THE VAN HORNE INSTITUTE

Performance Based Regulation of Utilities: Theoretical Developments in the Last Two Decades

C. R. (Sid) Carlson

Performance Based Regulation of Utilities: Theoretical Developments in the Last Two Decades

March 2010

C. R. (Sid) Carlson

C. R. Carlson is an adjunct professor in the School of Business at the University of Alberta. She holds a M.Sc. from the California Institute of Technology and a Ph.D. from the University of Iowa.

This paper was made possible by funding from the Van Horne Institute, Borden Ladner Gervais, ATCO, EPCOR, FortisAlberta, and AltaLink. The author is grateful for helpful comments and suggestions from Joseph Doucet and four anonymous reviewers from the Van Horne Institute.

Forward

The Van Horne Institute's Centre for Regulatory Affairs is pleased to release this background paper that reviews and discusses some of the key developments from the theoretical literature on performance based regulation (PBR). While both theory and practice continue to evolve, the objective of this paper is to present some of the issues in the literature that may be relevant to the design and implementation of incentive regulation in the coming years. This is of course particularly relevant today in light of the Alberta Utilities Commission's (AUC) rate regulation initiative. In light of the AUC's 2009 approval of Enmax' PBR plan and as the AUC moves forward with its initiative "to reform utility rate regulation in Alberta for electricity and natural gas distribution services" this survey is sure to be both timely and relevant to decision makers.

Joseph Doucet
Enbridge Professor of Energy Policy, The University of Alberta
Academic Director, Centre for Regulatory Affairs, The Van Horne Institute

Executive Summary

This paper reviews key developments in the theoretical literature on performance based regulation (PBR) over the last two decades. Theory has continued to develop since the insights of the 1980s: some approaches have been discarded, consensus has been gained on some recommendations, and many questions are still being explored. The intent is to lay out some of the issues in the literature that may be relevant to the design and implementation of incentive regulation today.

PBR has replaced traditional rate of return (ROR) regulation for utility rate regulation in many jurisdictions across the world. In Canada, the National Energy Board (NEB) and the CRTC have used some form of PBR since the mid-1990s. Interest in PBR may be increasing at the provincial level. Alberta, which has until now relied on ROR regulation, approved its first formal PBR plan last year for ENMAX Power. Ontario's recent announcement that it is considering privatizing its Crown Corporations will put it in much the same position as Britain during the 1980s, when it sold many of its public enterprises and needed then to devise new regulatory regimes for their governance.

The essence of PBR is to create incentives for firms to achieve social objectives by aligning their profits with those social objectives. Most PBR policies are based on two properties:

1. They give the firm some flexibility in how it meets performance goals.
2. They use a system of rewards or penalties for meeting those goals.

Flexibility is important because it allows the firm to use its superior knowledge to find better ways to achieve objectives. Flexibility gives firms the means to achieve goals; how vigorously the firm pursues those goals is determined by the strength of the incentives embedded in the policy. Specific policies contain many elements, often interrelated, that affect the overall incentives created for the firm. While PBR can be theoretically superior to ROR regulation, the details of implementation are key. It can be a delicate balance: the regulator must allow enough discretion for independently driven change and innovation to take place, but not so much that it leads to unintended negative consequences.

The conclusions of the early literature on the properties of good regulatory regimes differed radically from the philosophy of traditional rate of return regulation. Three key insights were:

1. The regulator can and must delegate some decision-making authority to the utility, even when its actions and costs are unobservable. The focus should be on creating the right incentives and giving the firm enough flexibility to respond to those incentives, not on prescribing specific behaviours.

2. The utility must be allowed to earn some amount more than normal profit. This is what provides the incentive for the utility to make socially efficient choices.

3. The regulator must be able to commit to allowing more than normal profit, even if it gains more information about costs and other parameters over time. This means that high earnings cannot simply be clawed back in subsequent regulatory reviews.

Over the last twenty years, analysts have tested these recommendations in more complex scenarios. They moved from static to dynamic settings, which raised such issues as whether firms would behave strategically to influence future regulatory decisions, the ability of a regulator to commit to policies over time and how that might affect the firm's willingness to invest in capital infrastructure, and the incentives for firms to adopt new and innovative technologies at an appropriate rate. More complex cost structures were examined, leading to questions about how best to allocate shared costs over multiple services. The possibility that the regulated firm might also compete in unregulated markets or control infrastructure necessary for the operation of rival firms was raised, which led to questions about how to design pricing and cost allocation policies to discourage anti-competitive behaviour.

The basic findings of the early literature have been found to be fairly robust. However, the exploration of more complex situations, along with a growing body of empirical evidence, showed that there is no "one size fits all" recommendation for designing PBR. Optimal policy depends on institutional factors, consumer preferences, availability of information, market dynamics and environment, and the nature of the industry.

Some specific recommendations from the earlier literature were reinforced. First, PBR creates stronger incentives when:

1. Performance goals are defined by factors exogenous to the firm, meaning that they cannot be influenced by the regulated firm's choices and behaviour.
2. The length of time between regulatory reviews in increased, since this increases the potential for both profits and losses.
3. The firm has reason to believe that the regulator will not renege on announced policies.

Second, PBR should not create undue risks for either regulated firms or consumers. Regulation should include adjustments to account for exogenous change, since firms should not be unduly exposed to factors outside of their control. Provisions to share profits and losses with customers reduce the firm's incentives to cut costs, but may provide assurance necessary to move from ROR regulation to PBR.

Commitment to allow an adequate return on investment is particularly important in sectors where capital investments are large and long-lived. Other findings on efficient investment in capital and new technology were:

1. Profit sharing when a multiproduct firm competes in both a regulated monopoly market and competitive markets leads to inefficient pricing and technology choices. It also creates incentives to misreport costs and under diversify into competitive markets.
2. PBR is superior to ROR regulation in driving efficient investment in capital and new technology if the regulator can commit to adequately rewarding investment. Otherwise, ROR regulation based on the “used and useful” principle is more likely to result in efficient investment.
3. Incentives to adopt innovative technology are increased if the regulator allows the firm to keep a greater share of the innovation benefits, even when regulatory commitment is weakened.
4. Yardstick competition, where some of the firm’s performance goals are based on the performance of other firms, can reduce the incentive to invest in more efficient technology. This is particularly true if the investment benefits other firms, as would be the case in network investments.

Regulation must also be designed with the properties of the overall regulatory regime in mind. For example, the regulation of service quality becomes more critical when PBR is used, since one way to cut costs is to reduce service levels. Few clear guidelines are apparent in the literature on combining price and quality regulation, except that it is a difficult proposition. On regulating retail service quality, it is noted that:

1. Minimum quality standards are easy to apply, but may or may not benefit consumers. The benefits depend on the market environment, consumer preferences, and whether regulators increase standards in response to quality improvements.
2. A system including both bonuses and penalties can be useful if quality is observable and verifiable.

Wholesale service quality raises different issues, since now one must consider how service quality levels can be used strategically to disadvantage competitors in retail markets. Parity standards, wholesale price regulation, and “carrots and sticks” bargaining have been explored, but none of these stand out as superior solutions.

One recurring theme in the literature was how practice and theory split from the late 1980s on, with new developments in the theoretical literature having little or no effect on the design and implementation of new regulatory regimes. Bridging the gap between theory and practice requires communication that is often lacking. The U.K. seems to have put more emphasis on this type of communication than other jurisdictions. For example, there were conferences in 2004 and 2005 that brought theorists and practitioners together to discuss efficiency estimates and capital investment. These led to special issues of *Utilities Policy* devoted to those questions, which disseminated the findings to the larger academic community. Similar activities could prove useful here, particularly for provinces that are moving to new regulatory regimes.

Contents

1. Introduction

Performance based regulation, in its broadest definition, is regulation based on *what* is achieved rather than *how*. As May (2003) observed, it is a change of mind set from prescribing specific behaviours in the hope of producing a desirable outcome to one where the goals are set, and entities are given discretion as to how they are achieved. It can be a delicate balance: the regulator must allow enough discretion for independently driven change and innovation to take place, but not so much that it leads to unintended negative consequences.

Performance based regulation (PBR) has replaced traditional rate of return (ROR) regulation for utility rate regulation in many jurisdictions across the world. As of 2002, PBR had become the norm for regulating telecommunication and energy utilities outside of North America. It has also been adopted widely in the U.S. for telecommunications, energy, and railroads (Lowry and Kaufmann, 2002). In Canada, the National Energy Board (NEB) and the CRTC have used some form of PBR since the mid-1990s.

Interest in PBR may be increasing at the provincial level. Alberta, which has until now relied on ROR regulation, approved its first formal PBR plan last year for ENMAX Power. Ontario's recent announcement that it is considering privatizing its Crown Corporations will put it in much the same position as Britain during the 1980s, when it sold many of its public enterprises and needed then to devise new regulatory regimes for their governance.

This paper reviews key developments in the theoretical literature on PBR – also called incentive regulation – over the last two decades. Theory has continued to develop since the insights of the 1980s: some approaches have been discarded, consensus has been gained on some recommendations, and many questions are still being explored. More than one regulatory economist has noted that there has been a divergence between theory and practice over the last twenty years. The intent here is to lay out some of the issues in the literature that may be relevant to the design and implementation of incentive regulation today.

1.1. Evolution of new approach

The theoretical literature on incentive regulation over the last thirty years followed two distinct approaches: Bayesian and non-Bayesian.[1] Bayesian analyses were based on highly stylized models of regulation, drawing from existing literature on games of incomplete information and principle-agent theory. Principle-agent theory was designed to solve a situation of asymmetric information between an entity (the principal) who needs to use another entity (the agent) to achieve some

[1] Material in this section is based on Acton and Vogelsang (1989), Schmalansee (1989), Laffont (1994), and Crew and Kleindorfer (2002).

desired outcome. The problem arises when the incentives of the principal and the agent differ, and the principal cannot directly observe the actions of the agent.

The solution, first applied to private contracting and governance issues, is to set up a mechanism that changes the agent's incentives to reflect those of the principal. This was the essence of the new approach to regulation: creating incentives that led firms, through their efforts to maximize profit, to willingly pursue socially desirable behaviours.

The merits of different regulatory regimes are judged on several criteria. The main questions in the theoretical literature have been:

1. Do they result in prices and output levels that maximize some measure of social benefit?
2. Do they induce firms to reduce costs, given the technology they have in place?
3. Are firms given the incentive to make efficient capital investment decisions in expanding infrastructure or adopting new technologies?
4. Will firms find it optimal to offer service quality at the right levels and prices?
5. If the firm is competing in both regulated and unregulated markets, is it limited in its ability to engage in anti-competitive strategic behaviour?
6. Does the policy result in arbitrary wealth redistribution or pose an undue amount of risk on either utilities or consumers?

Literature from the 1970's and early 1980's was primarily concerned with the first two questions. Maximizing social benefit involved asking if policies resulted in efficient pricing. Efficient resource use was evaluated based on the incentives created to reduce production costs. It was first established that, in settings with restrictive assumptions, it was possible to design policies that induced both efficient pricing and production even when the regulator had incomplete information about the firm's costs and technology possibilities.[2]

Following these early insights, analysts began exploring more complex scenarios. They moved from static to dynamic settings, which raised such issues as whether firms would behave strategically to influence future regulatory decisions, the ability (or lack thereof) of a regulator to commit to policies over time and how that might affect the firm's willingness to invest in capital infrastructure, and the incentives for firms to adopt new and innovative technologies at an appropriate rate. More complex cost structures were examined, leading to questions about how best to allocate shared costs over multiple services. The possibility that the regulated firm might also compete in unregulated markets or control infrastructure necessary for the operation of rival

[2] Baron and Myerson (1982) showed this for a single product regulated monopoly in a static setting, where demand is known to both regulator and firm, but the firm has better information about its own costs. Sappington (1983) extended this to a multiproduct monopoly in a static setting, where there is common knowledge about demand, and with the assumption that all shared costs were fixed costs.

firms was raised, which led to questions about how to design pricing and cost allocation policies to discourage anti-competitive behaviour.

More recently, service quality has been raised as an important issue. Pressure to reduce costs could create incentives to lower service quality; pricing flexibility might lead the firm to choose inefficient prices for different service levels, or to provide inefficient levels of quality. For vertically integrated utilities that have a monopoly on an upstream service, there could be an incentive to provide lower quality service to rivals competing with a downstream unregulated affiliate of the utility.

The conclusions of the Bayesian literature on the properties of good regulatory regimes differed radically from the philosophy of traditional rate of return regulation. Vogelsang (2002) notes three fundamental insights from the early Bayesian literature:

1. The regulator can and must delegate some decision-making authority to the utility, even when its actions and costs are unobservable. The focus should be on creating the right incentives and giving the firm enough flexibility to respond to those incentives, not on prescribing specific behaviours.
2. The utility must be allowed to earn some amount more than normal profit.[3] This is what provides the incentive for the utility to make socially efficient choices.
3. The regulator must be able to commit to allowing more than normal profit, even if it gains more information about costs and other parameters over time. This means that high earnings cannot simply be clawed back in subsequent regulatory reviews.

However, the Bayesian literature did not provide guidance on how exactly to incorporate these principles into actual regulatory policies. In addition to being highly abstracted from reality, the Bayesian mechanisms were criticized as requiring as much knowledge (although of a different kind) on the regulator's part as traditional regulation (Schmalansee, 1989; Liston, 1989; Crew and Kleindorfer, 2002). Much of the literature over the last twenty years has thus focused on non-Bayesian mechanisms, where analysts have sought to design policies that rely on observable data, yet retain the desirable incentives of the Bayesian approach. These are also, for the most part, the kinds of policies that have been adopted in practice around the world.

1.2. Overview of paper

The literature on performance based regulation is vast. The focus of this paper has therefore been limited to utility regulation, and the literature search to peer-

[3] "Normal profit" includes coverage of operating costs and just enough return on capital to justify its use in that market. Normal profit is also called zero economic profit.

reviewed journal articles that provided key insights or high quality summaries of specific branches of the literature.[4]

The paper proceeds as follows. Section 2 discusses the goals and components of PBR and compares price caps and revenue caps, the two most commonly used forms of PBR. The construction of price caps and findings on the relative merits of different index choices is discussed in Section 3. In Section 4, theoretical results on the effects of price cap regulation on capital investment are presented. The literature on service quality is discussed in Section 5, and Section 6 concludes with perspectives from noted regulatory economists on the evolution of theory in this area.

2. Performance Based Rate Regulation: Overview

Most PBR policies are based on two properties (Vogelsang, 2006):

1. They give the firm some flexibility in how it meets performance goals.
2. They use a system of rewards or penalties for meeting those goals.

Flexibility is important because it allows the firm to use its superior knowledge to find better ways to achieve objectives. As an example, consider the difference between controlling costs with either a technology standard or a performance goal. If firms use different technologies, then the most efficient technology may vary by firm. But it is usually prohibitively expensive to designate multiple standards, so only one technology will be chosen by the regulator. Efficiency would be increased if each firm was able to choose the technology best suited to its operations.

Flexibility gives firms the means to achieve goals; the strength of the incentives embedded in the policy determines how vigorously the firm will pursue those goals.
Specific policies contain many elements, often interrelated, that affect the overall incentives created for the firm. The longer the policy is in place, the greater the rewards of meeting or exceeding goals; conversely, the greater the losses of not meeting them. So increasing regulatory lag (the period between reviews) creates stronger incentives.

This incentive is weakened if earnings or losses are shared with customers. They are also weakened if the firm believes that higher than normal earnings will be clawed back or that losses will be covered in the next review. Resets that trigger automatic reviews in case of unexpectedly high profits or losses between regulatory reviews have the same effect.

[4] The notable exception is a reliance on a chapter on the theory of regulation written by Armstrong and Sappington (2003) for the *Handbook of Industrial Organization*.

However, sharing and reset provisions reduce risk to all parties. Reducing risk increases the acceptable regulatory lag. It also reduces the likelihood that the regulator will renege on its announced policy, either mid-stream or at the next review (Crew and Kleindorfer, 1996). So the effect of sharing and resets on incentive strength is mixed. As a practical reality, the reassurance they provide to all parties may be necessary to implement incentive regulation in the first place.

The general goals of utility regulation from the economic perspective were outlined in the introduction. The first objective – to maximize some measure of social benefit – bears a bit more explanation since it drives the theoretical evaluation of specific pricing policies discussed in later sections.

2.1. Maximizing social benefit

Social benefit, or "social welfare", is a measure of the net gains from trade in a market. Consumers' gains are called consumers surplus and they arise from the difference between the value consumers place on a good or service and what they have to pay to obtain it. A firm's gain, called producers surplus, is simply its profit in this context. In the absence of regulation, gains from trade are the sum of consumers and producers surplus. Regulation entails costs, so social welfare under regulation is the sum of consumers and producers surplus less the cost of regulation.

Most of the literature does not model regulatory costs explicitly. Instead, they assume that the regulator seeks to maximize social welfare, or sometimes a weighted average of consumers and producers surplus, with more weight put on consumers surplus.[5]

With a single product firm or a multiproduct firm where operating costs can be attributed unambiguously to the different products, gains from trade are maximized when prices are equal to the marginal cost of production. Such prices are defined as "efficient." Unfortunately, marginal cost pricing as a rule presents difficulties both in theory and in practice. To illustrate, consider the simplest case of a single product natural monopoly. A single product natural monopoly, by definition, has decreasing average costs at the output level serving the entire market. This means that marginal costs are below average costs at that output level, so marginal cost pricing results in losses.

There have been two general approaches to this problem. The first, often proposed as part of Bayesian mechanisms, is to set prices equal to marginal cost and then cover the losses with a transfer. Transfers are problematic on legal grounds, as most regulatory agencies do not have the authority to grant lump-

[5] The cost of regulation is considered, of course. For example, it would be used to rank policies with otherwise similar properties, or to reject a policy if the regulatory costs were prohibitive. It is also a primary consideration in determining whether a regulated monopoly is better than a deregulated, but imperfectly competitive, market (Armstrong and Sappington, 2006).

sum transfers to firms.[6] The alternative is to use a second-best rule called Ramsey pricing. The Ramsey pricing rule states that the percent markup of price over marginal cost should be inversely proportional to the elasticity of demand. This results in higher mark ups on goods where demand is relatively insensitive to price and lower markups on goods where demand is more sensitive to price. Intuitively, this gives the least distortion from the output levels that would have been demanded if prices had been set equal to marginal costs.

When there are shared operating costs across services, which is often the case, marginal cost is not defined, and third-best pricing solutions are then explored. However, the general idea is that analysts are looking for prices that converge to or approximate Ramsey prices when transfers are not an option.[7]

2.2. Caps: basic forms

Both revenue and price caps have been used for incentive regulation, although price caps are much more common in practice. Revenue caps are administratively easy to implement, but they have less desirable properties than price caps and have been heavily criticized in the literature (e.g., Crew and Kleindorfer, 1996; Cowan, 1997; Vogelsang, 2002; Armstrong and Sappington, 2003). Revenue caps are explained and compared to rate caps below, followed by an overview of price cap regulation.

2.2.1. Revenue caps

Most revenue caps restrict the rate of growth in average revenues, and they may include factors accounting for inflation, growth in productivity and output, and various exogenous factors. Caps may apply to revenues from all regulated products (comprehensive caps) or they may only apply to revenues used to cover specific costs (non-comprehensive caps).

In Canada, the National Energy Board (NEB) has used comprehensive revenue caps for oil pipelines operated by Enbridge and TransMountain Pipelines, and non-comprehensive caps for Westcoast Energy gas distribution.[8] Non-comprehensive caps have also been applied to BC Gas in 1994, NOVA Gas Transmission in 1998, and Consumers Gas of Ontario in 1998 (Lowry and Kaufmann, 2002). The U.S., Britain, and Australia have also used average revenue caps. Examples of revenue caps are shown in Table 1.[9]

Table 1. Revenue Cap Examples, by Jurisdiction[10]

Country	Utility	Type

[6] Meran and von Hirschhausen (2009) note that it was subsequently pointed out that a fixed access fee is mathematically equivalent to a transfer, and regulators do have the authority to set access fees.
[7] One reviewer pointed out that regulators appear to be more focussed on cutting costs than on achieving abstract goals like marginal cost pricing or efficient rate rebalancing, which could be an example of how the focus of theory can diverge from practice.
[8] The NEB appears to have since moved away from revenue caps (NEB, 2009).
[9] The data in Table 1 are not current.
[10] Sources: Cowan (1997), Lowry and Kaufmann (2002).

Canada	Enbridge Pipelines TransMountain Pipelines	Comprehensive
	BC Gas Westcoast Energy NOVA Gas Transmission Consumers Gas	Non-comprehensive
Britain	National Grid Transmission	Comprehensive
	Regional electric distribution British Gas pipelines Airports	Unknown
U.S.	Southern California Gas	Revenue per customer
	PacifiCorp electricity distribution	Comprehensive
	San Diego Gas & Electric	Non-comprehensive
Australia	Energy Australia transmission Powerlink Queensland transmission TransGrid transmission	Comprehensive

Revenue caps do create incentives to reduce costs and provide pricing flexibility (Crew and Kleindorfer, 1996; Lowry and Kaufmann, 2002). In its most basic form, an average revenue cap sets a maximum average price:

$$\frac{\sum_i p_i q_i}{Q} \le p^0$$

where p_i is the price of the service i, q_i is the quantity demanded at price p_i, Q is an aggregate index of outputs, and p^0 is the maximum average price (Vogelsang, 2002). If unlagged quantities are used, this becomes

$$\frac{\sum_i p_i^t q_i^t}{\sum_i q_i^t} \le p^0 \qquad (1)$$

where the t superscript denotes the time period. One immediate problem to note is that this cap is not known until actual output levels are determined at the end of the period. Hence it requires forecasts, with true-ups at the end of each regulatory period (Vogelsang, 2002).

The second more serious problem is that this may give the firm an incentive to set inefficient prices. Note that the firm can influence the value of the weighted average price by its pricing choices, since the quantity demanded depends on price. To illustrate the problem, suppose the firm was only selling two services, and that the demand for the first service is elastic and for the second is inelastic. Ramsey pricing would call for a lower markup of price over marginal cost for the first service, and a higher markup for the second service. However, the firm has the opposite incentive. Raising the price of the first service causes the quantity demanded to drop enough that the sum p_1q_1 decreases, which loosens the price cap. Raising the price of the second causes revenues from that service to increase, tightening the cap (Vogelsang, 2002).

One suggestion to address this problem has been to use a lagged average revenue cap, where prices are weighted by the previous period output levels:

$$\frac{\sum_i p_i^t q_i^{t-1}}{\sum_i q_i^{t-1}} \leq p^0$$

Cowan (1997) compares the dynamic welfare properties of the average revenue cap shown in (1), the lagged average revenue cap shown above, and the chained Laspeyres price cap given by:

$$\frac{\sum_i p_i^t q_i^{t-1}}{\sum_i q_i^{t-1}} \leq \frac{\sum_i p_i^{t-1} q_i^{t-1}}{\sum_i q_i^{t-1}}$$

The price cap imposed in this form requires that the average revenue from current prices, weighted by the previous period's output levels, be no greater than the actual average revenue of the previous period. Its interpretation is that under the new period t prices, consumers could still purchase the same output levels as in the previous period. Vogelsang (1989) proved that this price cap yields prices that converge to efficient levels, with rising consumer surplus over time. Cowan (1997) shows that the price cap results in higher consumer surplus and lower utility profits than the revenue cap unless the marginal costs and elasticities of demand in all services are identical. When elasticities vary, the firm with a revenue cap has an incentive to vary prices too much, and may actually set prices in markets with elastic demand higher than the unregulated monopoly price.

Comparing the price cap with the lagged average revenue cap, Cowan (1997) finds that consumer surplus is higher under the price cap in general. Consumer surplus under the lagged average revenue cap rises initially, but falls thereafter to less than its starting value. Utility profits are higher than with the price cap, and the prices set by the utility are inefficient unless the firm is myopic.

Lowry and Kaufmann (2002) point out two further difficulties with comprehensive average revenue caps. First, they may create stronger incentives to reduce quality than price caps. A utility with a revenue cap who loses customers because of poor service can make up lost revenue by raising price in markets with inelastic demand. This is not an option with a price cap. Second, regulatory costs are probably higher. An average revenue cap will still require standard regulatory hearings to determine how to allocate costs across different customer classes.

Non-comprehensive revenue caps may be a useful approach if the regulator wants to influence where cost-cutting is directed, or if too much controversy exists in setting a comprehensive cap. They may also mitigate concerns about transactions with unregulated affiliates, since the regulator can put a cap on revenues used to purchase goods or services from affiliates. However, Lowry and Kaufmann (2002) also note that non-comprehensive caps distort the utility's incentives and may limit their ability to cut costs overall. On the former, the utility may increase costs in other areas in order to lower costs in the regulated areas. On the latter, non-comprehensive caps may fail to harness the firm's superior knowledge of how best to reduce costs.

Crew and Kleindorfer (1996) are scathingly critical of revenue caps. They characterize them as a hangover from rate of return regulation that "…do not promote efficiency, and should be abolished in electric utility regulation" (Crew and Kleindorfer, p. 215).

2.2.2. Price caps

In competitive industries, market forces limit the ability of firms to earn more than normal profit. Bernstein and Sappington (1999) note that one goal of regulation is to do the same thing in the absence of competition and show how this assumption leads to the basic structure of a price cap. A firm earns only normal profit over time if the following holds:

Rate of change in output prices =
Rate of increase in input prices – Rate of increase in productivity.

Thus, if prices are set in the first regulatory period so that the firm is earning only normal profit, and then allowed to rise according to the difference above, then the firm will continue to earn only normal profit.

As it stands, though, it gives no incentive to the regulated firm to attempt to reduce the prices it pays for inputs or to invest in productivity improvements. These incentives can be created by substituting an expectation of changes in input prices and feasible productivity gains for the actual changes, leading to a price cap of the form:

Rate of change in output prices =
Rate of expected increase in input prices – Rate of expected increase in productivity.

A firm that can do better than the forecast is rewarded with higher profit, and one that does worse will suffer losses. However, there are additional factors that must be considered in a changing environment. The strength of the incentive to reduce costs increases the longer the regulatory period for which the price cap holds. But the risk of this leading to unacceptable profits or losses increases with a long regulatory lag if other factors are not included. Two approaches have been suggested to address this concern. The first is to include a pre-determined formula for sharing profits or losses with consumers, discussed earlier. The second is to build in adjustments to the allowed increase in regulated prices over time that use more current data to update initial forecasts and accounts for other exogenous changes in the industry:

Rate of change in output prices =
Rate of increase in input prices – Rate of increase in productivity
± Exogenous changes.

Price caps (also called rate caps or price or rate indexing) are the most commonly used form of incentive regulation (Vogelsang, 2002). They have been applied in many jurisdictions in telecommunications, power and gas distribution, oil and gas pipelines, water utilities, transportation, and other regulated utilities (Lowry and Kaufmann, 2002).

In Canada, the NEB began using negotiated settlements in 1995 to set oil and gas pipeline tolls over multiple years (Doucet and Littlechild, 2006). Settlements include incentives for cost efficiency and service quality improvement, and may allow for cost savings to be shared with customers (NEB, 2009). Price cap regulation was applied by the CRTC in 1997 in the telecommunications industry, and Ontario has used price caps for both gas and power distribution services (Lowry and Kaufmann, 2002). The Alberta Utilities Commission has relied entirely on ROR regulation until recently,[11] when it approved a price indexing plan for ENMAX Power's distribution services and a formula-based approach for its revenue requirements for transmission services (AUC, 2009).

Table 2. Examples of price caps[12]

Country	Sectors
Canada	Telecommunications, power and gas distribution, oil and gas pipelines
U.S.	Telecommunications, power and gas distribution, oil and gas pipelines, railroads
Britain	Telecommunications, power and gas distribution, water
Australia	Power distribution
Chile	Telecommunications

Under price cap regulation, the regulator sets a formula that determines how fast rates can rise for utility services. When a cap is set over a bundle of services (a "tariff basket"), it determines how fast a weighted average of those service prices can rise. The cap is automatically adjusted between regulatory hearings

[11] The AUC's predecesor, the Energy and Utilities Board, occasionally substituted negotiated settlements for traditional cost of service hearings, but the primary objective was to clear regulatory backlog.
[12] Source: Lowry and Kaufmann (2002).

according to the formula. The service bundles and formulae are reviewed periodically and may be reset (Acton and Vogelsang, 1989; Lowry and Kaufmann, 2002).

There are key differences between price cap regulation in the U.S. and the U.K., due in part to how the shift to the new regulatory regime came about. The U.S. started with private firms that had been operating under rate of return regulation for many years. In the U.K., utilities that had been owned by the Crown were privatized and then controlled via newly created regulation. Crew and Kleindorfer (1996) point out that it is often much easier to institute radical (or any) change if a regulatory regime is being designed from scratch. Changing an existing regime requires dealing with entrenched interests and a reluctance to move from the known to the unknown.

The U.K. approach is based more on Littlechild's original 1983 proposal for British Telecom when it was privatized: that prices should be allowed to rise by the general rate of inflation less an X factor that would force a decline in a weighted average of service prices (RPI – X, where RPI is the Retail Price Index, Britain's measure of general price inflation). The X factor is negotiated between the government and the utility. Crew and Kleindorfer (1996) call this a "purer" form than has been adopted in the U.S., which usually includes reset provisions, safety nets, and elaborate adjustment factors.

The advantages of price caps over ROR regulation, according to Beesley and Littlechild (1989), are:

1. Price caps provide greater incentives for efficiency, both in production and capital investment.
2. Price caps on tariff baskets allow the firm to move to more efficient prices, since the cap allows rate rebalancing within the tariff basket.

In addition, they state that RPI - X is simple to implement and lowers regulatory transaction costs, and that the focus on real price reductions makes it more acceptable to consumers.

In the U.S., price caps take the general form of:

$$\Delta PI \leq \Delta P - X \pm Y$$

where
$\Delta PI =$ the rate of change in the price index of regulated prices
$\Delta P =$ a measure of price inflation
$X =$ total factor productivity, or an index of expected efficiency gains
$Y =$ a factor capturing other variables.

The X factor in U.K. style price caps includes expectations of efficiency gains and the ability of firms to deal with industry-specific factors and exogenous changes.[13] In that sense, RPI - X has been characterized as having a Bayesian component (Vogelsang, 2002). The U.S. style uses indices to calculate the X and Y factors; determination of the X factor is discussed in detail in Section 3. The Y factor is used to incorporate exogenous changes that affect earnings, such as new tax regimes, and other variables unique to the industry that are not included in X. Lowry and Kaufmann (2002) note that a Y factor is useful in that it decreases risk that the utility has no control over without affecting efficiency incentives. However, the calculation of Y can be controversial. Y factors would be revisited at each regulatory review, so they could increase regulatory costs substantially.

Acton and Vogelsang (1989) point out that this more formulaic approach likely reflected political and institutional realities in the U.S., which started from cost of service regulation. It may not be surprising, then, that the price cap adopted for ENMAX follows the U.S. style, since cost of service regulation has been the norm in Alberta.

3. Constructing price caps

Price caps are made up of indices, and which specific indices are used can have an important effect on the properties of the price cap: the incentives it creates, how gains are shared with consumers, the risks created for the regulated firm, how long the price cap period can be, and cost of implementation. Advantages and disadvantages of commonly used indices are discussed below.

3.1. Choosing an output price index

The allowed rate of change in output prices, called the output price index, is a ratio of weighted averages of current and past prices for all services included in the price cap, where individual prices are weighted by quantities produced. The questions are: first, which past prices to use, and second, what quantity levels should be used to create the weighted average? There is no clear right answer from the literature, since choosing ideal weights requires that the regulator know the demand functions and be able to correctly forecast optimal quantities (Vogelsang, 2002). Four commonly used indices are:

Paasche index: $$\frac{\sum_i p_i^t \cdot q_i^t}{\sum_i p_i^0 \cdot q_i^t}$$

[13] One reviewer points out that the UK often accounts for industry-specific factors and exogenous changes in a separate Y factor (also called a Z factor.)

Laspeyres index: $\dfrac{\sum_i p_i^t \cdot q_i^0}{\sum_i p_i^0 \cdot q_i^0}$

Chained Laspeyres: $\dfrac{\sum_i p_i^t \cdot q_i^{t-1}}{\sum_i p_i^{t-1} \cdot q_i^{t-1}}$

Fisher Ideal index: $(Paascheindex \cdot Laspeyresindex)^{1/2}$

The chained Laspeyres index was recommended by Littlechild in his 1983 proposal (Vogelsang, 2002). A practical advantage of this index is that the past period prices and quantities are easy to identify. As mentioned earlier, Vogelsang (1988) showed that this converges to efficient pricing in a stationary environment. However, Neu (1993) shows that if demand is changing over time, and the demand for some services grows faster than others, then this form of pricing flexibility results in inefficient prices. Specifically, prices will be too low in slow growing markets, and too high in fast growing markets. Concerns have also been raised that using one period lagged variables may create incentives for the utility to manipulate the price index for the following year, since it can affect the quantities demanded in that year with its pricing decisions (Diewert and Fox, 2000).

Diewert and Fox (2000) compare the Paasche, Laspeyres, and Fisher Ideal indices. They prefer a Laspeyres index based on a single base year to the chained Laspeyres, with the base year preferably one before the issue of price cap regulation arose. They favour the Fisher Ideal index on the grounds that it is superior to the others in meeting tests of reasonableness as laid out in Diewert (1992). In addition, they see it as balancing the biases of the Paasche and Laspeyres indices, since the Paasche index understates inflation, while the Laspeyres overstates it. Vogelsang (2002) also favours a compromise between the Paasche and the Laspeyres indices, although he suggests using a simple average of the two.

One advantage of the chained Laspeyres index is that it may be easier to incorporate new services into the price cap – an important issue when technology and products are changing rapidly as in the telecommunications sector. Diewert and Fox (2000) note this is a problem with choosing a single base year, since it is hard to estimate what the prices "should have been" if those new services had been available in the past.

3.2. Choosing an inflation rate

There are three general types of indices used for the inflation rate:

1. Macroeconomic inflation rates

2. Industry-specific price inflation rates
3. Peer price measures

Macroeconomic measures, such as the retail or producer price indices, capture general price changes in the economy. They are used in telecommunication regulation world-wide and in energy sectors in Europe and the U.K. (Lowry and Kaufmann, 2002). They are simple to obtain and their calculation is relatively uncontroversial, but they may not capture price changes that are important cost drivers for the regulated firm.

Industry-specific inflation rates have been used in the U.S. for railroad and power and gas distribution caps, and for power distribution in Ontario (Lowry and Kaufmann, 2002). These are calculated as a weighted average of input price indices for that sector. Since they track prices in that sector more closely than macroeconomic inflation rates, they can reduce the risk that the price cap will be unduly harsh if input costs rise faster than general inflation. This allows for longer regulatory periods, which may increase efficiency incentives. However, Lowry and Kaufmann (2002) point out that they have some drawbacks. They must be calculated on a case-by-case basis, so they are not as simple as general inflation rates, and could be a source of controversy that increases the cost and complexity of regulatory hearings. They could also increase regulatory risk if not formulated properly.

Peer price measures create an index based on the prices charged by other similar utilities. The assumption is that pricing trends for the other utilities will reflect both changing input costs and general efficiency gains in the industry. This approach avoids the problem of measuring input costs and efficiency gains directly, and so may be less controversial than industry-specific inflation rates. On the other hand, the question of which utilities should be considered similar may increase controversy. They may also be difficult to apply if utilities have not all unbundled their services and prices in the same way (Lowry and Kaufmann, 2002).

3.3. Determining X

In general, the X factor in price caps has represented expected productivity growth, where productivity is measured by the quantity of input used per unit of output (Lowry and Kaufmann). When there are multiple inputs and outputs, productivity is an aggregate index of outputs divided by an aggregate index of the inputs necessary to produce those outputs. X may be preset for the regulatory period or it may vary from year to year.

Ideally, X is set according to factors that are exogenous to the firm. This ensures that the firm will not alter its behaviour in an attempt to change X and it can strengthen its incentive properties.

3.3.1. Theoretical considerations

Bernstein and Sappington (1999) provide a theoretical derivation of how an X factor should be determined. They assume multiproduct firms and a regulated sector with multiple firms, and examine four scenarios:

1. The prices of the regulated firms do not affect the general rate of inflation, and all services are included under a single price cap.
2. Some prices are not regulated, or prices are regulated but placed in different tariff baskets, each with its own cap.
3. The general rate of inflation is affected materially by prices in the regulated sector.
4. There is increased competition in some subset of the markets served by the regulated firms.

Case 1: Base case
Let

ΔPI = rate of increase in regulated output prices
ΔT = rate of increase in productivity in the industry
ΔW = rate of increase in input prices in the regulated industry

and ΔP^E, ΔT^E, and ΔW^E be the rate of increase in output prices, productivity, and input prices, respectively, for firms outside of the regulated industry. ΔP^E in this case is the general inflation rate. ΔT and ΔT^E are total factor productivities.[14] They show that the relationship between the rise in output prices in the regulated industry relative to external prices increases is:

$$\Delta PI = \Delta P^E - X_1 \qquad (2)$$

where X_1 is a function of changes in productivity, input costs, profits, and output levels in both regulated and external firms. If all firms earn only normal profit, then X_1 reduces to:

$$X_1 = (\Delta T - \Delta T^E) + (\Delta W^E - \Delta W)$$

and the allowable increase in regulated prices to maintain normal profit becomes:

$$\Delta PI = \Delta P^E - [(\Delta T - \Delta T^E) + (\Delta W^E - \Delta W)]$$

Thus, prices are allowed to rise at the rate of inflation, adjusted for relative changes in productivity and input costs. The X factor is increased if productivity in the regulated sector is higher than externally, or if their input costs are rising less rapidly.

[14] In their analysis, all rates of change are continuous, i.e., changes are derivatives of the prices and productivity with respect to time.

Case 2: Not all services capped
Let

ΔPI^C = rate of increase in regulated output prices
ΔPI^U = rate of increase in unregulated output prices
α = fraction of total revenues from capped services

They show that the firm will earn only normal profit now if the price cap is set so that:

$$\Delta PI^C = \Delta P^E - X_1 - \left(\frac{1-\alpha}{\alpha}\right)(\Delta P^U + \Delta T - \Delta W)$$

which is (2) with an adjustment factor. Bernstein and Sappington (1999) note that the implied adjustment to X may be either positive or negative, depending on how fast the prices of unregulated services are rising relative to changes in input costs and productivity for all the services produced. This means, for instance, that if price increases in unregulated services plus gains from productivity are outweighed by increases in input costs, then prices for capped services have to be allowed to rise more. Conversely, if prices for unregulated services are rising rapidly, input prices are falling, and productivity is increasing, then the cap on regulated services should be tighter.

Case 3: Regulated price increases affect inflation
If the general rate of inflation is affected by prices in the regulated industry, as might be the case with an energy sector, then things become complicated. Bernstein and Sappington produce a two step procedure for deriving the appropriate level of X. Their main conclusion is that the larger the share of GDP the regulated industry accounts for, the less weight should be put on the rate of inflation in the price cap.

Case 4: Structural changes
They note that problems can arise using historic industry data on costs and productivity if there are on-going or anticipated structural changes due to increased competition or to new incentives created by a change in regulatory regime. If a new regulatory regime creates stronger incentives for efficiency, then it may be appropriate to add a "stretch factor" to X to reflect expected improvements over historical performance. Bernstein and Sappington (1999) point out, though, that structural change due to increased competition may imply either an increase or a decrease in X. Competition may drive even greater productivity gains, so that X should be increased. But it could also decrease productivity in the short-run. If the regulated firm has economies of scale and competition results in the firm producing less, then the firm could operate less efficiently because it cannot optimally adjust fixed inputs down to match decreased variable inputs. A reduction in X would then be called for.

New entry could also increase the demand for scarce inputs, driving their prices up and changing the resource mix used by the regulated firm. Finally, competition could sufficiently reduce the profit margin on unregulated services enough to require greater escalation in regulated prices.

3.3.2. Benchmarking and yardstick competition

Benchmarking and yardstick competition are both efforts to create standards for costs, quality, and other factors in a price cap that are exogenous to the regulated firm, and which serve to separate the prices it charges from its own costs. Some authors use the terms benchmarking and yardstick competition synonymously, but there is an important distinction. Benchmarking defines performance expectations based on statistical engineering models that define what a firm "should be" able to achieve. In setting efficiency criteria, it tries to estimate the long run efficient costs of a hypothetical firm using best practices. Yardstick competition uses data on the costs, prices, and quality of other similar firms to evaluate a firm's relative performance. When included as a factor in setting a price cap, this creates strong incentives in theory to improve performance levels (Vogelsang, 2002).

Benchmarking is an example of how practice and theory have often split. It has been used extensively to guide regulatory practices in Chilean and European electricity sectors, by the FCC in the U.S. to estimate telecommunications network costs, and in the UK for airports, electricity distribution, and in the water and sewage sectors (Burns et al., 2005; Stern, 2005; Vogelsang, 2002). In the literature, however, views on the usefulness of benchmarking have been mixed, ranging "…from the strongly enthusiastic to the deeply sceptical" (Stern, 2005, p. 273).

Stern (2005) discusses the major issues being debated on the use of benchmarking efficiency estimates by UK regulators. Summarizing his points:

1. Should the same X be used for all firms in a sector, or should higher efficiency expectations apply to firms that appear to be less efficient, according to the benchmark? Using a higher X for less efficient firms would seem reasonable, since there ought to be more room for such firms to reduce costs. However, benchmark modeling is not an exact science. Unduly optimistic and unequal standards could create significant risks for firms identified as inefficient.
2. Should benchmarks apply only to operating costs, or to both operating and capital costs? Regulators in the UK have not used benchmarks for capital costs. Comparability and measurement of capital costs have been seen as a problem, as well as the fairness of punishing or rewarding current management and shareholders for decisions made by others in the past. There is a concern, though, that leaving them out may distort capital investment decisions. Burns et al. (2005) argue that a firm facing a yardstick only on operating expenses may choose a technology with low

operating costs and high capital costs, even if a more efficient alternative with high operating costs and low capital costs is available.

3. How should benchmarked costs be used: as explicit factors in price caps or as guides to judgment? Stern (2005) argues that estimation techniques are not good enough yet to use benchmarked estimates as explicit factors. However, they can provide valuable information to regulators on which cost drivers to focus on, and the size of potential efficiency gains.
4. How fast should companies be expected to reach benchmarked efficiencies? Stern (2005) reports that there is no way to determine this from theory, and suggests that it may involve many factors unique to the firm, sector, and jurisdiction in question.
5. Should benchmarking be done only for the initial price cap, or should it be repeated in subsequent regulatory reviews? This is another question for which there is no direction from theory, and empirical evidence has been mixed. After analyzing Ofgem's use of benchmarking, Pollitt (2005, p. 286) believes that "serious consideration" should be given to discontinuing their use.

3.3.3. Yardstick indices

There are many variations on yardstick indices used to set efficiency estimates in practice, and each has different implications for efficiency incentives, the risk created, and the difficulty of implementation. Diewert and Fox (2000) provide a thorough analysis of indices based on type and composition. Their results on two specific measures are presented below to illustrate the issues presented by index selection. Theoretical results on the effect of yardstick regulation on investment are presented in Section 4.

Diewert and Fox (2000) consider the choice between using growth in per unit costs and growth in productivity as part of an X factor. They show first that either can be used since minimizing the growth in per units costs is equivalent to minimizing the input price index for the firm divided by the productivity growth rate.

Per unit cost measures depend on a comparison of the cost and output indices for the current period relative to some past period.[15] The regulator can choose to cap per unit cost growth, set a pre-determined reduction, or set a reduction based on industry averages. The first two methods have the advantage of requiring only total cost and output data from the regulated firm. They also reduce the uncertainty the firm faces, since it likely has good information on its own costs and outputs, and any negotiated annual reduction will be known ex ante. However, neither of those methods account for exogenous changes in input costs, which increases the risk to the firm if those costs change.

[15] They recommend choosing a single base year from the past rather than using one-year lagged data. The latter approach may bias the firm's choices in the current year if it knows those choices will affect its indexed goals in the following year.

Using the industry unit cost growth to calculate the annual reduction does factor in changes in input costs. But it does so at a tradeoff: now, the firm faces uncertainty about the parameter, since it cannot be determined until after the fact.

Productivity is measured as the ratio of an index of total outputs to an index of total inputs. Using the firm's productivity growth as a substitute for per unit cost growth requires more data, since now the regulator also needs information on input quantities and prices. Yet more data is required if a yardstick of other firms' productivity is used, since now the regulator needs input data from industry.[16] The advantage of using exogenous data as a yardstick is the strength of the incentives it creates. However, the data could be costly to obtain, agreement on the appropriate peer group may be controversial, and the firm again faces uncertainty, since it must forecast industry productivity.

4. Price Cap Regulation: Commitment and Investment

4.1. Regulatory commitment

Creating good incentives for cost reductions and capital investment requires that the firm has faith that the regulator will not renege on its announced policy. A lack of credible commitment can arise in several ways. There may be pressure on the regulator (internally or externally imposed) to lower prices following successful cost reductions or higher earnings. In the case of a newly privatized utility, the regulator may also be new and not have an established history with the regulated firm (Armstrong and Sappington, 2003). If the firm believes that the profits earned from increasing efficiency this period will be clawed back in the next, or that capital investment will be inadequately rewarded, then it will be reluctant to put effort into reducing costs, invest in necessary infrastructure, or adopt new technology.

Commitment to allow a return on investment is particularly important in sectors where capital investments are large and long-lived. Unfortunately, long run regulatory commitment is not possible, since no regulatory authority today can bind the action of a board twenty years in the future (Lewis and Yildirim, 2002). One possible mitigating factor is that once a firm is privatized, the regulator must answer to some extent to shareholders as well as customers.

4.2. Capital efficiency and technology adoption

Weisman (1993) explores issues of technology adoption and diversification into unregulated markets for a multiproduct firm serving two markets, one regulated and the other competitive, where shared costs exist. He analyzes three forms of regulation:

[16] This may be less of a problem if the other firms are all under the jurisdiction of the same regulator.

1. Cost-based, where total revenues in the regulated market are constrained to be less than directly attributable incremental costs plus an allocation of shared costs.
2. A price cap which is imposed on the regulated market only, and there are no constraints on profit.
3. A modified price cap, where a price cap is imposed on the regulated market, and some profits from that market are shared with customers.

Profit sharing, like the cost-based policy, requires that the shared costs be allocated between the two products, since it is not otherwise possible to calculate the profits from the regulated market alone. Weisman shows that price cap regulation is efficient only if there is no profit-sharing, extending the earlier result by Brauetigam and Panzer (1989) that fully-distributed cost pricing is incompatible with efficiency.

He also finds that modified price caps can be inferior to cost-based regulation, resulting in inefficient technology choices, incentives to misreport costs, under-diversification into the competitive market, and purely wasteful expenditures if the utility believes that above normal profits will be clawed back in subsequent hearings.

Weisman notes that modified price caps were common in new telecommunication policy regimes at that time, and may have been seen as a way to make the transition to a price cap without sharing. Further, sharing may have been necessary to allow the regulator to commit to the new policy regime.

Gilbert and Newbery (1994) examine the question of capital investment in terms of sunk costs and risk. They compare ROR and price cap regulation under different degrees of regulatory commitment to compensate a utility for capital investment. They extend the earlier literature by modeling demand as a random variable and the investment/compensation decisions as a repeated game. The degree of commitment under ROR regulation is defined by the criterion used by the regulator to allow new capital investment into the rate base. They focus on the "used and useful" principle, which states that, ex post, the regulator should allow only that portion of a capital investment that is needed to meet demand. They model three commitment levels: all investment is allowed in, only "used and useful" is allowed, and some portion of "used" is not allowed.

Commitment under price cap regulation depends on prices, which are driven by demand. In high profit periods, the regulator will under reward past capital investment.

Under ROR regulation, they find that the "used and useful" criterion is superior to the others in supporting efficient capital investment. They conclude also that "used and useful" is more likely to result in efficient investment than price cap regulation. Their logic is that ROR regulation in that case yielded efficient investment under a wider range of the variable parameters in their model.

Biglaiser and Riordan (2000) ask how regulatory policy affects the firm's incentive to adopt more efficient technology over time. They assume that technology development is exogenous to the regulated firm, and that newer technologies reduce both operating and capacity costs. An optimal investment path is defined as maximizing the net present value of social welfare, subject to a breakeven constraint for the firm.

Under ROR regulation, prices are set to recover operating costs, a normal rate of return on capital, and depreciation. In practice, depreciation is treated as an expense using an accounting formula that only accounts for asset cost and expected life. They argue that this underestimates the rate of actual depreciation in the value of the asset when better technology is available. As a result, ROR regulation leads to early overinvestment in capacity and underinvestment in better technologies later.

They show that price cap regulation where price is set equal to long run marginal costs will lead to optimal investment. However, two problems with implementation arise. First, when there is a transition from ROR to price cap regulation, setting the price cap equal to long run marginal costs will create stranded costs. This occurs because the rate base assumed under ROR is greater than its actual value and the price cap would only reflect actual value.[17] They suggest that this could be resolved by treating stranded costs as an additional fixed cost to be recovered by a Ramsey markup on price.

The second problem arises from an inability of the regulator to commit to price cap regulation. If regulatory periods are too short, or if there is a concern that policy will revert to ROR regulation, then price cap regulation does no better than ROR regulation and results in much the same incentives to over invest early and under invest later.

Lewis and Yildirim (2002) also look at the firm's incentive to adopt innovative technology. They model a monopolist in a dynamic environment, where some costs are unknown, innovation is possible, costs may increase or decrease over time, and regulatory policy is renegotiated each period. In contrast to other analyses, they find that renegotiation and a lack of regulatory commitment does not eliminate the firm's incentive to invest in cost-saving technology under two conditions: exogenous costs are falling, and the regulator allows the firm to keep a greater share of the innovation benefits.

[17] Note that Biglaiser and Riordan assume that the firm continues as a regulated monopoly. This is different from the problem of changing asset value when a firm's market is deregulated. Moving to a competitive environment may either increase or decrease the value of the deregulated firm's assets.

This result is driven by an assumption that the firm gains experience over time, and that the experience increases the probability of successfully adopting better technologies. This, in turn, increases the probability of reducing costs at an increasing rate in the future. Both consumers and the firm benefit over time, although innovation takes place more slowly than is socially optimal.

The “experience” aspect of the model also leads them to conclude that a regulator should grant long-term franchises to a single firm when exogenous costs are falling, since accumulated learning increases the rate of innovation.

4.3. Yardstick competition

Shliefer (1985) was the first to propose yardstick competition, which uses other utilities’ costs as an external measure to set limits on a utility’s rates. He argued that ROR regulation provides no incentives to reduce costs, and that simply increasing the length of a price cap period is also ineffective, since the firm will anticipate lowered prices if they reduce costs. In a static setting, he showed that yardstick competition results in a socially efficient cost reduction.

Concerns have been raised about the effect of yardstick competition on capital investment. Sobel (1999) argues that firms may be unwilling to invest in more efficient technologies, since it means that the firm may not be able to capture as much of the gains from its investment. Absent the yardstick, it captures all gains from cost reductions not shared with consumers. With the yardstick, its gains may be reduced if it is still relatively less efficient than other firms.

Dalen (1998) examines the incentives of a firm to invest in more efficient technologies if it knows it will be subject to yardstick regulation, and the regulator cannot commit to how any realized gains will be shared with consumers after the investment is sunk. He considers two types of investment: industry-specific and firm-specific. Industry-specific technology reduces the costs of all firms in the industry; firm-specific technology benefits only the firm undertaking the investment. He finds that no investment in industry-specific technology takes place under a yardstick regime. This has particular implications for vertically-integrated utilities that own and operate networks, as investments in improving network technology convey benefits on all firms using the network. The results are the opposite for firm-specific technology: yardstick competition increases the incentive to invest. Intuitively, this is because the firm gains not only from lowering costs, but also from improving its position relative to other firms.

Another issue raised with Shliefer’s model of yardstick competition was that it imposed onerous information requirements on the regulator. Meran and von Hirschhausen (2009) propose a modified yardstick that uses only average output levels of all other firms, and constrains the total revenue of the firm to be no greater than the average of all other firms’ total costs. They find that it yields efficient prices in a static setting. It also converges to efficient prices over time, but slowly enough that the effect on consumers surplus is ambiguous.

5. Quality Regulation: Theoretical Results

The regulation of service quality levels has always been a concern. The issue has become more critical with the introduction of price cap regulation, since one way to cut costs is to reduce service quality.[18] Quality is much harder to regulate than price. Some types of quality, such as customer satisfaction, are hard to measure. If quality is unobservable, the regulator must find proxy variables or create indirect incentives. Customers may also not accurately gauge the level of service received, which makes it difficult to fairly assess performance and justify bonuses or penalties based on customer surveys.

Often, customers have different preferences over desired quality. How does the regulator choose the optimal level of quality if only one level can be provided? And if more than one level can be provided, how can incentives be created for the firm to offer the optimal menu of service levels at efficient prices?

Service quality is also often affected by factors outside of the firm's control. For example, the overall quality of a telecommunication network depends on the quality offered by each member of the network. Interdependence makes it difficult to measure any single firm's performance, again making enforcement of quality standards difficult.

5.1. Quality issues independent of price caps

Some regulatory issues are inherent in quality regulation, independent of incentives created by price caps. Given the current policy debate on whether to include quality benchmarks in price caps, it is important to understand how quality issues arise (Currier, 2007).

5.1.1. Market structure and service levels

Left to themselves, firms will generally not provide socially optimal levels of service quality. If only one service level can be provided, a single product monopoly may offer more, the same, or less than the efficient levels, depending on how customers value quality. So the regulator, without detailed knowledge of customers' preferences, will not know whether it should be setting a quality ceiling or a quality floor.

When customers have different preferences for quality and more than one level can be offered, a monopolist chooses neither efficient levels nor prices. Suppose there are two customer classes, one valuing quality more than the other. Then the monopolist will provide two service levels, but it will set the bar too low on its basic offering. Intuitively, this is because the willingness to pay for higher quality

[18] Except as otherwise noted, Section 5 is based on Sappington's 2005 review of the theoretical literature on regulation of service quality.

depends on both the absolute level and the difference between the quality levels. The monopolist can therefore increase the price it charges for the high quality service by decreasing the lower quality level. It must also reduce the price it charges for the lower quality service, but that is offset by the revenues gained from increasing the price of the high quality service.

Two different issues may arise under duopoly. First, the firms will tend to reduce quality levels in general, since high quality offerings by one firm will be met by price cuts from the other. Second, if the quality of services offered is similar, then consumers will base their choices on price. Firms thus have an incentive to differentiate their products and reduce price competition. The resulting mix of quality levels may not be optimal, and lack of competition results in higher prices.

5.1.2. Imperfect information

If consumers cannot accurately judge quality (either before or after purchasing the service) and firms cannot signal quality, then quality levels tend to be depressed. Decreasing quality reduces costs but not revenues, since demand is not affected by quality.

5.1.3. Networks

Network quality raises two distinct issues: externalities and strategic behaviour. Externalities arise because improving quality in one network can benefit users of other networks connected to the first one. It will usually be impossible to charge those other users for this benefit; hence, the quality level chosen will be less than socially optimal. The standard economics solution for this is to merge the networks, since then all benefits are internal to a single provider. However, this is problematic if the firms are investor-owned or if they span more than one jurisdiction.

If the firm that owns the network also competes downstream with other companies, then downgrading service to competitors could be a concern. Downgrading quality is also a concern if there is one provider with a large network interconnected with other small networks. If the large network downgrades its interconnections with the others, then users of the small networks are more likely to experience lower service quality than users of the large network. This happens simply because customers in the large network are less likely to connect with users on the outlying networks. The large network can thus make itself more attractive to users of the smaller networks without doing much harm to its existing customers.

5.2. Regulating Retail Service Quality

The welfare effects of price caps when quality is an issue can be either positive or negative. If a monopolist only offers one service level, then a fundamental problem arises if allowable rate increases are tied to exogenous costs. Raising quality increases the firm’s costs, so it has a disincentive to raise quality, and a strong incentive to lower it. A monopolist offering a single quality level will always provide less than the socially optimal service when price is not tied to its actual costs. Where quality disincentives exist, they can be ameliorated by allowing rates to reflect some changes in actual costs.

In the two quality level case, however, price caps may improve welfare. Price caps limit the ability of the monopoly to raise the price of the higher quality service. Its incentive to decrease the level of the low quality service is thus diminished, since it can no longer take advantage of the inelasticity of demand for the higher quality service.

A new issue arises when an industry is moving from ROR regulation to a regulatory regime which includes both price caps and quality regulation. The regulated firm may make changes in quality while it is still under ROR regulation and can recover those costs as a one-time capital expense. They then have the benefit of higher quality going forward into a price cap that is linked to quality. This may lead to overinvestment in quality in anticipation of moving to the new regulatory regime.

5.2.1. Minimum quality standards

When measurable and verifiable, minimum quality standards are easy to apply. However, whether they increase or decrease social welfare depends on many factors, few of which the regulator is likely to be able to observe or control. A minimum quality standard will increase welfare, all else equal, if:

1. It constrains the utility from overly downgrading the lowest service levels it offers.
2. It discourages duopolists from using quality differences to reduce price competition, in the case where quality increases affect fixed costs but not marginal costs.

Standards can result in decreases in welfare if:

1. The standard is set too high and results in too high a price
2. Firms are price-takers in an oligopolistic market. In equilibrium, the firms supply too much low quality and too little high quality, both at prices that are too high.
3. Incumbents can use them strategically to deter entry. For example, a minimum quality standard may impose higher costs on entrants than on the incumbent.
4. Innovations that reduce the cost of providing high quality may be met with higher standards in subsequent review periods.

5.2.2. Bonuses and penalties

The U.S. has used bonus and penalty systems to increase the incentive to meet quality standards in the telecommunication and electricity sectors. Few regulators have attempted to devise schemes that cover all dimensions of service, nor has the literature addressed the (fairly intractable) question of service with multiple dimensions (Tangerås, 2009).

If quality is measurable, penalties can be imposed for specific incidents. Sappington gives an example of requiring rebates of some or all of an installation charge if the installation of a service is delayed. Penalties combined with a complaint system can be used by the regulator to reduce monitoring costs. When quality is verifiable, complaint systems harness the customer's superior knowledge of the actual quality received.

Quality indices can also be built into the price cap formula, where decreases in quality below a given standard cause the cap to fall faster. The empirical evidence on the effectiveness of linking quality and price reductions is mixed.[19]

Often, quality is difficult or prohibitively expensive to measure. The regulator's best options then depend on several factors. If neither the regulator nor customers can observe quality, then the regulator should tie rates more closely to actual costs if high quality is important. Otherwise, the firm has too much incentive to reduce quality.

If customers can observe quality, but the regulator cannot, then the appropriate measure depends on whether customers view quality and quantity as complements or as substitutes.[20] If they are complements, then the regulator needs to allow greater price increases for the service. This creates an incentive for the utility to increase quality, since higher quality spurs higher demand and greater profit. The opposite action is called for if they are substitutes. Then the regulator needs to put downward pressure on price in order to create an incentive to increase quality. Lower allowed prices lead the firm to want lower customer demand (assuming that marginal costs fall as output decreases), and increased quality will reduce demand if quality and quantity are substitutes.

When the regulator observes quality, but customers cannot, then setting standards is again an option. The regulator might also consider ways of making quality information available to customers.

5.3. Regulating Wholesale Service Quality

The appropriate policies for regulating retail and wholesale service are different, since the latter must consider how service quality levels can be used strategically to disadvantage competitors in retail markets. A vertically integrated firm that is a regulated monopoly in the wholesale market, and which competes with other firms in a retail market, may have incentives to provide lower quality to retail

[19] See Ai et al. (2004) for an empirical analysis of quality standards in telecommunications.

[20] An example of complements is water purity: high confidence in the quality of tap water makes it less likely that consumers will buy bottled water for consumption. An example of substitutes might be more informative billing for water, heating, or electricity: bills showing monthly historic use might create an incentive to reduce usage during peak months.

competitors. This has been shown to be an issue regardless of whether firms engage in price or quantity competition in the retail market.[21]

The logic is as follows. Retail competitors' costs increase if the quality of the wholesale input is lowered, meaning that their prices must rise. Under price competition, this increases the demand for the retail affiliate of the vertically integrated firm, which in turn increases its profit. Under quantity competition, an increase in competitors' costs results in a reduction in their output. This drives the market price up, which again increases the profit of the retail affiliate.

An exception to this is if the vertically integrated firm also derives revenue from the sale of the wholesale service to the downstream rivals. Decreases in quality could depress wholesale demand enough to offset its gains in the retail market.

In theory, the regulator could apply the same type of policies discussed for retail service quality: benchmarks for key services, with a system of rewards and penalties. However, the regulator is unlikely to have the data to determine socially optimal quality levels. Alternatives that have been suggested are parity standards, wholesale price regulation, and bargaining with carrots and sticks.

5.3.1. Parity standards

Under a parity standard, the utility is obligated to provide no worse service to retail competitors than it gives its own retail affiliate. Failure to meet the parity standard results in penalties that are paid to the retail competitors; delivering quality above the standard is generally not rewarded. One problem with this approach is that it creates no incentives to increase quality above the standard. Other issues that need to be recognized are:

1. Exogenous factors can affect quality. If both penalties and rewards were associated with quality of service, then random changes that reduce quality might be expected to be offset by random changes that increase quality. When there are only penalties, the utility is punished even if it provides an appropriate level of service on average.
2. If quality can also be affected by the buyer, then it is possible that retail competitors would strategically lower quality if it received enough compensation from the penalties received.
3. Lowering quality across the board may be profitable if the profit resulting from damage inflicted on competitors more than offsets the losses of its own retail affiliate.

5.3.2. Regulation of wholesale prices

Sappington (2005) sees no clear direction from the literature on how to use price regulation to address quality concerns. There are some parallels with the previous discussion on retail service quality and whether quality and quantity of

[21] Under price competition, firms set prices and then produce the quantity demanded at those prices. Under quantity competition, firms are price takers: they determine their output levels, and the market price then depends on the total quantity produced by all firms.

the wholesale service are complements to retail firms. One suggestion has been to encourage higher quality by allowing higher wholesale prices. If the retail market is not competitive, though, higher wholesale service prices lead to higher than optimal retail prices.

5.3.3. Carrots and sticks

Regulatory authority can be wielded in more ways than setting standards and prices. Concerns about retail affiliate relationships have often led regulators to bar incumbent monopolies from competing in retail markets. One carrot that has been used in the U.S. telecommunications sector was to make access to long-distance markets conditional on the quality of wholesale services in the local market. Sappington (2005) points out that this requires careful implementation, citing an empirical study that showed wholesale quality decline after the incumbent gained access to the long-distance market.

The threat of forced divestiture for incumbents who offered retail services prior to deregulation has also been suggested as a stick. Even when this is politically feasible, it could be a costly exercise. In addition, it may result in a loss of economies of scope, decreasing the efficiency of production. It may also reduce the incentive to invest in innovation if some of the benefits of innovation are realized at the retail level.

6. Conclusion

One recurring theme in the literature was how practice and theory split from the late 1980s on, with new developments in the theoretical literature having little or no effect on the design and implementation of new regulatory regimes. Crew and Kleindorfer (2002) believed that this was driven, in part, by a difference in the objectives of economists and governments. Economists were seeking ways to minimize the inefficiencies associated with natural monopolies; politicians were seeking to reduce regulatory costs or to deregulate, subject to pressure from consumers and industry stakeholders. As a consequence, restructuring of regulatory regimes "…resulted in a piecemeal and selective implementation of changes and reforms…that [seemed] to be the most easily implemented without addressing some of the fundamental underlying problems" (Crew and Kleindorfer, 2002, p. 9).

Laffont (1994) had a different perspective. He believed that the split was developing because economists had failed to take up the challenge of adapting theoretical recommendations to institutional context and political constraints. While it is necessary to start with abstract, stylized models, it is also necessary to then develop them into relevant policy recommendations.

Bridging the gap between theory and practice requires communication that is often lacking. It is a mistake for economists to recommend dramatic policy changes if they do not fully appreciate the situation to which they will be applied, as California's experience with electricity deregulation illustrates. It is equally unconstructive for regulators to adopt policies that have been discovered to have serious flaws, or that can be used well, but only if implemented with care.

The U.K. seems to have put more emphasis on this type of communication than other jurisdictions. For example, there were conferences in 2004 and 2005 that brought theorists and practitioners together to discuss efficiency estimates and capital investment. These led to special issues of *Utilities Policy* devoted to those questions, which disseminated the findings to the larger academic community. Similar activities could prove useful here, particularly for provinces that are moving to new regulatory regimes.

References

Acton, J., & Vogelsang, I. (1989). Introduction. *RAND Journal of Economics*, 20(3), 369-372.

Ai, C., Martinez, S. and Sappington, D. Incentive Regulation and Telecommunications Service Quality. *Journal of Regulatory Economics*, 26(3), 263-285.

Armstrong, M. and Sappington, D. (June 2003). Recent Developments in the Theory of Regulation. Draft chapter for *Handbook of Industrial Organization*, Volume III. Retrieved January 10, 2010 from: http://www.ppge.ufrgs.br/giacomo/arquivos/regulacao2/armstrong-sappington-2003.pdf

Armstrong, M. and Sappington, D. (June 2006). Regulation, Competition, and Liberalization. *Journal of Economic Literature*, XLIV, 325-366.

Armstrong, M. and Vickers, J. (1991). Welfare effects of price discrimination by a regulated monopolist. *RAND Journal of Economics*, 22(4), 571-580.

Baron, D. and Myerson, R. (1982). Regulating a Monopolist with Unknown Costs. *Econometrica*, 50(4), 911-930.

Beesley, M. and Littlechild, S. (1989). The regulation of privatized monopolies in the United Kingdom. *RAND Journal of Economics*, 20(3), 454-472.

Berg, S. (2005). Introduction to the Special Issue on service quality. *Utilities Policy*, 13(3), 187-188.

Bernstein, J. and Sappington, D. (July 1999) Setting the X Factor in Price-Cap Regulation Plans. *Journal of Regulatory Economics*, 16(1), 5-26.

Biglaiser, G. and Riordan, M. (2000). Dynamics of price regulation. *RAND Journal of Economics*, 31(4), 744-767.

Braeutigam, R. and Panzar, J. (1989). Diversification incentives under "price-based" and "cost-based" regulation. *RAND Journal of Economics*, 20(3), 373-391.

Braeutigam, R. and Panzar, J. (1993). Effects of the Change from Rate-of-Return to Price-Cap Regulation. *American Economic Review*. 83(2), 191-198.

British Columbia Utilities Commission (December 2009). Terasen Gas Inc. Terasen Gas (Vancouver Island) Inc. Terasen Gas (Whistler) Inc. and Return on Equity and Capital Structure Decision. Retrieved January 10,

2010 from
http://www.bcuc.com/Documents/Proceedings/2009/DOC_24241_TUS_ROE-Decision-Web.pdf

Burns, P., Jenkins, C. and Reichmann, C. (2005). The role of benchmarking for yardstick competition. *Utilities Policy*, 13(4), 302-309.

Burns, P. and Riechmann, C. (2004). Regulatory instruments and investment behaviour. *Utilities Policy*, 12(4), 211-219.

Cowan, S. (1997). Price-Cap Regulation and Inefficiency in Relative Pricing. *Journal of Regulatory Economics*, 12(1), 53-70.

Crew, M. and Kleindorfer, P. (January 2002). Regulatory Economics: Twenty Years of Progress? *Journal of Regulatory Economics*, 21(1), 5-22.

Crew, M. and Kleindorfer, P. (1996). Incentive Regulation in the United Kingdom and the United States: Some Lessons. *Journal of Regulatory Economics*, 9(3), 211-225.

Currier, K. (2007). A practical approach to quality-adjusted price cap regulation. *Telecommunications Policy*, 31(8-9), 493-501.

Dalen, D. (1998). Yardstick Competition and Investment Incentives. *Journal of Economics & Management Strategy*, 7(1), 105-126.

Diewert, W. (1992). Fisher Ideal Output, Input, and Productivity Indexes Revisited. *The Journal of Productivity Analysis*, 3, 211-248.

Diewert, W., and Fox, K. (2000). Incentive Indexes for Regulated Industries. *Journal of Regulatory Economics*, 17(1), 5-24.

Doucet, J. and Littlechild, S. (2006). Negotiated settlements: The development of legal and economic thinking. *Utilities Policy*, 14(4), 266-277.

Gilbert, R. and Newbery, D. (1994). The dynamic efficiency of regulatory constitutions. *RAND Journal of Economics*, 25(4), 538-554.

Jamasb, T. and Pollitt, M. (December 2007). Incentive regulation of electricity distribution networks: Lessons of experience from Britain. *Energy Policy*, 35(12), 6163-6187.

Kahn, A. (January 2002) “The Deregulatory Tar Baby: The Precarious Balance Between Regulation and Deregulation, 1970-2000 and Henceforward.” *Journal of Regulatory Economics*, 21(1), 35-56.

Laffont, J. J. (May 1994). The New Economics of Regulation Ten Years After. *Econometrica*, 62(3), 507-537.

Lewis, T. and Yildirim, H. (2002). Learning by doing and dynamic regulation. *RAND Journal of Economics*, 33(1), 22-36.

Lewis, T. and Sappington, D. (1989). Regulatory options and price-cap regulation. *RAND Journal of Economics*, 20(3), 405-416

Lewis, T. and Sappington, D. (1991). Incentives for monitoring quality. *RAND Journal of Economics*, 22(3), 370-384.

Lewis, T. and Sappington, D. (1997). Penalizing success in dynamic incentive contracts: no good deed goes unpunished? *RAND Journal of Economics*, 28(2), 346-358.

Liston, C. (1993). Price-Cap versus Rate-of-Return Regulation. *Journal of Regulatory Economics*, 5(1), 25-48.

Littlechild, S. (2009). The bird in hand: Stipulated settlements in the Florida electricity sector. *Utilities Policy*, 17(3-4), 276-287.

Lowry, M. and Kaufmann, L. (2002). Performance-based regulation of utilities. *Energy Law Journal*, 23(2), 399-457.

Lyon, T. (1991). Regulation with 20-20 hindsight: “Heads I win, tails you lose”? *RAND Journal of Economics*, 22(4), 581-595.

Lyon, T. (1996). Evaluating the Performance of Non-Bayesian Regulatory Mechanisms. *Journal of Regulatory Economics*, 9, 41-60.

May, P. (2003). Performance-Based Regulation and Regulatory Regimes: The Saga of Leaky Buildings. *Law & Policy*, 25(4), 381-401.

Meran, G. and von Hirschhausen, C. (2009). A modified yardstick competition mechanism. *Journal of Regulatory Economics*, 35(3), 223-245.

National Energy Board (2009). The Regulation of Traffic, Tolls and Tariffs. Retrieved January 10, 2010 from http://www.neb.gc.ca/clf-nsi/rthnb/whwrndrgvrnnc/rgltntrffctllstrffs2007-eng.html

Navarro, P. (1996). The Simple Analytics of Performance-Based Ratemaking: A Guide for the PBR Regulator. *Yale Journal on Regulation*, 13, 105-162.

Neu, W. (1993). Allocative *In*efficiency Properties of Price-Cap Regulation. *Journal of Regulatory Economics*, 5(2), 159-182.

Pollitt, M. (2005). The role of efficiency estimates in regulatory price reviews: Ofgem's approach to benchmarking electricity networks. *Utilities Policy*, 13(4), 279-288.

Sappington, D. (1982). Optimal regulation of research and development under imperfect information. *Bell Journal of Economics*, 13(1), 354-368.

Sappington, D. (1983). Optimal Regulation of a Multiproduct Monopoly with Unknown Technological Capabilities. *Bell Journal of Economics*, 14(2), 453-463.

Sappington, D., Pfeifenberger, J., Hanser, P. and Basheda, G. (October 2001). The State of Performance-Based Regulation in the U.S. Electric Utility Industry. *The Electricity Journal*, 71-79.

Sappington, D. (2005). Regulating Service Quality: A Survey. *Journal of Regulatory Economics*, 27(2), 123-154.

Sappington, D. (2006). Regulation in Vertically-Related Industries: Myths, Facts, and Policy. *Review of Industrial Organization*, 28(1), 3-16.

Schmalensee, R. (1989). Good regulatory regimes. *RAND Journal of Economics*, 20(3), 417-436.

Shliefer, A. (1985). A theory of yardstick competition. *RAND Journal of Economics*, 16(3), 319-327.

Sibley, D. (1989). Asymmetric information, incentives and price-cap regulation. *RAND Journal of Economics*, 20(3), 392-404.

Sobel, J. (1999). A Reexamination of Yardstick Competition. *Journal of Economics & Management Strategy*, 8(1), 33-60.

Stern, J. (2005). UK regulatory price reviews and the role of efficiency estimates. Utilities Policy, 13(4), 273-278.

Stern, J. (2006). Capital efficiency and infrastructure regulation: Editorial introduction. *Utilities Policy*, 14(4), 219-223.

Tangerås, T. (2009). Yardstick Competition and Quality. *Journal of Economics & Management Strategy*, 18(2), 589-613.

Vogelsang, I. (2001). Price Regulation for Independent Transmission Companies. *Journal of Regulatory Economics*, 20(2), 141-165.

Vogelsang, I. (2002). Incentive Regulation and Competition in Public Utility Markets: A 20-Year Perspective. *Journal of Regulatory Economics*, 22(1), 5-27.

Vogelsang, I. (2006). Electricity Transmission Pricing and Performance-based Regulation. *The Energy Journal*, 27(4), 97-126.

Weisman, D. (1993). Superior Regulatory Regimes in Theory and Practice. *Journal of Regulatory Economics*, 5(4), 355-366.