Australasian Transport Research Forum 2010 Proceedings
29 September – 1 October 2010, Canberra, Australia
Publication website: http://www.patrec.org/atrf.aspx

Commercial road supply with incentive regulation

Mark Harvey

Bureau of Infrastructure, Transport and Regional Economics
GPO Box 501, CANBERRA, ACT, 2601
Email: Mark.Harvey@infrastructure.gov.au

Abstract

Commercial approaches to road supply have been advocated to secure both adequate funds in total and more efficient use of funds. However, unregulated supply leads to monopoly prices and under-provision of infrastructure. The proposed incentive regulation scheme uses performance-based financial incentives to control investment and maintenance levels by a public utility or private road supplier. Performance is measured by weighted average social generalised costs of road use.

A regulator determines the levels of road user charges, sets the performance targets and measures actual performance. It pays revenues raised from road users into a fund from which it remunerates the supplier by paying a shadow toll determined by a formula. A supplier that meets its targets will recover costs and earn a normal rate of return on capital. The shadow toll is reduced for underperformance.

It is shown that, under assumptions of perfect information, perfect divisibility, and malleable capital, profit maximising outcomes under incentive regulation, exactly match welfare maximising outcomes. This applies regardless of whether road user charges are at optimal levels, to both the congestion–capacity and the pavement damage–strength dimensions of road supply, and to networks of roads with inter-related demands. The regulator can engineer above- or below-optimum outcomes if desired. Some practical aspects of incentive regulation are also addressed.

1. Motivating policy issues

Commercial approaches to road supply have been advocated to address concerns about both the adequacy and the allocation of road funding. Inadequate funding prevents economically warranted investment and maintenance works from being undertaken. Compounding the problem, a portion of funds is allocated to uneconomic investment and maintenance works. Some over-spending from an economic viewpoint may be justified on social grounds. Some may be due to mistakes. Some may be motivated by political considerations.

It is argued that charges directly linked to road usage can improve investment decisions because price signals provide information on users' demand in relation to infrastructure capacity (AFTS 2009 p. 402, BITRE 2008 p 58). High prices reflecting high congestion or heavy vehicle damage to pavements draw attention to a possible need for additional investment. Profitability provides information about the expost net worth of a past investment and an incentive to invest wisely in the future. The profit motive incentivises technical efficiency—to provide a given output at the lowest possible cost, or the highest output for a given cost—and innovation. Congestion pricing and user charges to fund roads may more readily gain community acceptance if there is direct link between revenue and expenditure on roads.

The term 'commercial', in the roads context, can be interpreted in different ways. The most basic level is hypothecation of tax revenue collected from road users to government road

agencies. The next level of commercialisation involves a dedicated road fund into which revenues from road user charges are paid. Higher levels of commercialisation entail supply by a public utility or a private firm with spending decisions based on commercial criteria. (Productivity Commission 2006, pp. 267-8)

Due to the natural monopoly characteristic of roads, in the absence of regulation, commercial decision making would lead to monopoly prices and underspending on road provision compared with economically efficient levels.

For other network industries such as electricity and telecommunications, regulation is mostly concerned with controlling monopoly prices, for example, through price-cap regulation. Regulating service quality is of secondary importance because there is not a large range of different quality levels that can be economically efficient under different circumstances. For roads, a very wide range of service qualities is warranted in different circumstances, from a four-wheel drive track to a multi-lane freeway. Economically efficient service quality levels for roads are determined by cost–benefit analysis. A very different outcome could result from application of commercial investment criteria to road infrastructure and this is the barrier to allowing investment decisions to be made using commercial criteria. In order to secure a good outcome in terms of economically efficient resource allocation, the regulator needs to control investment and maintenance decisions as well as charges. So much control leaves limited room for commercial decision making.

2. Incentive regulation applied to roads

This paper introduces an idea for remunerating a public utility or private road supplier in a way that makes profit-maximising investment and maintenance decisions consistent with economically efficient outcomes. As such, it is a form of 'incentive regulation'. Incentive regulation uses financial rewards and penalties, instead of commands, to encourage good performance by a public utility or private supplier.

Under the proposed incentive regulation scheme, a regulator sets the charges levied on road users and deposits revenues into a road fund. The regulator pays the road supplier a shadow toll, which is determined by a formula. The shadow toll for each vehicle-kilometre need not be the same as the user charge. The formula includes a target service level. If the supplier meets the target, the revenue from the shadow toll is just sufficient to recover the supplier's costs including a normal return on capital. This is also the supplier's profit maximising position. Underperformance in relation to the target engenders a reduction in the shadow toll.

The incentive regulation scheme is developed in the context of three models of road supply. The first model concerns a single road that experiences congestion. User charges are set at economically optimal levels. The second model is more general, allowing for non-optimal pricing, non-optimal service-level targets, and the pavement strength–damage dimension of road supply. The third model applies to networks and allows for non-optimal pricing and non-optimal targets. In each case, it is demonstrated that incentive regulation, when correctly applied, produces welfare maximising outcomes subject to the constraints imposed.

All three models assume perfect information, perfect divisibility and malleable capital, and so are quite theoretical. The last section but one briefly considers some of the practical aspects. There remain many more theoretical and practical issues to address before the concept could be applied in practice.

3. Optimal congestion pricing model

Optimal pricing and investment

The demand curve for road use is $q(p)$ where *p* is the generalised price equal to the sum of average generalised cost *(c)* and a distance-related (variable) road user charge or toll *(t)*,

that is, $p = c + t$. The average generalised cost is determined by 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 be taken to include alignment and safety. Hence, x is used rather than w for width. Generalised cost consists of costs of vehicle operation, time taken, trip variability, and the expected value of crash costs borne by road users. Normally, it would include the entire social costs of crashes and environmental externalities, however, this would complicate the model by creating need to distinguish between social and private generalised costs.

The standard of infrastructure determines the level of capital invested in the road. The annualised capital cost K is comprised of the investment cost of the assets annuitised over the life of the assets at the rate of return on capital. The rate of return on capital is the normal rate that would be earned in a competitive industry with similar risk characteristics. Note that references below to 'zero profits' mean normal profits as defined here.

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 social welfare function to be maximised is

Social welfare (W) = road users' willingness-to-pay (WTP) – road users' costs (cq) – the road supplier's investment cost (K)

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

To determine the economically optimal charge on road users

$$\frac{\partial W}{\partial q} = p(q') - c - q\frac{\partial c}{\partial q} = 0 \tag{2}$$

Since $p - c = t$, the well-known result for the optimum congestion charge, $\hat{t}$, is obtained

$$\hat{t} = q\frac{\partial c}{\partial q}. \tag{3}$$

This is the volume of traffic times the slope of the average cost curve, equal to the gap between the marginal social cost and average cost curves.

The economically optimal road standard is found by differentiating equation (1) with restpect to x.

$\frac{\partial W}{\partial x} = -q\frac{\partial c}{\partial x} - \frac{dK}{dx} = 0$ which implies $-q\frac{\partial c}{\partial x} = \frac{dK}{dx}$ (4a and 4b)

$-q\frac{\partial c}{\partial x}$ is the downward shift of the average cost curve from a unit increase in x times the number of vehicle-kilometres. dK/dx is the additional annualised investment cost from a unit increase in x. At the optimum, the marginal benefit from investing in road standard equals the marginal investment cost.

Equation (4b) can be rewritten as $-q\frac{\partial c}{\partial x}\Big/\frac{dK}{dx} = 1$, (5)

that is, the marginal benefit–cost ratio (MBCR) equals one The MBCR is defined as the gain to society from investing an additional dollar in infrastructure.

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. Say the function c is homogeneous of degree zero, that is, a proportional increase q and x leaves c unchanged. This would be the case if x was capacity and c was a function of the volume–capacity ratio, $c = c(q/x)$. Then, by Euler's theorem,

$q\frac{\partial c}{\partial q} + x\frac{\partial c}{\partial x} = 0$. Substituting equation (3) and rearranging, $\hat{t} = -x\frac{\partial c}{\partial x}$ (6)

Let $\kappa = \frac{dK}{dx}\frac{x}{K}$ be the elasticity of investment cost with respect to road standard (as in Mohring and Verhoef (2007)). Substituting equation (4b), and rearranging, $\kappa\frac{K}{q} = -x\frac{\partial c}{\partial x}$. (7)

Given revenue $R = q\hat{t}$, from equations (6) and (7), $\frac{\hat{t}q}{K} = \frac{R}{K} = \kappa$. (8)

Constant returns to scale implies $\kappa = 1$, and revenue equals investment cost. Increasing returns to scale (or economies of scale) implies $\kappa < 1$ and there is under-recovery of costs. The converse holds for decreasing returns to scale.

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 Mohring and Verhoef (2007) for literature surveys).

There are economies of scale in road supply. Costs of infrastructure along the sides of roads (shoulders, signs, guide posts, drainage ditches) are the same regardless of the number of lanes. Also, because of the greater passing opportunities, a four-lane road has more than twice the capacity of a two-lane road (Mohring and Harwitz 1962, p. 87, Hau 1992). There are enormous economies in respect of pavement strength (Newbery 1989). For flexible pavements, a 10 per cent increase in pavement thickness results in a doubling of the traffic loading required to produce a given amount of damage. However, for major urban roads, the economies of scale are offset by diseconomies of scale. The number of intersections increases faster than the number of lane-kilometres of roads in a network in a given area. Intersections are land-intensive and often require traffic signals or grade separation (Hau 1992).

There is a consensus in the literature that approximate constant returns to scale apply for *major urban roads*. Small and Verhoef (2007, p. 112) reviewed a number of studies and conclude 'Altogether, the evidence supports the likelihood of mild scale economies for the overall highway network in major cities. Scale economies are probably substantial in smaller cities in which one or two major expressways are important, and may disappear altogether in very large cities where expanding expressways is extraordinarily expensive due to high urban density'. So, in the long term, revenue from optimal pricing of congested urban roads should approximately cover costs with optimal investment. For other roads, where economies of scale predominate, as discussed below, non-optimal pricing may be needed to recover costs.

Commercially optimal investment

In order to avoid charges at monopoly levels, assume a government regulator sets the user charge at the optimal congestion price given by equation (3). The commercial road supplier has no direct control over the price charged, but determines the road standard, *x*, to maximise its profit function

$\pi = R - K = \hat{t}q - K.$ (9)

Differentiating with respect to *x*,

$\frac{d\pi}{dx} = \hat{t}\frac{dq}{dx} + q\frac{d\hat{t}}{dx} - \frac{dK}{dx} = 0$ (10)

Taking the total differential of the user cost function $c = c(q, x)$, $dc = \frac{\partial c}{\partial q}dq + \frac{\partial c}{\partial x}dx$, dividing by *dx*, multiplying by *q* and substituting equation (3).

$q\frac{dc}{dx} = \hat{t}\frac{dq}{dx} + q\frac{\partial c}{\partial x}$ or $\hat{t}\frac{dq}{dx} = q\frac{dc}{dx} - q\frac{\partial c}{\partial x}$ (11a and 11b)

Substituting (11b) into (9), recalling that $p = c + t$, then substituting equation (4a), the profit-maximising road supplier will set

$$\frac{d\pi}{dx} = q\frac{dp}{dx} - q\frac{\partial c}{\partial x} - \frac{dK}{dx} = q\frac{dp}{dx} + \frac{\partial W}{\partial x} = 0 \quad (12)$$

Equation (12) shows that the profit-maximising level of investment differs from the welfare-maximising level. Since $\frac{dp}{dx} < 0$ (an increase in road standard reduces generalised price), $\frac{\partial W}{\partial x} > 0$ at the profit maximising level of investment. The unregulated supplier will under-invest relative to the social optimum.[1]

Figure 1 provides a graphical explanation. The top part graphs consumers' surplus and revenue against *K*. It is assumed that the regulator sets the user charge at the optimum level for the road standard *x* associated with the value of *K* on the horizontal axis. At low road standards, high user costs from congestion and a high charge discourage road use and revenue is low. As more capacity is provided, congestion and the charge fall and the revenue (*R*) increases to a maximum. Revenue drops thereafter as the loss of revenue from lower congestion charges predominates over the induced traffic effect. The curve reaches zero where capacity is so great that free-flow conditions are achieved and the optimal congestion charge becomes zero.

The 'consumers' surplus plus revenue' curve (*CS+R*) is total willingness-to-pay minus the total generalised costs of road use (*cq*). The curve rises at a diminishing rate, levelling off as free-flow, zero-congestion-charge conditions are approached and capacity has expanded to the point where further increases cease to lower users' generalised costs. The 45 degree line shows points where revenue equals investment cost (*R=K*). With constant returns to scale, from the Mohring–Harwitz theorem, the optimum level of investment in capacity occurs at *K**, where there is exact cost recovery with optimal pricing.

The lower part of the diagram shows the differences between the two curves and the 45 degree line. The *W = CS+R–K* curve shows total social welfare and the *π = R–K* curve, net profit to the road supplier. At the optimum level of investment in capacity, *W = CS+R–K* is maximised, and *π = R–K = 0*. The slope of the *CS+R* curve is one, that is, the MBCR is one.

An unregulated commercial road supplier would provide a road standard at cost level K_c where *π = R–K* is maximised. At this point, the slope of the revenue curve is one.

Incentive regulation scheme

Shadow toll

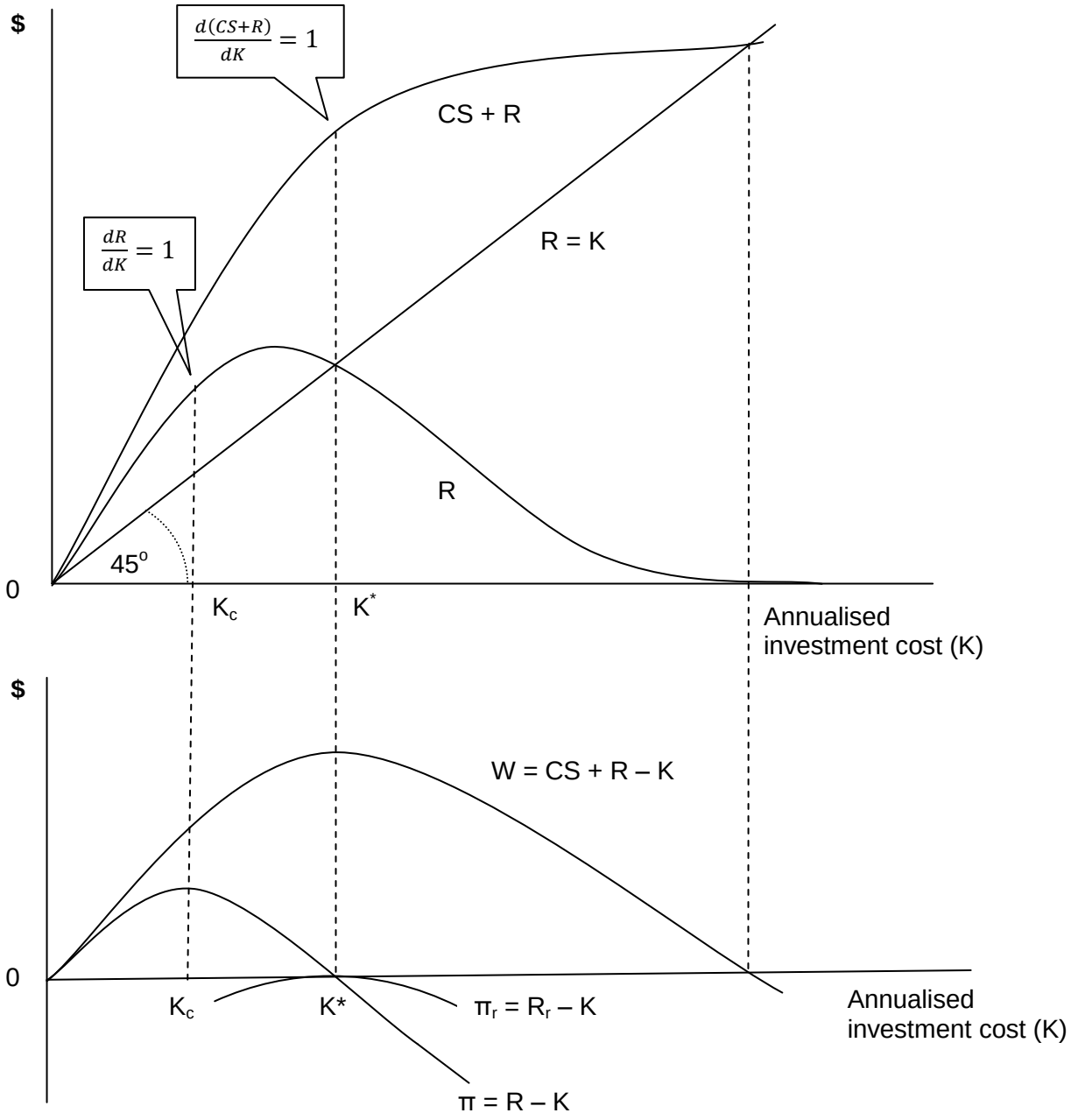
In the optimal pricing model, under the proposed incentive regulation scheme, road users pay to the regulator economically optimal charges for congestion.

The regulator, using data and models relating to demand, traffic flows, road construction costs and cost–benefit analysis, estimates the optimal or 'target' road standard along with the associated levels of

- annualised capital cost *K**
- traffic level in vehicle-kilometres, *q**, and
- average generalised cost, *c**, per vehicle-kilometre.

1. Small and Verhoef (2007, pp. 192-4) derive the profit maximising conditions for commercial supply of a congested road where both the user charge and capacity are unregulated. The profit maximising condition for capacity is shown to be the same as the welfare maximising condition, but capacity is optimised for a lower traffic volume compared with the welfare optimum because the higher, monopoly charge reduces traffic volume.

Figure 1: Consumers' surplus and revenue as functions of investment in road capacity with optimal congestion pricing



To exactly cover costs at the target road standard (including a normal return on capital), the road supplier has to receive the 'base shadow toll', $t^* = K^*/q^*$ per vehicle-kilometre. c^* becomes the supplier's target average generalised cost level.

The regulator pays the road supplier a shadow toll per vehicle-kilometre of $t^* - (c - c^*)$.

To the extent that the supplier provides a road standard below the target level, road users incur generalised costs above the target level ($c > c^*$). The road supplier is penalised by having the shadow toll reduced by $c - c^*$per vehicle-kilometre.

It seems reasonable to constrain the shadow toll so that it never becomes negative. As will be demonstrated below, in situations where $c < c^*$, it is not necessary to constrain the shadow toll to a maximum of t^*. The supplier has no incentive to exceed the target because the additional investment cost exceeds the gain from the higher shadow toll received. Also,

where the regulator faces uncertainly about target settings, preserving symmetry in the shadow toll formula is more likely to result in an expected value at the desired target level. To have the shadow toll remain at t^* for $c < c^*$ would impose a greater penalty on the supplier for exceeding the target than for falling short of it. The expected value of the outcome will be biased towards falling short of the target.

Say the road supplier invests in a project that reduces generalised costs from c_1 to c_2 where $c_1 > c_2$. The road supplier would gain $[t^* - (c_2 - c^*)] - [t^* - (c_1 - c^*)] = c_1 - c_2$ multiplied by the quantity of existing traffic. This is exactly the same as the economic benefit *for existing traffic* estimated in a cost–benefit analysis.

For generated traffic the supplier earns $t^* - (c_2 - c^*)$ for each additional vehicle-kilometre. This is not the same as the economic benefit from the generated traffic, but the two measures have the same sign (both positive for a road improvement) and, as is shown below, for small changes, converge as the target c^* is approached. For both cost–benefit analysis and financial analysis, in the absence of budget constraints, the decision to invest depends only on the sign of the net present value, not its size.

Supplier's profit curve

For the incentive regulation scheme to perform well, the supplier's profit function has to satisfy two conditions. First, it has to equal zero at the target road standard K^*.

The suppliers' profit function is

$$\pi = [t^* - (c - c^*)]q - K \tag{13}$$

Profit is zero at K^*, because $c = c^*$ and $t^*q^* = K^*$.

Second, the first derivative has to be zero at K^* so the suppliers' profit maximising level of investment corresponds to the welfare maximising level. Differentiating with respect to x,

$$\frac{d\pi}{dx} = [t^* - (c - c^*)]\frac{dq}{dx} - q\frac{dc}{dx} - \frac{dK}{dx} \tag{14}$$

Substituting equations (11a) and (4a)

$$\frac{d\pi}{dx} = [(t^* - \hat{t}) - (c - c^*)]\frac{dq}{dx} + \frac{\partial W}{\partial x} \tag{15}$$

As long as the target K^* and c^* are set to correspond with the economically optimal road standard, when the road is provided at the target standard, $c = c^*$ and $\partial W/\partial x = 0$. With constant returns to scale, $t^* = \hat{t}$ at the optimum. Hence, $d\pi/dx = 0$ at the economically optimal road standard.

At road standards different from the optimum, the slope of the supplier's profit curve differs from the slope of the welfare curve by $[(t^* - \hat{t}) - (c - c^*)]\frac{dq}{dx}$. The derivative dq/dx is the amount of generated traffic caused by a unit improvement in road standard. It is the product of the slope the demand curve, dq/dp, and the impact of a unit increase in road standard on the generalised price paid by road users, that is $dp/dx = dc/dx + d\hat{t}/dx$. The difference between the slopes of the welfare and the profit curves at non-optimal standards is due to the different values placed on generated traffic by the supplier and by cost–benefit analysis. The slopes converge to zero as investment approaches the optimum level.

In the lower part of figure 1, the incentive regulated supplier's profit curve is shown as $\pi_r = R_r - K$. It reaches a maximum at K^* with zero profit, but is flatter than the welfare curve.

4. More general model

The model just developed applies to congested roads with optimal pricing. The more general model developed in this section allows for non-optimal pricing, non-constant returns to scale, non-optimal targets, variable pavement strength, and variable pavement maintenance standard.

Non-optimal prices

On congested roads, non-optimal pricing could arise from absence of a congestion pricing scheme or from a cordon or area charging scheme that averages optimal congestion charges over many roads. On uncongested, low-volume roads, the optimal price is zero and a non-optimal variable charge may be levied to cover costs.

Most of the road system by length consists of uncongested or low-volume roads, that is, roads with volume–capacity ratios such that optimal congestion prices would be zero or low most or all of the time. Not only can most non-urban roads be categorised as uncongested, but also suburban streets and minor arterials in urban areas.

Walters (1968, p. 17) argued that such roads approximate to being pure public goods. The essential characteristic of a pure public good is that its enjoyment by one person does not in any way detract from its availability to others. Charging a price will inhibit some people from taking advantage of the good, which is wasteful because their consumption imposes no cost on society. In the transport context, this argument dates back to Dupuit (1844). For uncongested roads, $\partial c/\partial q = 0$, which implies, from equation (3), that the optimal price is zero.

Walters (1968) attributes the public good nature of low-volume roads to a combination of indivisibilities and economies of scale. A road must be at least as wide as the narrowest car plus a margin for safety, and wider still to allow trucks to pass over it. The number of lanes must be a whole number. There is a minimum pavement depth. A road must have an earth surface, a gravel surface or a sealed surface.

Some of the reasons for economies of scale in roads were noted above. For low-volume rural and inter-urban roads, Walters argued that another source of economies of scale is that roads are subject to jointness in the supply of capacity and quality. Supply of one automatically means that 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 the value of the savings in time, vehicle operating costs, and crash costs to road users. However, these improvements also add to capacity, keeping any congestion price to practically zero. For low-volume roads, the reason $\partial c/\partial x < 0$ is not that greater capacity reduces congestion as is the case for congested roads, but that a straighter, wider, smoother road gives users a faster and safer ride.

For low-volume roads, the pure economic approach is to charge zero prices (other than for pavement damage) and to recover the deficit from general taxation or from a land tax that does not affect resource allocation. Local roads, funded by rates paid to local governments are, in effect, 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 and are related to ability to pay because larger vehicles with higher operating and capital costs pay more in absolute terms.

Pavement damage and strength

The more general model features the following additions to the previous model.

- Road roughness, *r*, is included in the average generalised cost function $c = c(q, x, r)$ where $\partial c/\partial r \geq 0$.
- Pavement strength, *s*, is included in the investment cost function $K = K(x, s)$ where $\partial K/\partial s > 0$.

Maintenance costs are split into fixed and variable components. Fixed maintenance costs, that is, weather-related road deterioration, are incorporated into the annualised investment cost function $K = K(x, s)$, for simplicity. They could be accorded a separate function. Variable maintenance cost per vehicle, that is, pavement damage, is $m = m(x, s, r)$ where

$\partial m/\partial x > 0$, $\partial m/\partial s < 0$, and $\partial m/\partial r < 0$. Not including q in the function means we are assuming that $\partial m/\partial q = 0$. Hence, total variable maintenance cost, *mq*, is proportional to the number of vehicles. The fact that different vehicle types do different amounts of damage, in particular, cars do negligible damage, is addressed below. The supplier can reduce average variable maintenance costs by allowing roads to deteriorate to higher terminal roughness levels before rehabilitiating, which increases the average roughness over time, or by investing in stronger pavements that deteriorate more slowly.

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. The shadow toll formula becomes $t^* + m^* - (c - c^*)$ where $t^* = K^*/q^*$. If the supplier saves investment costs by constructing weaker pavements, the consequent additional variable maintenance costs are borne by the supplier without compensation. 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.

Non-optimal targets

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 not included in the welfare function, or wishes to engender over-provision of infrastructure for social reasons. The latter is often the case for low-volume roads in rural areas. Application of unmodified incentive regulation would lead to disinvestment in such roads. Since the aim is to transfer a benefit to the road users, the government would not want to set the user charge to achieve cost recovery. It would be desirable from both economic efficiency and public policy viewpoints for the additional financial costs of over-spending on roads for social purposes to be funded from general revenue rather than cross-subsidised by users of other roads.

To set a target road standard above or below the optimal level, it is necessary, but not sufficient, for the regulator to set the base shadow toll, *t**, and the target generalised cost, *c**, at levels consistent with the target road standard. In addition, a correction factor ψ, set below one for below-optimum targets and above one for above-optimum targets, must be added to the shadow toll formula. The formula becomes $t^* - \psi(c - c^*) - \psi(m - m^*) + m$. The reason *m* is added back is that the supplier incurs *m*, unlike *c*, which is incurred by road users.

The welfare function is maximised subject to a constraint that the level of service to users, measured by the sum of average generalised cost they incur and the pavement damage costs they pay, is equal to $c^* + m^*$.

Constrained welfare maximisation

In the absence of any constraints, it can be shown that the optimal charge is $\hat{t} = q\frac{\partial c}{\partial q} + m$, equation (3) with the addition of the short-run marginal cost of pavement damage. However, we assume that the road user charge *t* is set exogenously. It may or may not equal $q\frac{\partial c}{\partial q}$. Road damage (or variable maintenance) costs are charged to users at marginal cost (assumed to equal average cost), *m*. The generalised price faced by road users is then, $p = c + t + m$.[2]

2. An alternative model specification is to incorporate an exogenously determined variable road damage charge into *t*, which would allow for non-optimal pricing of pavement damage. The chosen option that assumes optimal pricing of pavement damage is considered more realistic because mass-

The road supplier optimises three variables, *x*, *s* and *r*. With *t* set exogeneously, and *m* determined by $m = m(x,s,r)$, *q* is endogenous to the model, determined by the levels of *c* and *m* via the demand curve.

The constrained social welfare function to be maximised is given by equation (16) where *λ* is the Lagrangian multiplier.

$$W = \int_0^{q'} p(q)dp - c\,q - mq - K + \lambda(c^* + m^* - c - m) \quad (16)$$

$$\frac{\partial W}{\partial x} = t\frac{dq}{dx} - q\left(\frac{dc}{dx} + \frac{dm}{dx}\right) - \frac{\partial K}{\partial x} - \lambda\left(\frac{dc}{dx} + \frac{dm}{dx}\right) = t\frac{dq}{dx} - qG_x - \frac{\partial K}{\partial x} - \lambda G_x = 0 \quad (17)$$

$$\frac{\partial W}{\partial s} = t\frac{dq}{ds} - q\left(\frac{dc}{ds} + \frac{dm}{ds}\right) - \frac{\partial K}{\partial s} - \lambda\left(\frac{dc}{ds} + \frac{dm}{ds}\right) = t\frac{dq}{ds} - qG_s - \frac{\partial K}{\partial s} - \lambda G_s = 0 \quad (18)$$

$$\frac{\partial W}{\partial r} = t\frac{dq}{dr} - q\left(\frac{dc}{dr} + \frac{dm}{dr}\right) - \lambda\left(\frac{dc}{dr} + \frac{dm}{dr}\right) = t\frac{dq}{dr} - qG_r - \lambda G_r = 0 \quad (19)$$

The amount $q\frac{dc}{dx}$ is the reduction in average generalised costs to existing road users from a unit increase in road standard times the number of users *q*. This is offset, in a degree, by an increase in variable maintenance costs $q\frac{dm}{dx}$. The amount $t\frac{dq}{dx}$ is the social gain from marginal generated traffic as a result of the unit increase in road standard. In the absence of a constraint, the optimal road standard is found where the sum of these benefits equals the marginal investment cost $\partial K/\partial x$.

Figure 2 illustrates 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, *t*, and, for simplicity, ignoring maintenance. (See ATC 2006, pp. 55-6 for a detailed explanation.) The welfare gain from marginal generated traffic is area A $\approx t\frac{dq}{dx}$. Area B $\approx -q\frac{dc}{dx}$ is the benefit to existing traffic from the fall in generalised costs. Figure 2 also shows $\partial c/\partial x$, which is the downward shift of the cost curve holding *q* fixed. It is different from dc/dx, which is the change in cost taking account of both the shift of the curve and the increase in *q*. In the case of optimal congestion pricing, equation (11b) shows that the area A equals area C. Hence the marginal benefit under optimal congestion pricing, given by equation (4b), is a special case of equation (17) ignoring maintenance costs, where the marginal benefit is areas A + B $\approx \hat{t}\frac{dq}{dx} - q\frac{dc}{dx} = -q\frac{\partial c}{\partial x} \approx$ areas B + C.

Additional investment in pavement strength reduces variable maintenance costs, which increases *q* via the demand curve, which in turn increases *c* on congested roads so that $dc/ds > 0$. On uncongested roads, there would be no change in *c* so $dc/ds = 0$. In the case of roughness, there is a trade-off between the higher *c* from increased roughness and a lower *m* from a lower maintenance standard.

We define the target MBCR with respect to *x* as $\mu_x^* = \left(t\frac{dq}{dx} - qG_x\right)/\frac{\partial K}{\partial x}$. Dividing both sides of equation (17), by $\frac{\partial K}{\partial x}$, $\mu_x^* - 1 = \lambda\frac{G_x}{\frac{\partial K}{\partial x}} = \lambda\left(\frac{dc}{dK} + \frac{dm}{dK}\right) = \lambda G_K$. In the same way, from equation (18), $\mu_s^* - 1 = \lambda G_K$, which shows that $\mu_x^* = \mu_s^* = \mu^*$, that is, the target MBCRs for road standard and pavement strength are identical.

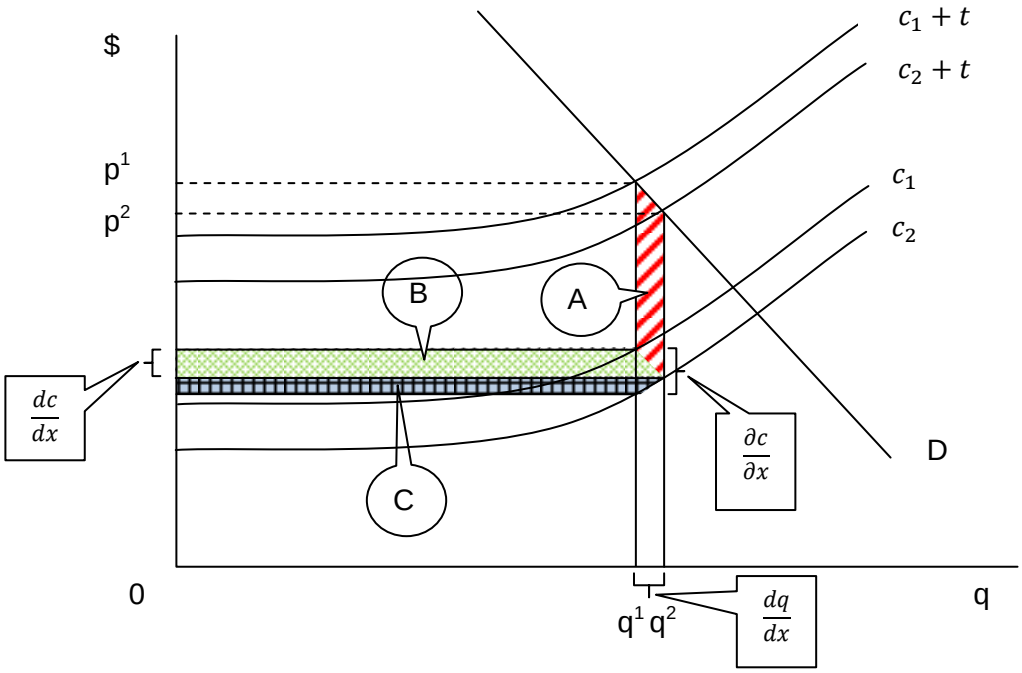
Substituting $\lambda G_x = (\mu - 1)\frac{\partial K}{\partial x}$ into (17) and $\lambda G_s = (\mu - 1)\frac{\partial K}{\partial s}$ into (18) gives the conditions for constrained welfare maximisation for *x* and *s*.

$$\frac{t}{\mu^*}\frac{dq}{dx} - \frac{q}{\mu^*}G_x - \frac{\partial K}{\partial x} = 0 \text{ and } \frac{t}{\mu^*}\frac{dq}{ds} - \frac{q}{\mu^*}G_s - \frac{\partial K}{\partial s} = 0 \quad (20a \text{ and } 20b)$$

---

distance-location charging of heavy vehicles is technically feasible and there are not the community acceptance difficulties faced by congestion pricing.

Figure 2: Welfare changes from a small downward shift in the average generalised cost curve



Unregulated commercial supply

An unregulated commercial supplier under the assumed pricing regime would maximise

$$\pi = (t+m)q - mq - K = tq - K \tag{21}$$

$$\frac{\partial \pi}{\partial x} = t\frac{dq}{dx} - \frac{\partial K}{\partial x} = 0 \; ; \quad \frac{\partial \pi}{\partial s} = t\frac{dq}{ds} - \frac{\partial K}{\partial s} = 0 \; ; \text{ and } \quad \frac{\partial \pi}{\partial r} = t\frac{dq}{dr} = 0 \tag{22a, 22b and 22c}$$

The unregulated supplier compares only the impact of marginal changes in generated traffic on revenue with marginal investment cost, ignoring the impacts on road users' generalised costs and variable maintenance costs. In figure 2, the supplier takes account only of area A, and ignores area B. Since the regulator passes increases in variable maintenance costs onto road users in the form of higher road damage charges, the supplier under-provides pavement strength because it is concerned only with the negative impact of reduced traffic on revenues.

Roughness is above optimum because the only benefit to the supplier from maintenance spending to reduce roughness is the increase in revenue from the effect of a lower value of *m* on quantity demanded. The benefit to existing road users is ignored.

Incentive regulation

The incentive regulated supplier's profit function is

$$\pi = [t^* - \psi(c - c^*) - \psi(m - m^*) + m]q - mq - K = [t^* - \psi(c - c^*) - \psi(m - m^*)]q - K \tag{23}$$

$$\frac{\partial \pi}{\partial x} = [t^* - \psi(c - c^*) - \psi(m - m^*)]\frac{dq}{dx} - \psi q G_x - \frac{\partial K}{\partial x} = 0 \tag{24}$$

$$\frac{\partial \pi}{\partial s} = [t^* - \psi(c - c^*) - \psi(m - m^*)]\frac{dq}{ds} - \psi q G_s - \frac{\partial K}{\partial s} = 0 \tag{25}$$

$$\frac{\partial \pi}{\partial r} = [t^* - \psi(c - c^*) - \psi(m - m^*)]\frac{dq}{dr} - \psi q G_r = 0 \tag{26}$$

At the socially optimal levels of investment and maintenance, $c = c^*$ and $m = m^*$ so the shadow toll becomes *t** and profit is zero. If $t^* = t$, $\psi = 1$ and $\lambda = 0$, then equations (24),

(25) and (26) are the same as equations (17), (18) and (19) respectively. Provided the base shadow toll equals the variable user charge and the target is optimal, incentive regulation causes profit maximising behaviour to correspond with welfare maximising behaviour. In figure 2, at the welfare optimum, a unit increase in road standard earns the supplier increased revenue of area A from generated traffic plus area B from the increase in the shadow toll applied to existing traffic.

In situations where $t^* \neq t$ and/or $\lambda \neq 0$ (hence $\mu^* \neq 0$), it is necessary to find the value of ψ that causes equations (24), (25) and (26) to equal zero at the same values of *x*, *s* and *r* as for equations (17), (18) and (19).

Since $p = c + t + m$ and *t* is fixed, $\frac{dp}{dx} = G_x$ and $\frac{dq}{dx} = \frac{dq}{dp}\frac{dp}{dx} = \frac{q}{p}\eta_D G_x$ (27)

where $\eta_D < 0$ is the price elasticity of demand.

Combining equations (20a) and equation (24), by eliminating $\frac{\partial K}{\partial x}$, and substituting (27)

$\psi = \left(t^* - \frac{t}{\mu^*}\right)\frac{\eta_D}{p^*} + \frac{1}{\mu^*}$ where $p^* = c^* + t + m^*$ is the generalised price at the target. (28)

The same result is obtained for pavement strength using equations (20b) and (25) and the pavement strength form of (27). For roughness, $\frac{\partial K}{\partial r} = 0$, so the MBCR cannot be defined. It can be demonstrated that the value of ψ that ensures $\frac{\partial \pi}{\partial r} = 0$ while equation (19) holds is the same as for value of ψ for road standard and pavement strength by combining equations (17), (18) and (19) with equations (24), (25) and (26) respectively. All three cases, *x*, *s* and *r*, give rise to the same expression $\psi = (t^* - t)\frac{\eta_D}{p^*} + 1 + \frac{\lambda}{q^*}$, which can be shown to equal equation (28), when the MBCR can be defined.

The unregulated profit curve, $\pi = t^* q - K$, and the incentive regulated profit curve intersect at zero profits at the target level of investment. However, when either or both $t^* \neq t$ and $\mu^* \neq 0$, the incentive regulated profit curve does not attain a maximum at that point. The supplier can earn above-zero profits by investing and maintaining at levels different from the target. The correction factor ψ shifts the incentive regulated profit curve so it reaches a maximum of zero profits where it intersects the unregulated profit curve to form an upside down Greek letter psi, as can be seen in the lower part of figure 1.

In the absence of any generated traffic, $\eta_D = 0$, ψ is the reciprocal of μ^*. With a target below the optimum, returns to the supplier from additional investment in the form of lowering users' generalised costs are too high, causing the profit curve to reach a maximum with above-zero profits at an above-target road standard. A correction factor below one lowers the incentive to invest to the point where maximum profits of zero are attained at the target level. With a target above the optimum, returns from additional investment are too low to induce the supplier to reach the target. The profit curve reaches a maximum with above-zero profits at a below-target road standard. A correction factor above one provides the additional incentive needed to induce the supplier to reach the target.

A non-unitary correction factor is required when $t^* \neq t$ because the social value of marginal generated traffic, *t*, differs from the private value, *t**, at the target. If $t^* > t$, the supplier is induced to exceed the target investment level, and conversely if $t^* < t$.

5. Incentive regulation in a network

Assume a set of *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 from the origin vector to the quantity vector, $(q_1', \cdots, q_n')$.[3] The welfare function is

$$W = \int_{0,\cdots,0}^{q_1',\cdots,q_n'} \sum p_i dq_i - \sum c_i q_i - \sum K_i \tag{29}$$

The vector of charges levied on road users, $(t_1, \cdots, t_n) = (p_1, \cdots, p_n) - (c_1, \cdots, c_n)$, is exogenously determined and need not be optimal. The condition for the optimal level of investment in segment 1 is

$$\frac{\partial W}{\partial x_1} = \sum p_i \frac{dq_i}{dx_1} - \sum c_i \frac{dq_i}{dx_1} - \sum q_i \frac{dc_i}{dx_1} - \frac{dK_1}{dx_1} = \sum t_i \frac{dq_i}{dx_1} - \sum q_i \frac{dc_i}{dx_1} - \frac{dK_1}{dx_1} = 0 \tag{30}$$

Equation (30), together with the partial derivatives for the other segments, constitute a set of simultaneous equations that could be solved to obtain the vector of optimal standards for all segments.

To interpret equation (30) (but not for comparing with equation (34) below), we make a substitution for $\sum q_i \frac{dc_i}{dx_1}$. Following the derivation of equation (11a) with optimal prices represented by $(\hat{t}_1, \cdots, \hat{t}_n)$, for all *i*, $q_i \frac{dc_i}{dx_1} = \hat{t}_i \frac{dq_i}{dx_1} + q_i \frac{\partial c_i}{\partial x_i} \frac{dx_i}{dx_1}$ (31)

where $\hat{t}_i = q_i \frac{\partial c_i}{\partial q_i}$. 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 $\partial W / \partial x_1$. Following the substitution, equation (30) becomes

$$\frac{\partial W}{\partial x_1} = \sum (t_i - \hat{t}_i) \frac{dq_i}{dx_1} - q_1 \frac{\partial c_1}{\partial x_1} - \frac{dK_1}{dx_1} = 0 \tag{32}$$

Equation (32) proves the axiom in cost–benefit analysis that, when prices in markets for substitutes or complements are at optimal levels (marginal social costs), $t_i = \hat{t}_i$ for any $i \neq 1$, changes in these markets due to shifts in demand curves between the base case and project case 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 marginal willingness-to-pay and marginal social cost 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. (See Harberger 1972, pp. 261-3 and ATC 2006 pp. 66-75 for expositions.)

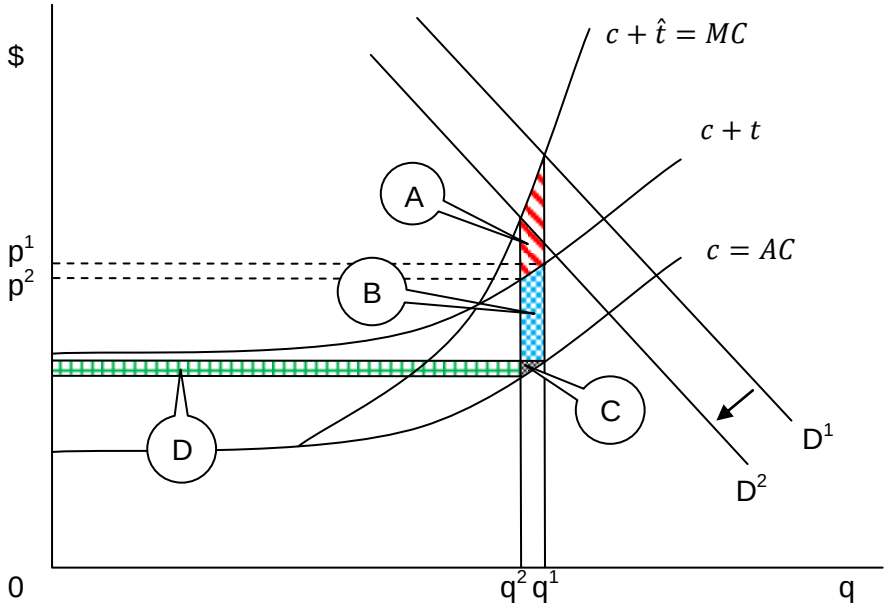
Figure 3 illustrates the effects of a small leftward shift in the demand curve for $q_{i \neq 1}$ as a result of a price fall in a substitute segment in the network q_1 due to a road improvement. The charge *t* is below the optimal level, $\hat{t}$. If the optimal price, $\hat{t}$, had been set, the loss of

3. By way of explanation, say the path between any two quantity vectors consists of a series steps in which only one quantity q_j is changed at a time. The other quantities, $i \neq j$, are held constant. Each quantity has an associated demand surface $p_j = p_i(q_1, \cdots, q_j, \cdots, q_n)$. For each step, the change in WTP is measured by integrating under the demand surface for the particular quantity that is changed, q_j, along the direction of the change. All the other demand surfaces are ignored, though they shift as q_j changes. Hence, we measure the area under the demand curve for q_j over the change in q_j. For the next step along the path, a change in quantity *k*, we take the area under the demand curve for q_k, notwithstanding the fact that the demand curve for q_k has shifted in response the previous change in q_j. The WTPs for all the steps along the path are added together. A path that is a smooth curve can be treated as a series of infinitesimally small steps, one quantity change at a time. The total WTP change between any two quantity vectors will differ depending on the particular path followed and the order of the steps taken unless the 'condition for path independence of line integrals' holds for the demand functions, that is $\partial p_i / \partial q_j = \partial p_j / \partial q_i$ for all $i \neq j$. For an infinitesimally small change in q_j, the change in WTP is the height of the demand surface above the value of q_j, that is, $\partial WTP / \partial q_j = p_j$.

WTP, would exactly equal the saving in resource costs $p\frac{dq}{dx_1} = MC\frac{dq}{dx_1} = (c+\hat{t})\frac{dq}{dx_1}$, leaving no net change in welfare. With the charge set below the optimal level, $t < \hat{t}$, as shown in figure 3, the loss of WTP $p\frac{dq}{dx_1} = (c+t)\frac{dq}{dx_1}$ is less than the saving in resource costs, $MC\frac{dq}{dx_1} = (c+\hat{t})\frac{dq}{dx_1}$, leading to a welfare gain of $(p-MC)\frac{dq}{dx_1} = (t-\hat{t})\frac{dq}{dx_1} \approx$ area A. Had the charge been set above the optimal level, $t > \hat{t}$, a welfare loss would have ensued.

In terms of equation (30), the welfare change is $t_i\frac{dq_i}{dx_1} - q_i\frac{dc_i}{dx_1} \approx$ – areas B+C + areas D+C, which equals areas A because for $i \neq 1$, areas D+C $\approx q_i\frac{dc_i}{dx_1} = \hat{t}_i\frac{dq_i}{dx_1} \approx$ areas A+B+C (see equation (31)).

Figure 3: Welfare changes for a competing road segment from a small leftward shift in the demand curve



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

$$\pi = \sum[t_i^* - \psi_i(c_i - c_i^*)]q_i - \sum K_i \qquad (33)$$

Differentiating with respect to x_1, then setting $c_i = c_i^*$ for all segments *i* and the derivatives equal to zero so profits will the maximised at the target K_1*

$$\frac{\partial\pi}{\partial x_1} = \sum t_i^*\frac{dq_i}{dx_1} - \sum \psi_i q_i\frac{dc_i}{dx_1} - \frac{dK_1}{dx_1} = 0 \qquad (34)$$

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 (34) is identical to the welfare maximising condition, equation (30), provided $t_i^* = t_i$ and $\psi_i = 1$ for all *i*. In terms of figure 3, as a result of a unit increase in road standard in segment 1, the supplier gains $q_i\frac{dc_i}{dx_1} \approx$ areas D+C in revenue from the increase in the shadow toll for segment *i* following the fall in average generalised cost, and loses $t_i^*\frac{dq_i}{dx_1} \approx$ areas B+C due to the reduction in traffic. These are the same as the welfare changes, which were shown above to equal area A. With optimal congestion pricing, $t_i^* = t_i = \hat{t}_i$, the supplier's gain in revenue, areas D+C would equal the loss of revenue, areas A+B+C. The

shift in the demand curve would be revenue neutral to the supplier under incentive regulation as well as welfare neutral to society.

To maximise constrained welfare assuming an identical target MBCR $= \mu^*$ for all segments

$$\sum \frac{t_i}{\mu^*} \frac{dq_i}{dx_1} - \sum \frac{q_i}{\mu^*} \frac{dc_i}{dx_1} - \frac{dK_1}{dx_1} = 0 \quad (35)$$

Combining equations (34) and (35) by eliminating $\frac{dK_1}{dx_1}$

$$\sum \left(t_i^* - \frac{t_i}{\mu^*}\right) \frac{dq_i}{dx_1} - \sum \left(\psi_i - \frac{1}{\mu^*}\right) q_i \frac{dc_i}{dx_1} = 0 \quad (36)$$

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.

For a network of road segments with inter-related demand curves, with the parameters correctly set by regulator, the profit maximising vector of road standards under incentive regulation will be the same as the welfare maximising vector constrained by a target MBCR.

It is essential that all the road segments with related demands be supplied by a single entity except in the case where there is optimal congestion pricing and unitary correction factors (so shifts in demand curves for related segments at the optimum are revenue neutral). Where there is non-optimal pricing, incentive regulation requires the supplier to include revenue changes on substitute and complement road segments in its financial analyses of investment decisions. Hence, incentive regulation would not work well if applied to a single toll road that competes for traffic with unpriced roads in a network.

6. Some practical aspects of incentive regulation

Measuring generalised costs

The target and actual generalised cost levels used to set shadow tolls would be estimates derived from computer models—the same models used to undertake cost–benefit analyses of road projects and to set congestion prices. These models 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, annual hourly volume distribution, directional splits) and include speed–flow relationships. Under incentive regulation, the models, parameters and data collection methods to estimate generalised costs would be included in the legal agreements between the regulator and the supplier.

As a transitional step to incentive regulation, performance indicators based on weighted average social generalised costs could be introduced for government road agencies for networks, sub-networks, areas and corridors. The average social generalised cost measure combines a wide range of characteristics of the output of road agencies (capacity, congestion, roughness, alignment, safety, environmental externalities) into a single number. The weights accorded to the different components reflect community valuations. Because of the central role they play in economic appraisal and price setting, average social generalised costs, as performance indicators, should be valuable aids to strategic planning.

Averaging across vehicle types and time periods for the same segment

For any road segment, average generalised costs vary between time periods with different demands and between vehicle type categories. Computer models estimate the generalised cost for each vehicle type and time period separately and then combine them to obtain a total cost for all vehicles over a whole year.

Equations (33) and (34) illustrate the situation after some modifications. The subscript *i* has to be redefined to refer to vehicle types in time periods. The variables *K*, *x* and *ψ* have no subscripts because there is a single road segment with a single road standard under consideration.

The weighted average target generalised cost per vehicle-kilometre could be defined as $\bar{c}^* = \sum c_i^* \, q_i^* / \sum q_i^*$. Since $\sum (c_i - \bar{c}^*) q_i^* = \sum (c_i - c_i^*) q_i^*$, use of the weighted average *c** has no effect on profits.

Alternatively, weighted average target generalised cost per vehicle-kilometre could be defined as $\bar{c}^* = \sum c_i^* \frac{dq_i}{dx} / \sum \frac{dq_i}{dx}$. Since $\sum (c_i - \bar{c}^*) \frac{dq_i}{dx} = \sum (c_i - c_i^*) \frac{dq_i}{dx}$, use of the derivative-weighted average *c** has no effect on the profit maximising level of investment.

The two sets of weights will be same if the proportions of vehicles of each type in each time period stay the same as the road standard is improved. In most cases, this is a reasonable assumption. On congested roads, peak narrowing as capacity is expanded would cause the proportions in different time periods to change. In such cases, the target weighted average *c** could be set using derivative weights (dq_i/dx) to ensure the correct profit maximising level of investment, and the effect on profits at the target countered by adjusting *t**. Alternatively, traffic weights (q_i^*) could be used along with a correction factor to ensure the first derivative of the profit curve equals zero at the target level of investment.

Being able to have a single weighted average generalised cost target for each road segment simplifies practical application of incentive regulation.

In the model above, variable maintenance charges and targets were expressed per vehicle. In practice, they would be expressed per equivalent-standard axle load (ESAL) kilometre, so averaging across vehicle types would be necessary only for setting correction factors.

Averaging across road segments

It is desirable to aggregate segments so they have a single average cost target, first, to simplify, second, to smooth out indivisibilities, and third, to have the regulator setting parameters at a strategic level rather than for individual road segments. Averaging targets across segments means that, at optimal investment levels, although profits from the group of segments together will be zero, roughly half the individual segments will make losses and the other half profits. So there must be no question of the supplier closing down or disposing of loss-making segments.

Care is needed when averaging target generalised cost levels across segments because the weighted average *c** value in the shadow toll formula for each individual segment is not the optimal value. There is no problem where generated and diverted traffic are negligible because with $dq/dx = 0$, *c** disappears from the first derivative of the profit function, $\pi = [t^* - \psi(c - c^*)]q - K$, $\frac{d\pi}{dx} = -\psi q \frac{dc}{dx} - \frac{dK}{dx}$. For many non-urban road segments with no viable alternative routes, an assumption of zero generated and diverted traffic in response to changes in road standard over the relevant range might be reasonable, permitting widespread aggregation and averaging.

Where there is significant generated or diverted traffic, the supplier will be induced to invest at above optimal levels on segments with below average target generalised costs, and below optimal levels on segments with above average target generalised costs. Investment patterns are distorted to shift traffic away from less remunerative segments towards more remunerative segments. Hence only segments with similar *c** values can be grouped together. By specifying bands of *c** values for grouping segments, with a single target *c** for each band, a considerable amount of aggregation should be possible without the actual *c** for any individual segment being very different from the group *c**.

An alternative approach would be to aggregate segments regardless of *c** but to apply different correction factors to groups of segments with differing targets to offset the tendencies to over- or under-invest. Larger correction factors would be needed for segments with above average generalised costs and conversely for segments with below average generalised costs. The formula is

$$\psi_i = \frac{\left(t_i^* - \frac{t_i}{\mu^*}\right)\frac{\eta_{Di}}{p_i^*} + \frac{1}{\mu^*}}{\left[\left(c_i^* - \bar{c}^*\right) + \left(m_i^* - \bar{m}^*\right)\right]\frac{\eta_{Di}}{p_i^*} + 1} \tag{36}$$

This is equation (28) over a denominator that includes $c_i^* - \bar{c}^*$, the difference between the *c** for the individual road segment *i* and the weighted average *c** for the group of segments, and likewise for variable maintenance costs.

Indivisibilities and new roads

If the target investment level lies within an indivisibility, the supplier would face a choice of making a loss from under-investing or loss from over-investing. The regulator should avoid setting the target road standard within an indivisibility. As already noted, aggregating segments smooths out indivisibilities.

The regulator can set parameters for roads that have not been built, on its own initiative or at the request of the supplier.

Leads and lags

The assumption of malleable capital implies instantaneous adjustment. Major road investments can take years to appraise, plan, design and construct. Parameters in the shadow toll formula could be set years advance as trajectories, with jumps at times when large investment projects are expected to be completed. The parties could negotiate for the supplier to undertake a specific investment project in exchange for agreed changes in the base shadow toll and target generalised cost to occur after completion of the project. Target generalised cost levels would have to change over time with changes in vehicle operating costs, the mix of vehicle types, and congestion due to traffic growth, less an adjustment for economically warranted investments. Annual adjustments might be made for inflation and improvements in efficiency as in price cap regulation.

7. Conclusion

The proposed incentive regulation scheme uses performance-based financial incentives to control investment and maintenance levels by a public utility or private road supplier. Performance is measured by weighted average social generalised costs of road use. The supplier is free to determine, using commercial criteria, which investment projects and maintenance works it will undertake to achieve its targets.

A regulator determines the levels of road user charges, sets the performance targets and measures performance. It pays revenues raised from road users into a fund out of which it remunerates the supplier by paying a shadow toll determined by a formula. At the target levels of investment and maintenance, the shadow toll is just sufficient for the supplier to fully recover costs and earn a normal return on capital. To the extent that actual investment and maintenance fall short of the target levels, the shadow toll is reduced.

It has been shown that, under assumptions of perfect information, perfect divisibility, malleable capital, identical private and social discount rates, and correct setting of parameters by the regulator, profit maximising outcomes under incentive regulation, exactly match welfare maximising outcomes. Welfare is still maximised, in a second-best sense, when road user charges are not at optimal levels. The scheme is equally applicable to congested urban roads and to low volume rural roads. It can deal with both the congestion–capacity and the pavement damage–strength dimensions of road supply. In a network of

road segments with inter-related demand curves, the profit maximising and welfare maximising sets of road standards are the same. The regulator can engineer above- or below-optimum investment and maintenance outcomes if desired.

The scheme has been developed at a highly conceptual level. There are many variations, elaborations and details to consider. Much more research is required into the theoretical and practical aspects before any incentive regulation scheme for road supply becomes a realistic proposition. In particular, the implications of relaxing the assumptions need consideration, that is, how the scheme could work in the presence of indivisibilities, risk and uncertainty, asymmetric information between the regulator and the supplier, and dynamics. Some of the issues have already been addressed in the regulatory economics literature. In the meantime, performance indicators based on weighted average social generalised costs can be introduced for road suppliers, and would serve as a transitional measure for introducing incentive regulation.

If incentive regulation can be successfully translated from theory into practice, it will expand the scope for commercial road supply without compromising efficient resource allocation. In doing so, it could offer solutions to some of the pressing issues in road economics faced in many countries.

References

Australia's Future Tax System 2009, Report to the Treasurer, Part Two, Detailed Analysis, December.

Australian Transport Council 2006, *National Guidelines for Transport System Management in Australia, volume 5, Background material*, ATC, Canberra.

Bureau of Infrastructure, Transport and Regional Economics 2008, *Moving urban Australia: can congestion charging unclog our roads?*, Working paper 74, Canberra.

Dupuit, J. 1844, *On the Measurement of the Utility of Public Works*, translated from the French *Annales des Ponts et Chaussées*, 2nd series, Vol. 8, by R.H. Barback 1952 for *International Economic Papers*, No. 2, pp. 83–110, reprinted in Munby, D. (ed.) 1968, *Transport: Selected Readings*, Penguin, Harmondsworth, Middlesex, pp. 19–57.

Harberger, A.C. 1972, *Project Evaluation: Collected Papers*, Macmillan, London.

Hau, T.D. 1992, *Economic Fundamentals of Road Pricing: A Diagrammatic Analysis*, World Bank, Policy Research Working Paper: Transport, WPS 1070.

Newbery, D.M. 1989, Cost Recovery from Optimally Designed Roads, *Economica*, New Series, 56(222), May, pp. 165-185.

Mohring, H. and M. Harwitz 1962, *Highway Benefits: An Analytical Framework*, Northwestern University Press, Evanston, Illinois.

Mohring, H. and Verhhoef, E.T. 2007, *Self-financing roads*, Tinbergen Institute Discussion Paper, TI2007–068/3.

Productivity Commission 2006, *Road and Rail Freight Infrastructure Pricing*, Inquiry Report 41, December.

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

Walters, A.A. 1968, *The Economics of Road User Charges*, International Bank for Reconstruction and Development, World Bank Staff Occasional Papers No. 5, John Hopkins Press, Baltimore.