B6) and (B7) by eliminating $\frac{dK_1}{dx_1}$

$$\sum\left(\tau_i^* - \frac{\tau_i}{\mu^*}\right)\frac{dq_i}{dx_1} - \sum\left(\psi_i - \frac{1}{\mu^*}\right)q_i\frac{dc_i}{dx_1} = 0 \qquad \text{(B8)}$$

Deriving this equation for all segments in the network gives rise to *n* simultaneous equations, the solution to which is *n* correction factors, one for each segment.

Acknowledgements

An early version was presented at the Australasian Transport Research Forum Conference, Canberra, September 2010. The author thanks David Hensher and anonymous reviewers for comments.

References

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

Dupuit J. 1844. On the measurement of the utility of public works. In: Munby D (ed), 1968. Transport: selected readings. Penguin: Harmondsworth Middlesex.

Geddes, RR. 2010. The road to renewal: private investment in U.S. transportation infrastructure. Rowman and Littlefield: Lanham, Maryland.

Guo X, Yang H. 2009. Analysis of a build–operate–transfer scheme for road franchising. International Journal of Sustainable Transportation 3(5):312-38.

Lyon TP. 1994. Incentive regulation in theory and practice. In: Crew MA (ed), Incentive regulation for public utilities. Kluwer Academic Publishers: Boston.

Mohring H, Harwitz M. 1962. Highway benefits: an analytical framework. Northwestern University Press: Evanston Illinois.

Mohring H. 1993. Maximizing, measuring, and *not* double counting transportation-improvement benefits: a primer on closed- and open-economy cost–benefit analysis. Transportation Research Part B 27(6):413-424.

Newbery DM. 1994. The case for a public road authority. Journal of Transport Economics and Policy 28(3):235-53.

Oum TH, Zhang A, Zhang Y. 2004. Alternative forms of economic regulation and their efficiency implications for airports. Journal of Transport Economics and Policy 38(2):217-246.

Roth G. 1996. Roads in a market economy. Ashgate Publishing Company: Aldershot Hants.

Roth G. 2006. Why involve the private sector in the provision of public roads? In: Roth G (ed), Street smart. Transaction Publishers: New Brunswick.

Sappington DEM. 1994. Designing incentive regulation. Review of Industrial Organization 9(3):245-72.

Sappington DEM. 2005. Regulating service quality: a survey. Journal of Regulatory Economics 27(2):123-54.

Schade J. 2003. European research results on transport pricing acceptability. In: Schade J. and Schlag B. (ed), Acceptability of transport pricing strategies. Elsevier: Oxford.

Semmens J. 2006. De-socializing the roads. In: Roth G (ed), Street smart. Transaction Publishers: New Brunswick.

Sheshinski E. 1976. Price, quality and quantity regulation in monopoly situations. Economica 43(170):127-137.

Small KA, Verhoef ET. 2007. The Economics of urban transportation. Routledge London.

Spence MA. 1975. Monopoly, quality, and regulation. Bell Journal of Economics 6(2):417-429.

Tan Z, Yang H, Guo X. 2010. Properties of Pareto-efficient contracts and regulations for road franchising. Transportation Research Part B 44(4):415-433.

Turvey R. 1969. Marginal cost. Economic Journal 79(314):82-299.

Ubbels B, Verhoef ET. 2006. Acceptability of road pricing and revenue use in the Netherlands. European Transport \ Trasporti Europei XI(32):69-94.

Ubbels B, Verhoef ET. 2008. Auctioning concessions for private roads. Transportation Research Part A 42(1):155-172.

Verhoef ET. 2007. Second-best road pricing through highway franchising. Journal of Urban Economics 62(2):337-361.

Verhoef ET. 2008. Private roads: auctions and competition in networks. Journal of Transport Economics and Policy 42(3):463-493.

Verhoef ET, Mohring H. 2009. Self-financing roads. International Journal of Sustainable Transportation 3(5):293-311.

Walters AA. 1968. The economics of road user charges. International Bank for Reconstruction and Development. World Bank Staff Occasional Papers No. 5. John Hopkins University Press: Baltimore.

Xiao F, Yang H, Han D. 2007. Competition and efficiency of private toll roads. Transportation Research Part B 41(3):292-308.

Yang H, Meng Q. 2000. Highway pricing and capacity choice in a road network under a build–operate–transfer scheme. Transportation Research Part A 34(3):207-222.

Yang H, Meng Q. 2002. A note on 'highway pricing and capacity choice in a road network under a build-operate-transfer scheme'. Transportation Research Part A 36(7):659-663.

Zhang L, Levinson DM, Shanjiang Z. 2008. Agent-based model of price competition, capacity choice, and product differentiation on congested networks. Journal of Transport Economics and Policy 42(3):435–461.

Zietlow GJ. 2006. Role of the private sector in managing and maintaining roads. In: Roth G (ed), Street Smart. Transaction Publishers: New Brunswick.

Zmud J, Arce C. 2008. Compilation of public opinion data on tolls and road pricing. NCHPP Synthesis 377. Transportation Research Board. Washington D.C.