Price-cap regulation

Simon Cowan[*]

Summary

■ This paper assesses the theory and practice of price-cap regulation for utilities. Price caps fix the time paths of prices and thus provide incentives for cost efficiency. Agency theory suggests that there is a trade-off in general between providing incentives for efficiency and extracting rents. A pure price cap is optimal when producer and consumer interests have equal weight in the regulator's objective function. More generally the use of yardstick competition and franchises or competition for the market can alleviate the regulatory trade-off. The paper also discusses how much pass-through of costs there should be, which depends on the risk aversion of the consumer and the firm and on whether the costs are endogenous or exogenous. Issues involved in the setting of price caps are discussed, including the lag between formal price reviews, quality incentives, multi-product price regulation, access pricing and price caps and competition. The key role of regulatory commitment in promoting long-term investment incentives is highlighted. Finally the small literature on the empirical evidence for the effects of price-cap regulation is discussed. The conclusion for policymakers is that while the experience of price-cap regulation has generally been favourable there remains concern about whether investment can be promoted under such a system of regulation.■

JEL Classification: D42, L51, L98.
Keywords: Utility regulation, price-cap regulation, incentive regulation, monopoly.

[*] *Simon Cowan is university lecturer in economics at the Department of Economics at Oxford University.*

Price-cap regulation

Simon Cowan[*]

Price-cap regulation is an innovation in regulatory policy that was developed in the 1980s and has been applied around the world. In this paper I assess the theory and practice of price-cap regulation. The general problem is as follows. A firm with partial or complete market power must be regulated to prevent it from abusing its dominant position through excessive pricing. An unregulated firm can be expected to set its prices to maximize profits, leading to both deadweight losses and transfers of purchasing power from consumers to the firm, both of which are costly to the regulator. At the same time the regulator wants to encourage the firm to be efficient. A price cap tackles these problems by the very straightforward solution of fixing the firm's price (or the price path over time). The firm thus bears the risks associated with varying exogenous input prices and shifting demand. At the same time the firm has full incentives to reduce its costs, as the price is not adjusted downwards when it succeeds in cutting its own costs. The key feature of a price cap is that the price level is not responsive to anything. The standard alternative form of regulation is based on capping the firm's realized rate of return on capital and is known as rate-of-return regulation or cost-of-service regulation. With rate-of-return regulation prices track observed costs closely, and thus the firm faces lower risks than under a price cap and is unlikely to earn excess profits for long. The downside is that the firm has little incentive to pursue efficiency gains because the profitability of such activity is reduced by the expectation that prices will be cut by the regulator in response.

Price-cap regulation (PCR) in practice is, of course, not as straightforward as simple theory would suggest. In particular price caps do not last indefinitely—in the UK they are revised after four or five years, in the US the lag between price reviews is typically rather shorter at around two years. At the price review the regulator will typically seek to recover in lower prices for consumers some or all of

[*] *I would like to thank an anonymous referee, the editors and Tore Ellingsen for helpful comments on earlier versions of this paper.*

the efficiency gains that the firm has made in the previous period, and aim to ensure that the firm does not earn an excessive rate of return in the future. At first glance, then, PCR can look rather similar to rate-of-return regulation.

There are, though, significant differences between the ways in which the two systems of regulation operate and in their outcomes. As well as having a long lag between price reviews, PCR is characterized by certainty about when the price level will be reviewed. The regulator commits during the period before the next price review to avoid interfering. For example suppose that in the second year of a five-year price cap the regulator acquires new information showing that the firm has managed to make larger efficiency gains than were anticipated at the previous review and is thus earning excess profits. Under PCR the regulator cannot use this information until the next price review, and must allow the firm to continue to earn the high profits allowed under the regulatory contract. Without this commitment the incentives for efficiency would be reduced or even eliminated. This type of commitment is very difficult for politicians to make, so an important condition for PCR to operate is that the regulatory agency is independent of the government and has a clear set of duties while remaining ultimately publicly accountable. There is a close analogy between the frameworks within which independent regulatory agencies and independent central banks operate.

The commitment goes both ways. A firm facing PCR bears the risk that adverse external shocks will occur. In the case of the Californian electricity system in 2000 and 2001 the regulated distributors had a price cap for sales to retail customers, an obligation to supply and a requirement to purchase all their power in the spot wholesale market.[1] When the price in the spot wholesale market rose well above the retail price the distributors were quickly bankrupted and the state had to enter the market. In the energy sector this type of problem is usually dealt with in advance by allowing the firm to pass through specified costs such as the costs of purchasing fuel or wholesale power. This type of pass-through is consistent with PCR as long as the costs that are passed through are outside the firm's control. Of course if all costs are passed through then there are no incentives for efficiency.

[1] For analysis of the problems in California see Joskow (2001) and Borenstein (2002).

One complication with PCR is that by providing strong incentives for cost reduction it can also lead to a degradation of quality. Price caps are typically supplemented by additional controls on quality levels. Another complication is that regulated firms typically sell many products and operate in several distinct markets. In keeping with the philosophy of decentralizing decisions as far as possible a firm under PCR often has a very large degree of flexibility about how to choose relative prices as long as the average price, measured by a price index, satisfies the cap. Finally when it is feasible and desirable to let the dominant firm face competition PCR appears to be helpful in ensuring that entry can occur without distortion.

1. The origins of price-cap regulation

The first application of PCR in the UK was for contraceptive sheaths in 1982, which at the time were produced by an unregulated monopoly (Monopolies and Mergers Commission, 1982, Armstrong et al., 1994, p 166). To eliminate the problem of excessive pricing the Monopolies and Mergers Commission recommended that the annual rate of price increase for sheaths should be no more that the growth rate of a cost index less 1.5 per cent and that this control should apply for five years. The cost index was outside the control of the regulated firm.

The main catalyst for the development of PCR for utilities in the UK was the report by Professor Stephen Littlechild to the Department of Industry in 1983 on the regulation the profitability of British Telecom (BT) once it was privatized (Littlechild, 1983). Littlechild argued that the focus of regulation should be on prices and not profits, and that US-style rate-of-return regulation generated poor incentives for cost efficiency and encouraged regulatory capture. He recommended PCR for prices in the markets where BT faced no competition. An index of prices is allowed to grow by at most the rate of growth of the Retail Prices Index, which is the main measure of consumer price inflation, less a predetermined factor initially set at 3 per cent for BT and known as the *X* factor (or the productivity offset in some countries). PCR is also known as RPI-*X* regulation. Thus real prices had to fall by at least 3 per cent each year. The *X* factor remained fixed and was independent of observed costs and profits during the five-year period before the cap was reviewed.

Price-cap regulation has since been applied in the UK to gas, airports, water, electricity and the railways. As competition in telecommunications, electricity and gas has developed price caps in retail supply have been eliminated. Price caps remain in place, however, for the network services parts of these industries where complete or substantial market power is still present. Other countries have followed a similar pattern. Typically telecommunications is the first industry to be subjected to PCR. In the US a price cap was imposed on the dominant long-distance carrier, AT&T, in 1989 by the Federal Communications Commission. Regulation of local telecommunications in the US is done at the state level, and Table 1, from Ai and Sappington (2001) shows the recent rapid move to PCR.[2] Many other countries around the world have implemented PCR for utilities.

Table 1. Number of states employing different types of regulation for telecommunications

Year	Rate of return regulation	Rate case moratoria	Earnings sharing regulation	Price-cap regulation	Other
1985	50	0	0	0	0
1986	45	5	0	0	0
1987	36	10	3	0	1
1988	35	10	4	0	1
1989	31	10	8	0	1
1990	25	9	14	1	1
1991	21	8	19	1	1
1992	20	6	20	3	1
1993	19	5	22	3	1
1994	22	2	19	6	1
1995	20	3	17	9	1
1996	15	4	5	25	1
1997	13	4	4	28	1
1998	14	3	2	30	1
1999	12	1	1	35	1

Source: Ai and Sappington (2001), Table 1.

[2] Rate case moratoria are similar to price caps since they involve freezing prices until a new regulatory system is established. Earnings sharing regulation is a hybrid of rate-of-return regulation and PCR that allows the firm's earnings to vary freely within a band but as earnings move outside the band there is some sharing with consumers.

2. The principles of price-cap regulation

In this section I outline the theory behind price-cap regulation. For comprehensive analyses of price regulation see Laffont and Tirole (1993), Armstrong et al. (1994) and Armstrong and Sappington (2002). We address a number of questions:

- How much cost pass-through is optimal?
- How should quality be regulated?
- When the firm sells many products how should relative prices be regulated?
- If the firm is vertically integrated and supplies network services as an input for rivals how should its prices be regulated?
- What effect does competition have on price-cap regulation?

2.1. Incentives versus rent extraction

In the main model of price regulation there is a trade-off between providing incentives and reducing excess profits (or rents).[3] I shall illustrate the trade-offs in regulatory design, characterize the general solution to the regulatory problem and find the conditions under which pure PCR is optimal. The model used here is a version of that presented in Cowan (2001). The main point is that the regulator has less information about the components of costs than the firm, and thus optimal regulation should take full account of the firm's information advantage.

The consumer buys one unit of the product and wants the price, p, to be as low as possible. Consumer surplus is $U - p$ where U is gross utility. Total cost has three parts. First, there is a random factor, θ, whose value is known by the firm but not by the regulator. The regulator does, however, know the distribution from which θ is drawn. The highest value that θ can take is denoted by $\overline{\theta}$ and the expected value of θ by $E[\theta]$. Second, when the firm puts in an effort level of e its cost falls by e. The regulator cannot observe effort. The difference between the first two elements of cost, $c = \theta - e$, is the accounting cost and is observed by the regulator, though neither component is known. Finally when putting in effort of e the firm incurs the unobserved effort cost $\phi e^2/2$, which implies that the marginal cost of effort

[3] Laffont and Tirole (1986, 1993) developed the general model.

is increasing (when $\phi > 0$). Total cost is thus $TC = \theta - e + \phi e^2/2$. The firm's profit or rent is $\Pi = p - TC = p - \{\theta - e + \phi e^2/2\}$.

The regulator's objective function is a weighted sum of consumer surplus and profit, $W = U - p + \alpha\Pi$. Here α, which satisfies $0 < \alpha \leq 1$ is the weight on profits. The regulatory objective may be rewritten as

$$W = U - TC - (1-\alpha)\Pi. \tag{1}$$

When $\alpha < 1$, so the regulator puts less weight on shareholders' interests than on consumers, equation (1) shows that the regulator wants to eliminate excess profits as well as to ensure that total costs, TC, are minimized. If the regulator had complete information there would be no problem in achieving both objectives. The firm would be told to set the effort level that minimizes TC and would in return receive a price that just covers total costs, so $p = TC$, $\Pi = 0$, TC is as low as possible and W is maximized.

With an asymmetry of information, however, it is usually not possible to achieve both objectives. Suppose that the regulator sets the price as a linear function of the (observed) accounting cost, i.e.

$$p(c) = a + bc, \tag{2}$$

where the coefficients a and b are chosen by the regulator. b is the coefficient of cost pass-through. When $b = 0$ a pure price cap operates. Alternatively when $a = 0$ and $b = 1$ there is pure rate-of-return regulation, as the price tracks accounting cost one-for-one. The higher b is in general the more price tracks cost and thus the less incentive there is for the firm to invest in effort.

When choosing a and b to maximize the objective given in (1) the regulator takes account of two constraints. First, the firm must be willing to participate. This is guaranteed as long as the firm covers its costs even when the highest possible value of θ is realized.[4] When the firm draws a value of θ that is below $\overline{\theta}$ it earns a positive profit. The expected value of the firm's profit is thus positive. This rent arises

[4] Note that this constraint requires that the firm is willing to participate in each state of the world rather than on average. Such an *ex post* constraint is appropriate when θ, once drawn, does not change.

because of the information asymmetry. The second constraint is that the regulator must take account of the firm's incentive to choose *e* to maximize its profit given any particular price regulation scheme defined by equation (2). The marginal benefit of effort to the firm is 1-*b*, which represents the cut in the accounting cost of 1 less the resulting price reduction as *c* falls. The marginal cost of effort is ϕe so (privately) optimal effort is $e^* = (1 - b)/\phi$. Note that effort declines as the pass-through coefficient rises and as ϕ increases.

The optimal pass-through coefficient is

$$b^* = (1-\alpha)\left[\overline{\theta} - E(\theta)\right]\phi. \tag{3}$$

Equation (3) implies that as long as consumer interests are weighted more heavily than the firm's, there is some variation in θ, and the marginal cost of effort provision is strictly positive, the optimal degree of cost pass-through is positive. If any one of these conditions does not hold then there should be no cost pass-through—in other words a pure price cap is optimal. For example when $\alpha = 1$ equation (1) implies that rent extraction is not a regulatory concern. Cost minimization is then the only objective so full incentives for cost efficiency are provided, through a price cap.[5] When $\overline{\theta} = E[\theta]$ there is full information, so there is no need to relate *p* to *c*. Finally when $\phi = 0$ the marginal cost of effort to the firm is zero, so providing incentives for cost reduction is costless to the regulator and the incentives should be as high-powered as possible.

When a pure price cap is used effort is at the full-information level $1/\phi$. The firm's total cost in the worst case is $TC = \overline{\theta} - 1/2\phi$, so to ensure participation the regulator sets *a* to just cover all costs when $\theta = \overline{\theta}$. This yields a rent to the firm of $\Pi = \overline{\theta} - \theta$, and thus expected profits with a price cap are $E[\Pi] = \overline{\theta} - E[\theta]$. For an analysis of how to set a price cap in practice see Bernstein and Sappington (1999).

There are two options for the regulator who wants to improve the trade-off: yardstick competition and franchising. Yardstick competi-

[5] Loeb and Magat (1979) present a model of regulation with asymmetric information where the full information outcome is achieved even though information is asymmetric because the regulator values consumer and producer surplus equally.

tion entails comparing the costs of similar firms and using the correlation in the information to improve regulation. The typical application of yardstick competition is for a regionally separated industry (such as electricity distribution or water supply) that is nationally regulated. Shleifer (1985) has the canonical model of yardstick competition. Suppose there are two firms, 1 and 2, and that their exogenous cost parameters, θ_1 and θ_2, are perfectly correlated. The fact that the cost parameters are perfectly correlated allows the regulator to achieve the full-information outcome of full productive efficiency and zero rents. To see this note that if firm 1 has a higher observed accounting cost than 2, the regulator immediately knows that the effort level of firm 1 is lower than that of firm 2 by the difference in costs c_1-c_2. By cutting p_1 by exactly this amount the regulator eliminates the incentive for firm 1 to reduce its effort below that of firm 2. The way to implement the full-information outcome is to use the following pricing rules:

$$p_1 = \frac{1}{2\phi} + c_2;\, p_2 = \frac{1}{2\phi} + c_1. \tag{4}$$

By decoupling the price that each firm receives from its own cost the regulator provides an incentive for each firm to put in the full-information effort level $1/\phi$. In the Nash equilibrium of the game the accounting costs are identical. The term $1/2\phi$ in (4) covers the unobserved cost to the firm of putting in optimal effort, and each firm thus just breaks even. Thus the full-information outcome obtains. If the exogenous cost parameters shocks are less then perfectly correlated the regulator can set the price as a linear function of the two accounting cost levels and the trade-off between rent extraction and cost reduction is partially improved relative to the case when there is no correlation so comparative information is of no use.

Franchising is a method of auctioning the right to supply a monopoly service.[6] It has been used for train operating services in the UK, for the water sector in France and for infrastructure projects in many developing countries. Because an incentive contract generates expected rents firms compete to earn the prize, and this competition dissipates the excess profitability. In a Chadwick-Demsetz auction potential suppliers compete over the price that they are prepared to

[6] For recent models of franchising see Engel et al. (2002) and Caillaud and Tirole (2001).

charge customers (and perhaps also over quality). The lowest price (or the best price-quality package for consumers) wins. Suppose that each firm knows the values of both its own and the other firm's θ variable and that $\theta_2 > \theta_1$, while the regulator remains ignorant of these parameters. The regulator fixes a value of *b* in advance. For simplicity suppose that $b = 0$ so there is no cost pass-through. The winning firm is the one that offers the lower fixed part of the price function, *a*, and thus the lower price to consumers. Clearly firm 1 will win the franchise, and the competition between it and firm 2 has exactly the same outcome as an asymmetric cost Bertrand game. Each firm chooses $e = 1/\phi$ and incurs effort costs of $1/2\phi$ (in the event of winning). Firm 2's profit is $\Pi_2 = a_2 - \theta_2 + 1/2\phi$, so its lowest possible bid is $a_2 = \theta_2 - 1/2\phi$. Firm 1 knows this, and will set its bid fractionally below a_2, giving it profits of $\Pi_1 = \theta_2 - \theta_1$. The key point is that compared to the case of a pure price cap the *expected* rent of the winning firm is lower when there is an auction. By auctioning the contract the regulator can achieve the same level of efficiency as with a pure price cap but lower rents. More generally the regulator can ask the firms to bid on both *a* and *b* and the competition between them improves the regulator's trade-off. The greater the number of firms bidding the better the outcome for the regulator, since the expected gap between the lowest two θ levels will be smaller. Similarly the more correlated are θ_1 and θ_2 the smaller the expected gap between them and thus the smaller the rent for the winner.

2.2. Incentives versus insurance

An alternative model when there is asymmetric information also has the firm having private information about its effort, but neither side knows the random factor affecting accounting costs. The model is a standard principal-agent one where there is a trade-off between providing incentives to the firm and offering it insurance. Here I discuss a version of the model in Armstrong et al. (1994, Chapter 2) that is generalized in two directions. First, I allow the consumer, as well as the firm, to be risk averse in order to find how relative attitudes towards risk determine how prices move. Second, we distinguish between exogenous and endogenous costs. Many price caps in practice do not completely stabilize the price that consumers face but allow some specified exogenous costs to be passed through to consumers. Utilities supplying electricity or natural gas typically face volatile input prices (the wholesale price of electricity or underlying fuel prices) and some cost pass-through is allowed, though the Californian case was an infamous exception.

Again the consumer has perfectly inelastic demand for the product. The consumer's utility is

$$V_c = -E[p] - \frac{\delta}{2} Var(p),$$

so price volatility is disliked when the coefficient of absolute risk aversion, δ, is positive. The firm's (observed) accounting cost is

$$c = c_1 + c_2 = \theta - e + c_2,$$

where $c_1 = \theta - e$ is observable, but not its components, while c_2 is *ex post* observable and is exogenous to the firm. θ and c_2 are random variables with means of $E[\theta]$ and $E[c_2]$, variances σ_1^2 and σ_2^2 and covariance σ_{12}. The regulator sets price as a linear function of the two observable parts of marginal cost

$$p = a + b_1 c_1 + b_2 c_2. \tag{5}$$

Profit is $\Pi = a + (b_1 - 1)c_1 + (b_2 - 1)c_2 - e^2/2$, which is the price less the accounting cost and the cost of effort term (we have set $\phi = 1$ here to simplify the notation). The firm's utility function is

$$V_f = E[\Pi] - \frac{\gamma}{2} Var(\Pi) - \frac{e^2}{2}$$

which is a standard mean-variance utility function. When the absolute risk aversion coefficient, γ, is positive the firm prefers to be insured against profit variation.

Faced with any given incentive scheme defined by (5) the firm chooses e to maximize its utility. The resulting level of effort is $e = 1 - b_1$. The regulator takes account of this relationship between effort and the pass-through parameter b_1 and chooses a, b_1 and b_2 to maximize the sum of the two utility functions subject to the constraint that the firm's utility equals a reservation level. Note, though, that the participation constraint here is an *ex ante* one—the firm participates as long as expected utility meet the reservation level. Define *r* as the correlation between θ and c_2: $r \equiv \sigma_{12}/\sigma_1 \sigma_2$. The general expressions for the two pass-through coefficients are

$$b = \frac{\gamma\sigma_1^2\left(1-r^2\right)}{1+\left(\gamma+\delta\right)\sigma_1^2\left(1-r^2\right)}, \tag{6}$$

and

$$b_2 = \frac{\gamma}{\gamma+\delta} + \frac{\gamma}{\left(\gamma+\delta\right)} \frac{\left(\sigma_{12}/\sigma_2^2\right)}{1+\left(\gamma+\delta\right)\sigma_1^2\left(1-r^2\right)}. \tag{7}$$

To interpret these suppose first that $\sigma_{12} = 0$ so $r = 0$ and the exogenous cost element provides no information to the regulator about the value of the random variable θ. Equations (6) and (7) become

$$b_1 = \frac{\gamma\sigma_1^2}{1+\left(\gamma+\delta\right)\sigma_1^2} \tag{6'}$$

$$b_2 = \frac{\gamma}{\gamma+\delta}. \tag{7'}$$

Equation (7′) is the standard formula for sharing risky income between two risk-averse parties since γ and δ are the coefficients of absolute risk aversion. In general with finite risk aversion on both sides there should be some cost pass-through. Only when the firm is risk neutral (so it makes sense for it to bear all the risk) or when the consumer is infinitely risk averse (so the price has to be fixed for the consumer to participate) is it optimal to have no cost pass-through. Full pass-through of the exogenous costs is optimal when the consumer is risk neutral.[7]

[7] For a general analysis of risk sharing between consumers and the firm see Cowan (2003).

Equation (6′) gives the pass-through coefficient for endogenous costs. Compared to the analysis in Armstrong et al. (1994) the difference is the presence of the consumer's risk aversion coefficient, δ, in the denominator, which naturally dampens the responsiveness of price to c_1. With independent cost shocks b_1 is zero again if the firm is risk neutral or if the consumer is infinitely risk averse.

Now suppose that the cost terms are perfectly correlated so $r = 1$. From (6) and (7) we see that it is optimal to have zero pass-through of the endogenous costs, but that there should be positive pass-through of the exogenous costs. The reason it is zero pass-through of the exogenous costs is optimal is that if c_1 rises, without c_2 also increasing, the regulator knows that this is because of a fall in e and thus the firm should obtain no reward.

2.3. Regulatory lag

So far the models have made no reference to time. In practice PCR requires price reviews after fixed intervals, with the gap between reviews known as the regulatory lag. A critical factor determining the long-run incentive properties of a price regulation scheme is the length of the regulatory lag. In the UK this is four or five years (though in the water industry it can be ten years if neither side requests a review after five), but in the US the lag is usually two years. Note that these match the lengths of time between General Elections for Parliament in the UK and Congressional elections in the USA. If the regulator resets the price at the review so that expected future profits are zero then clearly the long-run incentives for efficiency are lower the shorter is the regulatory lag. Very short lags would make the system look like rate-of-return regulation. In practice in the UK regulators have not always followed the practice of eliminating excess returns at the review. For example the first reviews of both the telecommunications and water price caps allowed the firms to continue to earn supernormal profits into the second period in order to enhance the incentive effects for the future. Although this practice might at first sight seem desirable it also illustrates the main weakness of PCR (to be addressed in the next section), which is that typically the regulators have a large degree of discretion. An additional point is that a very long regulatory lag might not be credible. Suppose that the lag is twenty years (which is the typical life of a patent). Both sides would probably then expect that the regulator would intervene any-way well before twenty years have passed when external shocks happen (for example a sustained oil price change in the case of an energy supplier).

2.4. Quality

One drawback of PCR is that by providing incentives for cost reduction it might encourage the firm to reduce its quality level. In practice, though, while regulators have to be aware of the potential for quality degradation this does not seem to have been too much of an issue. In the UK the one exception has been the railways. It is not clear in this case that PCR itself is to blame rather than the industry restructuring at both the horizontal and vertical levels. Regulators typically monitor several measures of quality, publish the results (in league tables for regionally organized industries) and threaten to set lower prices at the next review for firms with poor quality performances. The problem with this approach is that it is not clear that regulators can appropriately trade-off price against quality. A preferable scheme is to require firms that do not meet quality targets to give rebates to customers (this applies for example to the train operating companies), though again the setting of the targets themselves is likely to be rather arbitrary.

2.5. The structure of the formula for multi-product firms

One of the characteristics of PCR is that the firm has flexibility over the choice of relative prices when it serves several markets. Optimal prices for a multi-product natural monopoly are known as Ramsey prices and maximize consumer surplus subject to a profit constraint. The price constraints that have applied to British Telecom, the water companies in England and Wales and to AT&T are of the form

$$\sum_{i=1}^{n} p_i^t q_i^{t-1} \leq \left(1+RPI-X\right)\sum_{i=1}^{n} p_i^{t-1} q_i^{t-1} \tag{8}$$

where *RPI* is the percentage growth in the Retail Prices Index, p_i^t is the price of product *i* in time period *t* and *q* is the quantity sold. This is the same as capping a Laspeyres index of prices. Vogelsang (1989) shows that in the limit this type of constraint leads to Ramsey prices, though the rents will only be extracted from the firm if *X* is set appropriately. Intuitively this formula guarantees that the cost to consumers as a whole of purchasing the year $t-1$ quantities at the year t prices is X per cent lower in real terms than the actual cost in year t-1.The firm then seeks out the price that maximize profits subject to the constraint that consumers are at least X per cent better off.

In energy supply and airports a different formula based on a cap on average revenue has been used. For example in gas in the UK the average revenue per therm was controlled and this was supplemented by a subsidiary cap on the standing charge (the quarterly lump-sum charge that each customer paid to finance network costs). In theory at least an average revenue cap can lead to significant distortions compared to Ramsey prices—see Bradley and Price (1988) and Cowan (1997)—though in practice this does not seem to have been a concern.

Network firms in energy such as transmission and distribution companies often face forms of incentive regulation that are not simple price caps. For example the national electricity transmission company for England and Wales, and two Australian grid companies, have caps on total revenue, not on their prices. The idea here is that the costs of a transmission company are largely fixed. When demand shifts a firm with a price cap would face large fluctuations in profits and might then require a higher cost of capital. A revenue cap requires the price to fall when demand shifts out and thus dampens the extent of profit variation (though at the same time it increases the price variability that consumers face).

2.6. Access pricing and global price caps

The access pricing issue arises when a firm that operates the network as a natural monopoly also offers services over the network that may be subject to potential competition. Competitors at the retail stage need access to the natural monopoly services and the policy problem is to find the correct price for these services. I shall not present a full discussion of access pricing here.[8] A regulator setting access prices wants to ensure that efficient rivals are given reasonable access terms, that inefficient entry is discouraged and that the network firm is able to cover all of its costs. Whether these objectives are achievable depends on the instruments available to the regulator (for example

[8] See Armstrong et al. (1996) and Laffont and Tirole (1996) for models of access pricing.

whether the retail price is regulated) and on the information conditions. Laffont and Tirole (1996) recommend what they call a "global price cap". This treats the access pricing issue as a special case of multi-product pricing, where the firm has two types of customer, final consumers and intermediate goods producers, who consist of the rival retailers and its own retail division. The way to implement a global price cap is to use formula (8). In practice, though, regulators have tended to treat retail and access price regulation rather separately.

2.7. Price-cap regulation, entry and competition

Littlechild's original recommendation for BT's price cap in his 1983 report was that it should apply only to the monopoly element of BT's services, which he thought would just be local calls since long-distance calls were potentially competitive. He envisaged that even for local calls the price cap could eventually wither away as competition came in, and thus the point of the price cap was to "hold the fort" until competition arrived. The UK Government did not move as quickly as Littlechild had expected to allow competition in the telecommunications market, and it has taken almost two decades for the original vision to be realized. BT remains a very dominant player in network services so these are likely to remain regulated for a long time. Indeed PCR was expanded rather than contracted early on as international calls were brought into the control. In the energy industries retail price caps are now being eliminated as domestic supply competition has grown, though the incumbent suppliers retain dominant positions. Price regulation in these industries will in the future focus on the network services of transmission and distribution.

For countries starting out on the process of price regulation it is worth noting that it appears to be easier to introduce direct product market competition under PCR than under rate-of-return regulation. One reason for this is that under rate-of-return the regulator has an obligation to ensure that the firm covers its cost of capital, and competition makes this difficult to achieve. The progress of competition under PCR depends largely on the margin that entrants are allowed, which is the gap between the regulated retail price of the incumbent and the price of network services (which might themselves be provided by an integrated incumbent). To some extent there is a trade-off between the short-run interests of customers, who want the dominant firm to set a low retail price, and their long-run interest, which is more likely to be achieved with a high retail price that encourages entry and thus lower prices in the end.

3. Regulatory commitment

Probably the greatest challenge in implementing PCR is to establish long-term credibility. Within the period of regulatory lag the incentives for efficiency depend on the regulator sticking to the agreed price path irrespective of any new information that arrives. But for industries where asset lives are much longer than the regulatory lag period (for example in the water industry assets such as reservoirs might be depreciated over 150 years) the problem is that future prices are not guaranteed. Indeed once an asset is in place the regulator might be tempted to allow the firm to recover only its avoidable costs and disallow any return on the (sunk) capital. Anticipating such expropriation the firm will hold back on its investment.

Salant and Woroch (1992) and Gilbert and Newbery (1994) present game-theoretic models of this hold-up problem for investment. Gilbert and Newbery argue that rate-of-return regulation can be preferable to PCR when it comes to investment promotion. Sometimes a large shock to the whole economy will make it impossible to keep to existing contracts, as was the case in Argentina early in 2002 when the currency board collapsed and there was a large uncontrolled devaluation. At this point guarantees to foreign investors in utilities that specified that it was the Producer Price Index in the USA that would be used as the base index for PCR were torn up.

Institutional design is part of the solution to the problem of long-term investment incentives. Levy and Spiller (1996) discuss the importance of the institutional endowment of a country in determining the degree of commitment. At a minimum the regulatory agency needs to act independently of the government (though there must be some accountability). It is difficult to see PCR working at all when a government ministry is the regulator because of the day-to-day political pressure that would exist. Levy and Spiller point out that it is important that the politicians are not able to change the rules of the game arbitrarily, so the way the legislature works, the degree of independence of the judiciary and the availability of appeal mechanisms are also important. The flexible type of PCR seen in the UK and, to some extent in the USA, will be difficult to implement in countries with less favourable institutional endowments.

In the end, though, it does seem that some kind of guarantee about the rate-of-return on assets is inevitable if the concerns about sub-optimal investment are to be alleviated. In the UK for example the water industry, which has and has had a large investment programme, receives some comfort from the fact that the regulator's primary duty under the legislation establishing regulation is to ensure that the firms earn reasonable returns on their capital. The critical problem is that regulators cannot commit their successors to particular policies. Without a legislative constraint limiting regulatory discretion investment might not be forthcoming.

4. Empirical evidence on performance under price-cap regulation

There is now a small literature examining the empirical evidence on the performance of firms facing PCR. Mathios and Rogers (1989) examined AT&T's long-distance prices and found that most prices were significantly lower in states that allowed pricing flexibility than in states that used rate-of-return regulation. They used a simple dummy variable technique to capture PCR, and of course this evidence does not prove that PCR caused prices to be lower. More recently Resende (2000) applies a combination of data envelopment analysis (DEA) and econometric techniques to US local telephony and finds that incentive regulation (including PCR) is associated with greater productive efficiency than rate-of-return regulation. Ai and Sappington (2001) examine the impact of three types of state incentive regulation (PCR, rate case moratoria and earnings sharing) and of traditional rate-of-return regulation on various performance measures for local telecommunications services in the US. They find that under the three forms of incentive regulation there is greater network modernization. Costs tend to be lower under rate case moratoria, and are reduced under earnings sharing and PCR when there is sufficient local competition. While local call rates for business are lower under PCR the form of regulation appears to make no difference to residential call rates. Alexander et al. (1996) examine the evidence on the cost of capital for regulated industries and find that, as expected, firms facing incentive regulation have higher systematic risk than firms subject to rate-of-return regulation.

5. Conclusion

Price-cap regulation has proved to be popular, both as a method of holding the fort until competition arrives and as a long-term procedure for regulating the prices of network services that will remain monopolized. It appears to be successful in its main aim of establishing incentives within the regulatory period for cost efficiency. The long-term concern, and a challenge for policymakers, is how to promote long-term investment. Would the large infrastructure investments of the 19th and 20th centuries (railways, sewerage and clean water, electricity, natural gas, telecommunications) have been made if the firms had been subjected to price-cap regulation? In practice such investments were funded by the state or by private firms with guarantees about their rates of return. Price-cap regulation may be more appropriate for industries without substantial investment requirements where there is excess capacity than for expanding industries with large investment plans. If this is so then price-cap regulation may simply be a passing phase rather than a universal panacea for infrastructure industries.

References

Ai, C. and Sappington, D. (2001), The impact of state incentive regulation on the U.S. telecommunications industry, Mimeo, University of Florida.

Alexander, I., Mayer, C. and Weeds, H. (1996), Regulatory structure and risk: An international comparison, World Bank Policy Research Working Paper 1698.

Armstrong, M., Cowan, S. and Vickers, J. (1994), Regulatory Reform: Economic Analysis and British Experience, MIT Press, Cambridge, MA.

Armstrong, M., Doyle, C. and Vickers, J. (1996), The access pricing problem: A synthesis, Journal of Industrial Economics 44, 131-50.

Armstrong, M. and Sappington, D. (2002), Recent developments in the theory of regulation, Mimeo, Nuffield College, Oxford.

Bernstein, J. and Sappington, D. (1999), Setting the *X* factor in price-cap regulation plans, Journal of Regulatory Economics 16, 5-25.

Borenstein, S. (2002), The trouble with electricity markets: Understanding California's restructuring disaster, Journal of Economic Perspectives 16, 191-211.

Bradley, I. and Price, C. (1988), The economic regulation of private industries by price constraints, Journal of Industrial Economics 37, 99-106.

Caillaud, B. and Tirole, J. (2001), Essential facility financing and market structure, Mimeo, IDEI, Toulouse.

Cowan, S. (1997), Tight average revenue regulation can be worse than no regulation, Journal of Industrial Economics 45, 75-88.

Cowan, S. (2001), Developments in regulatory principles: The UK experience, in C. Henry, M. Matheu and A. Jeunemaître (eds.), Regulation of Network Utilities: The European Experience, Oxford University Press, Oxford.

Cowan, S. (2003), Optimal risk allocation for regulated monopolies and consumers, Journal of Public Economics, forthcoming.

Engel, E.M., Fischer, R.D and Galetovic, A. (2002), Competition in or for the field: Which is better?, Yale University Cowles Foundation Discussion Paper 1358.

Gilbert, R. and Newbery, D. (1994), The dynamic efficiency of regulatory constitutions, RAND Journal of Economics 25, 538-554.

Joskow, P.L. (2001), California's electricity crisis, Oxford Review of Economic Policy 17, 365-388.

Laffont, J-J. and Tirole, J. (1986), Using cost information to regulate firms, Journal of Political Economy 94, 614-41.

Laffont, J-J. and Tirole, J. (1993), A Theory of Incentives in Procurement and Regulation, MIT Press, Cambridge, MA.

Laffont, J-J. and Tirole, J. (1996), Creating competition through interconnection: Theory and practice, Journal of Regulatory Economics 10, 227-256.

Levy, B. and Spiller, P.T. (1996), A framework for resolving the regulatory problem, in B. Levy and P.T. Spiller, (eds.), Regulation, Institutions and Commitment, Cambridge University Press, Cambridge, UK.

Littlechild, S.C. (1983), Regulation of British Telecommunications' Profitability, Department of Industry, HMSO, London.

Loeb, M. and Magat, W. (1979), A decentralized model for utility regulation, Journal of Law and Economics 22, 399-404.

Mathios, A.D. and Rogers, R.P. (1989), The impact of alternative forms of state regulation of AT&T on direct-dial, long-distance telephone rates", RAND Journal of Economics 20, 437-453.

Monopolies and Mergers Commission (1982). Contraceptive Sheaths, Cmnd 8689, HMSO, London.

Resende, M. (2000), Regulatory regimes and efficiency in US local telephony, Oxford Economic Papers 52, 447-470.

Salant, D. and Woroch, G. (1992), Trigger price regulation, RAND Journal of Economics 23, 29-51.

Shleifer, A. (1985), A theory of yardstick competition, RAND Journal of Economics 16, 319-317.

Vogelsang, I. (1989), Price cap regulation of telecommunications services: A long-run approach, in M.A. Crew (ed.), Deregulation and Diversification of Utilities, Kluwer, Boston.