THE FIFTY MOST IMPORTANT PAPERS IN THE ECONOMICS OF REGULATION

Darryl Biggar


ACCC/AER WORKING PAPER SERIES
About the Author

Dr Darryl Biggar is an economist and a full-time consultant to the ACCC. His fields of interest are the economics of wholesale electricity markets and the economics of public utility regulation, particularly questions related to the foundations of public utility regulation. In addition to providing economic advice on regulatory issues, Dr Biggar is involved in the training programme of the ACCC/AER and the Australian Energy Market Commission.

He is very active as a presenter and discussant in the diverse forums that the ACCC promotes to stimulate thinking on the theory and practice of regulation.

Dr Biggar has published several articles in economics journals and in the engineering literature. He has co-authored papers with academics from the University of Adelaide and the Royal Institute of Technology in Sweden. In 2010 one of his papers won the Inaugural Distinguished Scholar prize from the International Confederation of Energy Regulators.

Prior to joining the ACCC in 2002, Dr Biggar worked with the OECD in Paris, on matters related to economic regulation and competition. He has also worked for the New Zealand Treasury and as an academic at University College London. He holds a PhD in economics from Stanford University and an MA in mathematics from Cambridge University, UK.
## Contents

**Foreword** ............................................................................................................................................... v

I. Debates over the Foundations of Economic Regulation ................................................. 3

II. Pricing I: The Neoclassical Approach .................................................................................. 10

III. Pricing II: Cost Allocation ..................................................................................................... 17

IV. Pricing III: Rate Base and Depreciation ............................................................................. 19

V. Averch–Johnson and Incentives under Rate-of-Return Regulation .................................... 23

VI. The Rise of ‘Incentive Regulation’ ....................................................................................... 27

VII. Pricing IV: The Weighted Average Price Cap ................................................................. 34

VIII. The Rise of ‘The New Economics of Regulation’ ........................................................... 38

IX. Yardstick Regulation and Benchmarking .......................................................................... 43

X. One-Way Access Pricing ....................................................................................................... 49

XI. The Regulation of Quality .................................................................................................... 55

XII. Vertical Integration and Non-Price Discrimination ........................................................... 61

XIII. Risk, Uncertainty and Cost of Capital Issues ................................................................. 67

XIV. Two-Way Interconnection ................................................................................................. 72

XV. The Design of Regulatory Institutions ............................................................................ 75

Commentary .................................................................................................................................... 80

Conclusion ....................................................................................................................................... 81

Appendix: Top Fifty Papers in Regulatory Economics ........................................................... 82

References ........................................................................................................................................ 86
FOREWORD

Many economists working full-time in the field of public utility regulation – either in regulatory agencies, economic consultancies, or in regulated firms – face an uphill battle to keep up with the ever growing literature in the field. It is thus with great pleasure that I welcome Darryl Biggar’s mapping of the landmark papers in regulatory economics. This is a quick and readable introduction to the fifty most significant papers in the field of regulatory economics in the past forty or so years. It brings the reader up to speed on both the key debates in regulation and, to an extent, the state of the art. From a historical perspective, this is an excellent roadmap to the major issues of regulatory theory and practice. It is also a nice summary of the on-going and unresolved problems that will continue to be the focus of regulation in the years ahead.

The selection of the fifty most significant papers in any field is naturally a subjective exercise. Although we would expect substantial overlap, one economist’s list of the fifty most important papers will inevitably differ from another’s. In addition, the selection of the most significant papers in a field will evolve over time, as certain lines of thinking become dominant while others die out. This survey, therefore, partly reflects the perspective of one person at a point in time. Dr Biggar has chosen to view the landmark papers of the past through the lens of the transactions cost approach to regulatory economics. Although the field of transactions cost economics has grown exponentially since it first emerged in the 1970s, and has been recently highlighted by the 2009 Nobel Prize in Economics to Oliver Williamson, the application of transactions cost economics to public utility regulation is only one of many lenses that might be used to analyse regulation and regulatory institutions. Nevertheless, as Dr Biggar makes clear in this survey, this is a valid and useful lens through which to view the literature on public utility regulation.

The field of public utility regulation is large, and several themes were of necessity excluded. Dr Biggar has not, for example, discussed the issues of reliability of service (which is particularly important in the electricity sector) or universal service obligations (which are particularly important in telecommunications and the postal sector). Nevertheless, what Dr Biggar has accomplished is to provide a concise and informative account of some of the landmark contributions to regulatory economics of the past 40 years, organised into a readable summary of the major themes that have occupied the profession in this period. This is a most welcome contribution to the field.

Paul Kleindorfer
Anheuser-Busch Professor of Management Science, Emeritus, the Wharton School of the University of Pennsylvania; and Paul Dubrule Professor of Sustainable Development, INSEAD
The Fifty Most Important Papers in the Economics of Regulation

Regulatory economists and practitioners typically seek to keep abreast of developments in their field by scanning journals’ content pages or reading summaries of recent papers. But only a few of the hundreds of papers published in any one year will have the influence or significance of the major or landmark papers from the past. This paper seeks to fill a gap by summarising the fifty most important papers in the field of regulatory economics. In each case an attempt has been made to put the paper in its historical context, to explain the key contribution of the paper, and to show the impact of the paper on subsequent practical and theoretical developments.

Of course, the selection of the fifty most important papers involves a degree of subjectivity and will inevitably be somewhat idiosyncratic. To help the reader understand the choices made, I make the following comments.

- First, I have taken ‘most important’ to mean the most valuable or most useful papers to the staff of a regulatory authority. This does not necessarily reflect the importance or significance of these papers within the academic economics profession, or the number of citations to the paper in academic journals. A paper which introduces a new theoretical mathematical technique may be important to the academic literature and widely cited, but may not be directly relevant to the day-to-day activities of a regulatory authority. Some of the papers included in this list are themselves survey papers, which are not likely to attract a large number of citations, but remain valuable to regulatory practitioners. I have deliberately excluded papers which are highly technical and which would not normally be considered accessible to a staff member of a regulatory authority.

- Second, an attempt has been made to cover most of the fields relevant to a regulatory authority. In practice, the economics literature has followed trends in thinking, fashions, and what were, at the time, promising lines of enquiry. The economics literature does not focus equal attention on all of the issues that confront a regulator in practice. In the case of some topic areas there are a large number of relevant economics papers – in other areas, very few. Some topics of interest to a regulatory authority (such as mechanisms for minimising the impact of the five-year regulatory cycle on the cycle of incentives on the regulated firm), have simply no relevant academic papers at all. To an extent, the number of articles under each heading here is more a reflection of the number of relevant articles in the economics literature than a reflection of the importance of the topic to a regulatory authority.

- Third, to an extent the selection and presentation of the articles set out here reflects my view that public utility regulation is best viewed through the lens of transactions cost economics. Specifically, public utility regulation is best viewed as a form of long-term contract. That contract is designed to solve the hold-up problem arising from the need to make material sunk, relationship-specific investments, particularly by the customers. As the papers in this survey show, this perspective was first proposed in the 1970s. Although there are a number of papers that recognise the importance of sunk investments by service providers, very few papers have explored the consequences of sunk investments by
customers. Nevertheless, in my view, this perspective offers strong promise as a coherent approach to understanding the foundations of public utility regulation.

Inevitably, some potential topics have not been covered here. For example, I have deliberately excluded papers on universal service obligations or on the liberalisation and market structure decision. I have also excluded papers which address issues unique to a particular industry.

The papers here are primarily drawn from economic journals. However I have also included articles published in the Handbook of Economics series and at least one unpublished survey article. Many of the survey articles from the Handbook of Economics are valuable, integrated presentations which summarise the state of knowledge of an entire topic area at a particular point in time. In some instances the most valuable reference for a particular topic will be an economics textbook, but these were excluded from this survey.

For each paper included in the top fifty, there are typically two or three related papers which might have made it into the top fifty and are at least worthy of mention. The full citations to these related papers are included in the references at the end.

The fifty selected papers have been grouped into fifteen topic areas. Each of these topic areas is discussed in turn below.
I. Debates over the Foundations of Economic Regulation

From the late 1950s, through the 1960s and early 1970s, economists at the University of Chicago produced a number of articles criticising all forms of economic regulation. This critique was given voice in the pages of the Journal of Law and Economics – first published in the late 1950s.

For example, in a 1962 paper titled ‘What Can Regulators Regulate?’, George Stigler and Claire Friedland presented the results of a study of regulated price structures and share prices immediately before and after the establishment of state public utility commissions in the first decades of the 20th century. Stigler and Friedland claimed that the introduction of public utility regulation of electricity utilities had no discernible impact – on price structures, on prices to small consumers, or on share prices of regulated firms. Their conclusion was that regulation was not only ineffective, but unnecessary, as electricity faced effective competition from other sources of energy (such as oil or coal).

A few years later, also in the Journal of Law and Economics, Demsetz (also reviewed below) fundamentally challenged the prevailing wisdom of the need for public utility regulation of natural monopolies – arguing instead for a form of tendering or franchising for monopoly services. Demsetz hinted that the real purpose of natural monopoly regulation was to serve the economic interests of the monopolies themselves.


George Stigler took up this latter idea and ran with it. His 1971 paper was not just about natural monopoly regulation. Rather, it asks the broader question: why do governments directly intervene in some industries and not in others? The most common view at the time was that government intervention was primarily directed at the correction of certain market failures. It was recognised, however, that the precise form of that intervention was the result of the haphazard, imperfect, and, at different times, both high-minded and self-seeking political processes.

In contrast, Stigler suggested that there is a discernible underlying order to government intervention in the economy. Specifically, he argued that economic regulation will be supplied by policy-makers to meet the economic interests of the regulated industry – in the form of subsidies, controls on entry, limitations on the rate of expansion of new firms, or direct controls on prices (particularly floors on prices). If the costs imposed on the rest of society are small enough, he argued, it will not be worthwhile for individuals to oppose the regulation through the political process, even though it reduces economic welfare overall. Stigler was the first to suggest the idea that regulation was not an exogenous force, directed at the public good, but endogenous to the economic system, potentially directed at the economic interests of firms themselves.

Stigler’s paper had an immediate and substantial impact. The notion that regulation and law-making may be endogenous to the economic system spawned the development of the new field of theoretical political science. This paper comes in at number fourteen in the top fifty most cited papers in the entire economics literature¹ and was explicitly mentioned when Stigler received the Nobel Prize for economics in 1982.

¹ Kim, Morse and Zingales (2006).
Two subsequent papers responding to Stigler’s idea are also among the most cited papers in economics: a paper by Richard Posner (‘Theories of Economic Regulation’, 1974) and by Sam Peltzman (‘Towards a more general theory of regulation’, 1976). Posner’s paper provides a bit more historical context and directs certain criticisms at Stigler’s main premise, but is broadly supportive. Peltzman mathematically formalises Stigler’s argument.

With the benefit of more than thirty years of hindsight Stigler’s fundamental premise looks somewhat shakier than it did in 1971. If regulation is primarily intended to promote producer interests, what can explain the substantial movement towards deregulation (in airlines and trucking, for example) within a decade of Stigler’s paper? It seems more plausible to argue that Stigler’s article highlighted an oversight for policymakers: that the regulations they thought were promoting the public interest were in fact not doing so. This resulted in both a roll-back in economic regulation and the establishment of new mechanisms (such as requirements for all new regulations to pass an economic cost-benefit test) to prevent inefficient regulations being implemented in the future.

Stigler’s paper does not directly address natural monopoly regulation. Despite the significant roll-back in economic regulation in the intervening decades, natural monopoly regulation remains largely intact. More importantly, natural monopoly regulation is not evidently in the interest of the regulated firms – at least it seems to place an upper limit, rather than a floor, on the prices the regulated firm can charge. Given the complaints of regulated firms, it would seem that Stigler’s fundamental premise that economic regulation is primarily implemented to satisfy producer interests just doesn’t apply.

Empirically, Stigler’s claims have met with mixed success. In a survey article published in 1989 in the Handbook of Industrial Organization, Roger Noll summarises the empirical evidence as follows:

> Taken together, the empirical studies … are broadly consistent with, but do not really prove, the political theory of regulation … Organized interests not only seem to succeed, but usually they do so at the cost of economic efficiency, at least as far as the data can tell us. Yet the evidence is still far from conclusive. One major weakness is that, except for simple cases involving one or two interest groups, the relationship between the stakes of groups and their political strengths remains a mystery … The second weakness is the lurking danger of tautology, i.e., of attributing causality to an inevitable consequence of any public policy action. … Until fundamental measurement problems about stakes, power, and gains are overcome, analysts will not be able to fully predict and to explain the details of regulatory policy. Only when they do can it reasonably be argued that interest group theories of regulation have been fully tested.4

---

2 Supporters of the interest-group or ‘capture’ theories claim that this deregulation is primarily the result of technological changes shifting the costs and benefits of regulation.

3 Priest (1993) suggests that the deregulation movement was directly a result of Stigler. He claims that Stigler’s paper ‘significantly contributed to the movement toward broad deregulation of industry which, along with the parallel adoption of the economic approach toward antitrust law, constitutes the greatest intellectual triumph in U.S. public policy of modern times’ (page 290).

In the 1960s, even amongst those economists who were sceptical about some forms of economic regulation, the case for regulation of natural monopolies seemed watertight. This view was challenged in 1968 by another Chicago economist in yet another path-breaking paper published in the *Journal of Law and Economics*.


In a provocatively-titled paper, Harold Demsetz argued that the existence of economies of scale did not automatically imply the need for traditional public utility regulation. Anticipating the later notion of a ‘contestable’ market, Demsetz argued that a firm didn’t need be currently producing a service in order to bid for the right to provide the service. Put another way, a firm need not be currently active in the market to undercut an existing seller. In particular, Demsetz argued that if the right to provide a given service is put out to tender, competition between bidders will ensure that the service is provided at a price just sufficient to cover the total cost.

In addition, Demsetz directly targeted the prevailing wisdom of the time that posited a strong connection between market structure and market performance outcomes. He argued that ‘we have no theory that allows us to deduce from the observable degree of concentration in a particular market whether or not price and output are competitive’.

In effect, Demsetz’s argument is that regulation by public commission can be replaced by auctioning of a contractual arrangement to provide the natural monopoly service. As we will see below, Demsetz’s challenge stimulated an intellectual response which, in the long run, established an entirely new field of economics and clarified the precise nature and role for public utility regulation.


Two key papers challenging Demsetz’s perspective appeared in 1976. The first, by Oliver Williamson, argued that in the presence of uncertainty about future demand or cost conditions, the transactions costs of writing a complete contract, contingent on all future outcomes is prohibitively costly. A once-for-all auction for the provision of a natural monopoly service is, in practice, totally impractical.

On the other hand, periodic re-tendering introduces its own problems. Most public utility industries require substantial sunk, long-lived investment – whether in distribution wires, rail lines, or telecommunications conduits. Where the life of this investment exceeds the life of the franchise, contractual arrangements must ensure continued investment and maintenance of the sunk asset. It may be difficult to verify the quality of the maintenance of the asset ex post. This is particularly the case where the quality dimension includes maintaining the human capital of the staff required to ensure the continued operation of the asset. In addition, the incumbent provider of the service is

---

5 Priest (1993) observed that in these industries: ‘In addition to any noneconomic goals, regulation seemingly can achieve the powerful economic goal of constraining monopoly pricing toward marginal cost in order to enhance social welfare. Indeed, this social welfare justification is so widely accepted that the natural monopoly case is viewed as the prototypical context for government regulation’ (page 296).

6 Page 60.
likely to have better quality information about the likely cost and demand characteristics of the service, providing an informational advantage over potential rival bidders.

To illustrate these ideas, Williamson used a case study based on the experience of tendering cable television franchises in Oakland, California. He concludes:

That franchise bidding for cable television has superior properties to regulation is not transparent … Not only is simple franchise bidding … beset with numerous transactional difficulties, but the institutional infrastructure that predictably develops to check dysfunctional or monopoloid outcomes has many of the earmarks of regulation.7

Using modern language, Williamson highlights and emphasises the importance of transactions costs. Williamson introduced for the first time the notion that natural monopoly regulation can be viewed as a form of long-term contracting. That long-term contract incorporates mechanisms to allow for adjustment to changes in the environment:

At the risk of over-simplification, regulation may be described contractually as a highly incomplete form of long-term contracting in which (1) the regulatee is assured an overall fair rate of return, in exchange for which (2) adaptations to changing circumstances are successively introduced without the costly haggling that attends such changes when parties to the contract enjoy greater autonomy.8


In the same year as Williamson’s critique of franchise bidding, Victor Goldberg published an even more fundamental critique of the conventional view of natural monopoly regulation. Goldberg’s insight was to emphasise that both regulated firms and their customers must make substantial relationship-specific sunk investments. These investments, once made, are subject to the risk of ‘hold-up’ – that is, the threat that the service provider will raise its prices in order to capture some of the value of the investment by the customers. In order to protect these investments both the regulated firm and its customers actively seek to enter into long-term arrangements governing the terms and conditions of supply. These governance mechanisms include vertical integration and long-term contracting.

Public utility regulation, according to Goldberg, does not differ in substance from an explicit long-term contract:

If the regulatory relationship were replaced by a private contract, the problems faced by the private agent would differ mainly in degree rather than in kind from those that plague regulators and provide a field day for their critics.9

Goldberg was the first to emphasise the incompleteness of long-term contracts and the key role of long-term contracts in establishing a process for adjusting the terms of the agreement to changing market conditions over time. He also foresaw the essential

---

7 Page 101.
8 Page 91.
9 Page 441.
similarity between the issues that arise in public utility regulation and the issues in administering a long-term contract:

In searching for a rationale for regulation we should look not at the shape of the long-run average cost curve, but instead at the complexities involved in devising and administering such a contract. Indeed, natural monopoly industries will be characterized in this paper not by their alleged decreasing average costs, but by the features which make long-term relationships between consumers and producers desirable and which further make it extremely difficult to determine at the outset the specific terms of that relationship.10

This problem of relationship-specific investments is, of course, not specific to the field of regulation – it can arise in any economic relationship. In fact, the problems faced by monopoly firms and their customers have close parallels in the relations between private firms and their suppliers, and their decision to vertically-integrate, or to enter into a long-term contractual arrangements (the ‘make or buy’ decision). These two papers by Williamson and Goldberg, although arising in the context of debates over the regulation of public utilities, sparked off the new field of transactions-cost economics.


Writing twenty years after Goldberg, Crocker and Masten surveyed developments in the field of transactions-cost economics and its application to natural monopoly regulation. Crocker and Masten summarise as follows:

Complex or uncertain transactions requiring durable, specialized investments require long-term incomplete, relational contracts. The choice between franchise bidding and regulation thus becomes largely a question of whether court enforcement or regulator administration is the more effective means of governing those agreements.11

According to Crocker and Masten, Demsetz was wrong because he failed to take into account the transactions costs arising from relationship-specific investments:

The choice between market and regulatory governance of public utility transactions has direct parallels to the theory of the firm. In each case, the hazards of renegotiations in the presence of relationship-specific investments at the time of contract renewal make anonymous spot markets unattractive. The shortcomings of franchise bidding – the public analogue to private long-term contracts – are identical to those of its private counterpart. In complex or unpredictable environments, transactions costs are likely to be minimized by removing exchange decisions from the market context and substituting the continuous oversight that regulatory mechanisms entail.12

As Crocker and Masten observe, the perspective on public utility regulation that emerges from the Williamson and Goldberg papers is somewhat at odds with the prevailing view even amongst economists today. Public utility regulation, according to this perspective, is not about determining the correct price, or setting the price that would prevail in a

10 Page 431.
11 Page 12.
12 Page 13.
hypothetical competitive market. Rather, public utility regulation is better viewed as the design of a governance mechanism which provides sufficient certainty for the parties to make their investments, while allowing adaptation to changing market conditions over time:

Given the need to make [relationship-specific] investments, the governance of public utility transactions becomes less a problem of discovering the right price than of identifying the governance structure that economizes on the cost of arriving at and adjusting the terms of trade over the course of the relationship.\(^\text{13}\)

Crocker and Masten emphasise that this perspective sheds some light on contemporary regulatory debates such as the design of appropriate incentive structures in regulatory contracts, the appropriate scope for regulatory discretion versus prescription and on the value of a move to ‘more mechanical and, hence, predictable forms of governance’. Crocker and Masten argue that if it were possible to move to more mechanical forms of regulation (such as regulation based primarily on changes in total factor productivity), that regulation could be replaced by franchise bidding.


As noted above, a 1962 paper by Stigler challenged the view that the establishment of public utility commissions had any impact on prices or profits, and his 1971 paper claimed that regulation was primarily established to protect the interests of the regulated firms. A 1978 paper by Jarrell (also published in the *Journal of Law and Economics*) further argued that the early history of electricity regulation in the US shows a pattern of competition in each municipality, followed by the creation of state regulatory commissions which the utilities could capture to prevent further competitive entry.

These claims were challenged by a seminal paper by George Priest. The early history of the electricity industry is not one of competition followed by state-regulated monopoly – rather, it is one of experimentation with regulation-by-contract followed by regulation-by-commission. Priest argues that prior to the creation of state regulatory commissions there were many years of experimentation with a system of municipal regulation by franchise. According to Priest, this regulation-by-franchise proved problematic – especially when it came to adjusting the terms of the contract over time, and monitoring and enforcing the terms of the contract (such as the quality of service). In other words, the problems with regulation by franchise were exactly the problems you would expect when attempting to design a long-term framework to provide sufficient certainty for investment while allowing flexibility to adapt to market developments over time.

Initially the franchise contracts were flexible, but this allowed the regulated firms to exploit their power to raise prices. Later, attempts were made to tighten up the terms and conditions the firms were allowed to charge. But this proved inflexible and unworkable in the face of on-going developments in the industry. Eventually the stakeholders settled on a system of broad franchise contracts administered by council subcommittees or boards of arbitrators. Priest argues that:

\(^{13}\) Page 30.
the adoption of regulation by commission cannot be claimed to differ qualitatively from the regulation by city council … or by specialized committee that preceded it in many jurisdictions. As a consequence, in contrast to the implication of the 'theories of regulation' debate, the introduction of regulation by state commission may not represent obviously either the imposition of new and more burdensome controls on public utilities or their escape from excessive municipal competition or regulation. … It is not clear that regulation by commission can be distinguished from an advanced stage in the evolution of municipal franchise contractual form.

Overall, the notion that utility regulation should be viewed as a governance mechanism, designed to solve a hold-up problem arising from the need to make relationship-specific investments, has only partially been accepted in the mainstream economics literature. More precisely, the perspective that monopoly service providers must make substantial sunk immobile investments and are therefore subject to hold-up by the regulatory process is widely accepted. This has led to a reasonably substantial literature on the design of regulatory institutions, limits on regulatory discretion and mechanisms to enhance the ability of regulators to commit to not expropriate the investment of the service provider.\(^\text{14}\) Much of this literature has direct implication for utility regulation in developing countries.

However, the need for customers to make their own sunk relationship-specific investments has been mostly overlooked. Lyon and Huang (2002) is one of the very few papers which argue the importance of taking into account the sunk investments by customers of the monopoly service provider. More recently, Biggar (2009) uses this perspective to argue that the conventional rationale for public utility regulation does a poor job of explaining the behaviour of regulators in practice. Instead, Biggar argues that a perspective which emphasises the sunk investments made by customers of the service provider helps to understand the importance of, say, price stability in regulatory practice, and offers promise as a way to better understand regulatory behaviour more broadly.

\(^{14}\) See Levy and Spiller (1994) and section 3.4 in the survey by Armstrong and Sappington (2007).
II. Pricing I: The Neoclassical Approach

The next few sections address key questions in the design of pricing provisions in the long-term regulatory contract – how should we structure prices, how should those prices evolve over time, and what are the implications of pricing for incentives. These questions are fundamental to public utility regulation – it has been said that the objectives of regulatory pricing are the objectives of public utility regulation itself.

We will focus first on questions of how the regulated prices should be structured. As we will see, the economics literature on this question goes back at least to the 1930s – well before the development of the transactions-cost approach to regulation. This literature has been hugely influential in the economics community and can be seen reflected in virtually all textbooks on regulation. It has, however, been somewhat less successful in regulatory practice, as discussed below.

The starting point of this theory is the question: If the regulator had perfect information about cost and demand functions, and if we ignore the need for customers to make sunk investments to extract the most value from the monopoly services, what level and structure of prices should the regulator set? As we will see, the analysis is typically focused on conventional allocative efficiency and the minimisation of deadweight loss.


Much of the economic thinking on regulatory pricing dates back to work by Hotelling in the 1930s (and, if we want to go further back, to the work of Dupuit and Marshall in the previous century). Hotelling argued that economic efficiency (which he refers to as the ‘general welfare’ – equal to the simple sum of producers’ surplus and consumers’ surplus) was best served by setting regulated prices equal to marginal cost. Hotelling recognised that in industries with increasing returns to scale, setting all prices at marginal cost will not normally yield sufficient revenue to cover all costs (including fixed and common costs). To solve this problem he recommended recovering any shortfall in revenue from wider taxation revenue. According to Hotelling, regulators should set the regulated prices and tariffs for railways and other public utilities equal to marginal cost, financing any shortfall with non-distortionary taxes on land, inheritances, and, to an extent, on incomes.

Hotelling recognised that his proposal was radical and a significant departure from the prevailing view at the time that each industry should earn sufficient revenue to cover not only its marginal cost but also the fixed costs of production (as summarised in the saying: ‘each tub must stand on its own bottom’). But, at the same time, Hotelling’s argument is a simple and logical implication of the objective of maximising the sum of producers’ surplus and consumers’ surplus – or, equivalently, the minimisation of deadweight loss – which we would now call allocative efficiency.

Hotelling recognised that short-run marginal cost – and therefore prices – might vary between those times when there is spare capacity and those times when demand exceeds the capacity of the regulated facility. Anticipating the later theory of peak-load pricing, Hotelling noted:

in cases in which the available equipment is actually used to capacity, and it is not feasible … to increase the amount of equipment, something in the nature of a rental charge for the use of facilities should … be levied to discriminate among different users in such a way that those willing to pay the most, and therefore in accordance with the usual assumptions, deriving the most benefit, would be the ones obtaining the limited
facilities which many desire. This rental charge for equipment … should never be so high as to limit travel to fewer persons than can comfortably be accommodated, except for unpredictable fluctuations.

Hotelling argued that the insistence by railroads on covering their fixed costs in the face of declining demand perversely led to increases in real railroad tariffs during the Great Depression – a policy which reinforced and prolonged that downturn.

One of the biggest criticisms of Hotelling’s proposal is that new investment would not be subject to a market test for feasibility – rather, decisions about whether or not to undertake major new investments must be made by what we would call a ‘central planner’. Hotelling is dismissive about the problems that might involve:

Defenders of the current theory that the overhead costs of an industry must be met out of the sale of its products or services hold that this is necessary in order to find out whether the creation of the industry was a wise social policy. Nothing could be more absurd.

Instead, Hotelling recommends that before undertaking a new investment a detailed study of demand and cost functions should be undertaken by a team of economists, statisticians and engineers. With the benefit of hindsight this element of his proposal seems hopelessly optimistic.

Hotelling also highlights the close relationship between the problem of setting regulated rates and the problem of optimal taxation – a link whose implications were not fully worked out until many years later.

Naturally, Hotelling’s proposals were controversial. One of the more important academic criticisms in the following decade came from Ronald Coase (a later Nobel prize-winner). Coase (1946) pointed out that charging two-part or multi-part tariffs may allow the regulator to both set the marginal or variable charges close to marginal cost, while using the fixed charges to raise sufficient revenue to cover fixed costs.

Hotelling’s proposals were never fully taken up in the regulatory community. Kahn’s classic 1970 text emphasises pricing at short-run marginal cost (SRMC) as the theoretically correct outcome, but worries about the possibility that SRMC might fluctuate widely as demand varies while available capacity remains constant. Kahn suggests that constantly changing price schedules would be ‘highly vexatious to buyers’ and offers, as a compromise, pricing at some form of average long-run incremental cost or long-run marginal cost (LRMC). But why, exactly, might customers value price stability? Isn’t this a sign that something is missing from this theory?


Interestingly, perhaps the most significant contribution to the conventional theory of public utility pricing in the latter half of the twentieth century was not a new idea, but the re-exposition of an old one. Baumol and Bradford take as the starting point the framework set out by Hotelling – that the primary objective of the regulator should be the selection of prices/tariffs which maximise social welfare. Like Hotelling, these authors recognise the close link between the selection of optimal prices and the theory of optimal taxation. They depart from Hotelling, however, in recognising that all taxes are inevitably distortionary – taxes essentially force the prices on at least some goods and
services to depart from marginal cost. If we accept this proposition, then setting regulated prices in some industries equal to marginal cost is not necessarily the best outcome – rather, we are in a second best world. In this world it turns out, unsurprisingly, that departing from pricing at marginal cost is second-best-optimal. The extent to which pricing should depart from marginal cost in any one case depends on the shape of the demand curves. Deadweight loss is minimised when the distortion of prices away from marginal cost is smallest on those services which are the most elastic (i.e., the most responsive to changes in prices). In their words:

> the damage to welfare resulting from departures from marginal cost pricing will be minimized if the relative quantities of the various goods sold are kept unchanged from their marginal cost pricing proportions. ... If we propose to (say) contract the demands for A and B each by precisely \( k \) percent, then if A’s demand is much more elastic than B’s, clearly A’s price must be raised by a substantially smaller percentage than that of B.

Baumol and Bradford present the problem generally as a problem of optimal taxation – how should a government distort prices away from marginal cost to raise a given amount of revenue? But they recognise the immediate application of this theory to public utility pricing – specifically, the answer to the question: If marginal cost pricing leaves a regulated firm with insufficient revenue to cover total cost, how should it distort its prices away from marginal cost so as to yield sufficient revenue to cover its total costs? The answer has, of course, come to be known as Ramsey–Boiteux pricing.

Baumol and Bradford go to great lengths to argue that their proposal is not a new one. They point out that railway rate theorists from the late nineteenth century often argued for what they called ‘value of service’ pricing, which involved charging different freight tariffs for different products according to what the market would bear. Bonbright (1961) cites several similar ideas in the history of public utility regulation. However, the primary theoretical contributions were made by Frank Ramsey in 1927 and M. Boiteux in 1956. Baumol and Bradford conclude that they are:

> at a loss to explain how a set of results flowing from the pens of so distinguished a set of authors and appearing and reappearing in the profession’s leading journals, should so long have escaped general notice. This is all the more curious in light of the importance of the subject and the elegance of the theory.

Baumol and Bradford’s paper seems to have had an immediate and significant impact amongst regulatory economists, sparking something of a revolution in thinking, at least among regulatory economists. Suddenly Ramsey–Boiteux pricing was all the rage. Ramsey–Boiteux pricing seemed to be a complete and theoretically elegant solution to virtually all the questions that were being asked about optimal public utility pricing. What more could a regulator want?

With hindsight, the reception of Ramsey pricing in the regulatory community has been cool. Laffont and Tirole (2000) comment that ‘… it is fair to say that participants in the current regulatory debate are on the whole suspicious of Ramsey access pricing’. Part of the reason seems to be the information requirements. Ramsey pricing requires information on demand elasticities (and potentially cross-elasticities) which may be difficult to measure. Ramsey prices also sometimes seem to be morally objectionable or ‘unfair’ – such as when Ramsey pricing would require the price to be raised on a
customer which has just made an investment which increases his demand for the monopoly service.\textsuperscript{15}


One of the applications of the theory of marginal cost pricing is to circumstances where demand changes more rapidly than the capacity of the monopoly facility, and the services of the natural monopoly facility cannot be stored. This applies, of course, to most public utilities, including electricity transmission and distribution services, telecommunications services, postal services, port and rail services, and so on. In every case, demand for services is variable but capacity is limited. At off-peak times the SRMC of supplying an additional unit of output is low or zero, whereas at peak times, when all capacity is utilised, the SRMC of supplying an additional unit of output can be very high indeed.

In effect, when the services of a monopoly cannot be stored, purchases of the monopoly services at different times of the day or year should be treated as different products, with potentially different prices. Since services at peak and off-peak times share substantial common costs (the fixed cost of the capacity of the facility), this pricing problem can be viewed as a problem of allocating common costs.

Under certain simplifying assumptions, the pricing arrangement which maximises the sum of producers’ surplus and consumers’ surplus (ignoring the need for sunk investments) is to set different prices for peak and off-peak time periods, with the prices at off-peak times just equal to short-run marginal cost, and the prices at peak-times high enough to ration demand to be equal to the available capacity, combined with an investment rule that ensures that the available capacity is augmented to the socially-optimal level. Under constant returns to scale, the higher prices at peak times yield sufficient revenue to cover all of the fixed costs of the natural monopoly facility. In effect, all of the common costs of the monopoly facility are loaded onto peak periods. This is known as ‘peak-load pricing’.

The theory of peak-load pricing was primarily developed in the 1960s and 1970s, starting from work by Boiteux (1949, 1951) and Steiner (1957), and extended by Turvey (1968) and Crew and Kleindorfer (1976).\textsuperscript{16} Rather than select one or more of these papers for this list of the fifty most important papers, I have chosen merely to cite the valuable 1995 survey by Crew, Fernando and Kleindorfer.

There are many extensions to this basic peak-load pricing model. Perhaps the simplest extension deals with the case where the demand at peak periods is sufficiently elastic that applying the simple peak load pricing rule would end up shifting the peak demand to the off-peak period. According to Joskow (2007) this problem was realised in practice in a number of countries that instituted simple peak-load pricing models for electricity during the 1960s and 1970s:

\textsuperscript{15} Laffont and Tirole (2000) provide an example.

\textsuperscript{16} We might also mention papers by Panzar (1976), Waverman (1976), and Bailey and White (1974). In fact sometimes it feels like you can’t be a regulatory economist of any standing unless you have published a paper on peak-load pricing!
The off-peak price was set so low that many consumers installed electric heating equipment that on cold winter nights led electricity demand to reach its peak levels. The solution to the problem is to share the joint capacity costs in a way that reflects the relative valuations of peak and off-peak consumptions. ... The off-peak price still is lower than the peak price, but now all consumption pays a share of the network’s capacity costs. Note as well, that the implementation of efficient prices now requires the regulator to have information about the elasticity of demand in different time periods.

There are important variations to the standard peak-load pricing problem which are important to the current work of regulatory authorities in Australia. Crew and Kleindorfer (1976) extended the original work to incorporate the possibility of a range of different production technologies with different fixed and variable costs. This work gives rise to the idea of a ‘merit order’ and notion of an optimal technology mix – that is, an optimal mix of ‘baseload’, ‘intermediate’ and ‘peaking’ capacity. Other extensions include the possibility of stochastic variation in production capacity – that is, the notion of unplanned outages, which, in turn, leads to the notion of optimal levels of reserve capacity and the associated optimal prices.  

The idea that efficient prices should vary by time-of-use, with prices higher at peak periods, seems to be very widely accepted, although the acceptability of such pricing arrangements to regulatory authorities seems to vary across industries. While some form of peak/off-peak differentiation is common in many industries, it is not clear that prices are used to efficiently ration demand in all circumstances, as peak-load pricing would require. For example, mobile phone networks are routinely saturated at certain peak times, with no apparent attempt to ration scarce capacity through prices; toll roads in Melbourne do not differentiate between peak and off-peak times; peak pricing of airport take-off and landing slots has found limited acceptance around the world.

In their survey of the contribution of economists to real-world policies, Baumol and Faulhaber (1988) spend some time discussing the impact of peak-load pricing theory on regulatory decisions. Their summary is that time-of-use pricing practices existed prior to the contribution of economists, and their implementation, in practice, doesn’t follow the theoretical prescriptions:

While economists’ arguments have again made time-of-day rates an issue for discussion in the electric power industry, its modern resurgence has had the most effect in telecommunications, where time-differentiated billing appeared to have been much easier to institute than in electric power. Prices differentiated by time of day and day of week have proliferated in telecommunications since the 1960s; however, the methods used to determine peak and off-peak prices owe little to economists’ formulas and more to accountants’ methods in allocating capacity costs. In other areas of public utility application, such as transportation, peak-load pricing has found little application, possibly because of preoccupation with ‘fairness’. ... Peak-load pricing, hailed as an example of the practical application of new economic theory, appears to be something a bit less. Although the effect of economists’ modern writings on the topic on the behaviour of regulatory commissions and public utilities appears to be important, this work is neither an economists’ invention nor is it new to industry.

17 Further variations on this theme include issues related to the technology required to set time-of-use prices (smart meters) and the theoretically-optimal arrangements in the absence of detailed time-of-use prices and therefore the need for some form of non-price rationing in certain circumstances.
When it comes to the textbook neoclassical theory of public utility pricing, one of the most valuable references is a chapter prepared for the *Handbook of Industrial Organization* by Ronald Braeutigam. Despite being more than twenty years old this chapter usefully summarises the theoretical work on utility pricing from most of the twentieth century.

In addition to surveying the theory of Ramsey pricing and peak-load pricing mentioned above, Braeutigam goes on to discuss what are known as ‘non-linear outlay schedules’. We noted earlier that simple two-part pricing might allow a firm to recover its fixed costs through fixed charges while, at the same time, reducing its marginal charges down to marginal costs. But why stop at simple two-part charges? Why not have a structure of charges whereby the marginal price varies with the quantity consumed? Such tariff structures are common in electricity distribution and retail water supply. Importantly, theory shows that if any marginal price differs from marginal cost it is possible to design a multi-part pricing schedule which leaves all parties better off.

Finally, what if consumers differ in their demand for the monopoly service, but the demand curve of each consumer is not directly observable by the monopolist? In this case the optimal pricing arrangement is a menu of tariffs where each consumer can select the tariff structure that he/she desires. Such arrangements are common in, say, mobile telecommunications where customers are offered a range of options, some with high fixed fees and low variable fees, and others (‘pre-paid services’) with low fixed fees and high variable fees.

In addition to ignoring the need for sunk investments, most of this classic work on pricing assumes perfect information on the part of the regulator, and puts aside the central problem of asymmetric information and the focus on incentives which so dominated regulatory thinking in the 1980s and 1990s (discussed in later chapters). Writing almost four decades after the paper by Baumol and Bradford, Ingo Vogelsang (2006) helpfully puts Ramsey pricing into context in the following reminiscences:

> Until about 1970, many of us believed in truly optimal or first-best regulation, which meant marginal-cost prices as proposed by Hotelling (1938). However, over time, the adjective ‘optimal’ has received more and more qualifications. The first was that losses incurred under optimal prices in the presence of economies of scale led to second best Ramsey pricing associated with Boiteux (1956), a movement that peaked around 1980. The main insight here was that prices should deviate from marginal cost prices by mark-ups that are inversely proportional to demand elasticities … The deficiency of Ramsey prices was the regulator’s lack of information about cost and demand functions. Thus, the next wave was third-best regulation under incomplete information. The main insights from this wave were (a) that regulated firms might need to be able to make economic profits in order to reveal private information and (b) that such profits can be limited by giving firms a choice from a menu of regulatory options. This wave probably peaked with the publication of the Laffont and Tirole (1993) book on incentive regulation. What is the next step away from optimal price regulation? Is it fourth best regulation that makes theoretical models of regulation applicable under political and practicality constraints? In any case, regulation economists have moved further and further away from what was once perceived as optimal price regulation.

A few papers have attempted to assess whether or not the neoclassical economics literature has, in practice, had an impact on regulatory policy, including Baumol and Faulhaber (1988) and Berg and Tschirhart (1995). Berg and Tschirhart (1995) point out
that neoclassical economics has not been as influential in regulatory practice as is usually thought: time-of-use pricing has been around for over a century and was an invention of engineers, rather than economists; Ramsey pricing has been discussed by regulatory commissions but the information required to use it has thwarted implementation; stand-alone cost tests are costly and result in such a high ceiling price that the FCC has rejected the approach; rate-of-return regulation is still the dominant form of regulation for electricity; and embedded cost is still seen by many commissions as the only practical means to determine prices.

What might explain this lack of enthusiasm in policy circles for many of the conventional prescriptions of neoclassical economics? Berg and Tschirhart (1995) argue that the problem is not that economists are not being heard:

If neoclassical economics has been less influential than its practitioners would like, it is not because the lines of communication between the economist and policymaker are few and fragile. Academic economists often serve as commissioners at both the state and federal level. Commissions and utilities frequently sponsor conferences where academics present current theoretical and empirical results. Universities house public utility institutes that provide courses, research, forums, and well-trained graduate students to commissions and industries.\(^{18}\)

Berg and Tschirhart (1995) suggest that this lack of influence is due to the lack of concern amongst economists for issues of equity or fairness. In my view, an alternative explanation for this lack of influence is that neoclassical economists have focused on the wrong foundation for regulation. Regulators do not, in practice, behave as though their central focus is the minimisation of deadweight loss, as the neoclassical approach assumes. Instead, I have suggested that regulators behave as though their central concern is the protection and promotion of sunk investments by both the service provider and its customers. Protecting the sunk investments of customers requires careful attention to issues such as price stability, and the impact of pricing structures on individual consumers – issues which are largely neglected in the neoclassical analysis, but which make sense in the context of protecting sunk investments by customers.

\(^{18}\) Page 322.
III. Pricing II: Cost Allocation

One of the fundamental decisions regulators must make is how to structure prices and how to allow those prices to evolve over time. Since a common regulatory objective is to ensure that the regulated firm earns sufficient revenue to cover its long-term costs, the question of how to structure regulated prices is equivalent to a cost allocation question – How do we allocate joint and common costs across services? The question as to how prices should evolve over time is equivalent to the question of how we should allocate costs which are common over time. This latter question is closely related to the question of the choice of depreciation, which is discussed further in the next section.

In practice, it is common for regulators to establish cost allocation mechanisms which relate changes in costs to changes in prices. These cost allocation mechanisms have been heavily criticised by economists as yielding prices which have no economic significance (See Bonbright 1961, Braeutigam 1980, Baumol, Koehn and Willig 1987). The resulting prices do not necessarily minimise deadweight loss. Nevertheless the use of cost allocation mechanisms persists and is a staple component of regulatory practice. Other than asserting that the resulting prices are meaningless, the economics literature has had little to say on why cost allocation mechanisms exist or how they should be designed.

The previous section addressed the question of how to structure regulated prices in such a way as to minimise the deadweight loss. But the transactions-cost approach to regulation emphasises that there are other, more important, considerations. Customers who must make a sunk investment require some kind of assurance that regulated prices will not increase in the future. In other words, one of the features that we would expect of regulated prices is that they would be broadly stable over time. This is, in fact, consistent with observed regulatory practice.

But there is more that we can say. In the long-run, as the cost of providing a given set of services changes, those changes must eventually be reflected in prices. Furthermore, in order to ensure stable prices, the regulator must ensure that prices for a particular set of services do not increase when the costs for a different set of services increases (perhaps due to an increase in the range of services or the quantity or quality of the other services). This implies that the regulator must ensure that there is no ‘cross-subsidisation’ between service categories. These considerations of price stability, reflecting long-run cost changes, and eliminating cross-subsidisation between services sheds some light on how cost allocation mechanisms should be designed.


In a classic paper from 1975 Faulhaber proposes one particular definition of what might constitute ‘subsidy-free’ prices (or, equivalently, a ‘subsidy-free cost allocation’). He proposes the notion that a set of prices (or a cost allocation) is subsidy free if the revenue from any service (or group of services) exceeds its ‘stand-alone cost’. Faulhaber proves the famous result that a set of prices is subsidy-free if and only if the revenue from any service (or group of services) exceeds the incremental cost.

---

19 Berg and Tschirhart (1995) observe that ‘this area may be one of the more under-analyzed in the field of regulatory economics’.
This result has directly influenced the work of regulatory authorities around the world to this day. But it should be emphasised that this paper does not make any claim that the proposed restriction on the range of possible prices (or the range of possible cost allocations) minimises the deadweight loss. It is well-known that Ramsey-optimal prices can sometimes exceed stand-alone cost (or conversely fall short of marginal cost).

Faulhaber himself does not claim that his approach maximises textbook economic efficiency, but merely proposes his approach as a solution to the concern of public policy makers to the question of equity: He asks: ‘does a proposed price structure … “unduly” favour the consumers of one commodity at the expense of the purchasers of another commodity?’. Faulhaber’s paper spawned its own mini-literature on ‘axiomatic’ approaches to cost-allocation. The approach in this literature is to list a set of criteria that we would like a cost-allocation to satisfy, and then to explore the characteristics of the rules which satisfy those criteria. This body of work was summarised in a 1985 book edited by H. Peyton Young.


Young highlights several possible extensions to the basic Faulhaber model. The most well-known is the so-called ‘Shapley Value’. The Faulhaber approach argues that each customer, or group of customers, should pay a revenue which exceeds the incremental cost of the services that he or she consumes. The Shapley Value takes this further and argues that the revenue paid by a customer should be equal to the average of the incremental cost of adding that customer to every possible coalition of other customers. The Shapley Value is the unique cost allocation methodology which satisfies a simple set of unobjectionable axioms.

Faulhaber, Young, and the other papers in this literature do not embed the cost allocation problem within the task of designing an optimal long-term regulatory contract. They merely suggest a set of axioms which they suggest prices should satisfy, without explaining why those axioms are necessary. This literature appears to have died out. The recent comprehensive survey paper by Armstrong and Sappington (2007) makes no reference at all to this literature. There is some evidence of the Shapley Value concept being used in practice, particularly in airport pricing. I suggest that this literature, reinterpreted in the context of the design of a regulatory contract, has the potential to make a larger impact in the future.
IV. Pricing III: Rate Base and Depreciation

We noted earlier that the pricing problem has two elements – the optimal design of price structures and the design of the way those prices evolve over time. As noted in the previous section, both these problems can be viewed as a form of cost allocation. The recovery of the sunk costs of investment over time is one of the most central tasks for regulation in practice. Much of the regulatory history of the first half of the 20th century involves disputes over the precise mechanism for ensuring that the regulated firm covers the costs of its sunk investments. In Australia, this cost allocation is usually achieved through a form of the building block model – that is, through the maintenance of a regulatory asset base and through the choice of depreciation.

For most of the 20th century the standard approach to rate-making simply involved setting a level of prices that would yield the utility a cash-flow (after deducting operating expenditure and tax) equal to a ‘fair’ rate of return on the regulatory ‘rate base’ plus depreciation. But there was no consensus on how that rate base should be set. There were at least three possible candidates: (a) the historic cost of the assets of the firm; (b) the replacement cost of the assets of the firm (which may be above or below the historic cost); and (c) the historic cost adjusted for inflation.

(13) ‘Rate Base Selection and the Structure of Regulation’, Bruce Greenwald, Rand Journal of Economics, 15(1), Spring 1984, 85-95

Greenwald’s paper focuses on the evolution of the level of the asset base (or ‘rate base’) over time. There is, of course, a close relationship between the evolution of the asset base and depreciation – normally the change in the asset base is equal to any new capital expenditure less the depreciation.20 A regulator can either choose a path for depreciation or a path for the regulatory asset base (RAB) but not both at the same time. Greenwald focuses on the choice of a path for the regulatory asset base – in effect, depreciation in his model is the residual that falls out from the change in the asset base from period to period.

Greenwald shows that provided (a) depreciation is set equal to the change in the asset base each period, and (b) the regulator sets the allowed return equal to the firm’s cost of capital, then any methodology for revaluing the asset base each regulatory period is consistent with investors earning a fair and normal return. In particular, any of the three major approaches to setting the RAB mentioned above are consistent with the regulated firm earning normal returns over its life. If we want some way to choose between these approaches for setting the RAB we must look elsewhere. This result gives enormous scope for choosing a path of the RAB over time. Greenwald also argues that:

(a) the regulatory asset base cannot be negative (otherwise investors could simply ‘shut-up shop’);

---

20 Normally the asset base in the following period is set equal to the asset base in the previous period plus any new investment less depreciation. This close relationship is emphasised by Rogerson (1992) amongst others. I will follow the normal terminology and use the word ‘depreciation’, although ‘amortisation’ – that is, the ‘spreading of costs over time’ – is a technically more accurate phrase. Depreciation suggests the physical wearing out of capital which is simply not relevant here.
(b) the regulatory asset base must at all times be less than the market value of a profit-maximising unconstrained monopolist (no firm can earn more than it could if it were unregulated); and

(c) the regulatory asset base should start at zero before any investment is sunk and should finish at zero when the firm ceases to exist.

Finally, Greenwald makes the important argument that the concept of the regulatory asset base helps the regulator resolve the problem of time-consistency – it allows the regulator to commit to ensuring that investors will be adequately compensated in the long-run, no matter what methodology the regulator chooses. The regulator may change its mind as to how it regulates – that is, what procedures or mechanisms it uses, but as long as the regulatory regime ensures that the change in the asset base over the regulatory period is equal to the present value of the payments to investors (discounted at the appropriate cost of capital), investors will be adequately compensated for their investments in the long run.

This latter observation provides an important justification for the use of the building block model in the long-term regulatory contract: the use of the building block model provides an important assurance to the regulated firm that, no matter how interest rates, costs, and demand change in the future, the regulated firm will be able to earn a normal return on its investments.


Because of the close mathematical relationship between the evolution of the RAB over time and the level of depreciation, it is not surprising that if the level of the RAB is essentially arbitrary, then the choice of the path of depreciation will also be arbitrary. This is the main point of a subsequent paper by Richard Schmalensee.

Schmalensee’s concise note proves that, provided the regulator sets the allowed return equal to the firm’s true cost of capital, virtually any path of depreciation is consistent with investors receiving a normal return on capital. When it comes to ensuring that investors are adequately compensated, the choice of the path of depreciation is essentially arbitrary. This remains true even if the firm doesn’t know what level of depreciation the regulator will choose at any point in time in the future.

Of course, Schmalensee recognises that this does not imply that all depreciation schedules are equally socially desirable. The choice of depreciation schedule affects how regulated prices evolve over time. As Baumol said in another seminal paper almost twenty years earlier (which almost made it into our list of the top fifty):

there may be many alternative revenue streams, each of which can give investors their required returns. The choice of depreciation policy may then be defined as the selection of one of these intertemporal patterns of prices which will yield one of the revenue streams adequate to compensate investors … It should now be clear in what sense one depreciation policy may be better than another. One set of intertemporal patterns of product prices may yield a better allocation of resources … than another.²¹

²¹ Baumol (1971).
Some authors have suggested that the optimal choice of depreciation is the choice which results in 'inter-temporal Ramsey prices'. This approach is suggested, for example, by Rogerson (1992). Burness and Patrick (1992) show that if demand and costs are constant, this approach leads to prices that are constant over time. In my view, however, as I have emphasised above, this approach neglects the importance of sunk investment by customers. When we take into account the need to protect sunk investment by customers we find that it may be desirable to achieve long-term price stability and to avoid undue discrimination as we noted earlier.


A further piece of the jigsaw on depreciation/amortisation was suggested by Crew and Kleindorfer. This paper focused on the possibility of an external constraint on the ability of the firm to recover its costs in the future.

Greenwald noted that the regulatory asset base could not increase above the present value of the future revenue stream for an unregulated monopolist. In the Crew and Kleindorfer paper, the present value of the future revenue stream for the unregulated monopolist is declining exponentially over time, perhaps due to forces of competition or technological change. This places a declining upper limit on the path of the regulatory asset base over time. The result, unsurprisingly, is that front-loading of capital recovery is essential if the regulated firm is to remain viable.

In essence, when the regulated firm will be constrained by other forces in how much it can recover in the future, the regulator must take this into account in the present, and allow the firm a higher rate of depreciation. This is the origin of the tilted annuity concept used by some regulatory authorities in telecommunications regulation. Crew and Kleindorfer point out that traditionally there has always been a sense among regulators and utilities that problems could be put right "at the next rate case". However, they emphasise that this is clearly not always true. If some other constraint – such as changes in demand or technology – prevents the regulated firm from earning a normal return in the future, the regulator must take that into account in its depreciation policy today.


Michael Salinger's starting point is the observation that the US Telecommunications Act of 1996 mandates cost-based prices. The US FCC interpreted this to mean the 'forward-looking economic cost'. Salinger starts his paper highlighting that, in the presence of substantial sunk costs (and therefore high levels of depreciation), the notion of 'cost-based prices' and 'forward-looking costs' does not determine a unique set of prices.

Salinger points out that if 'cost-based prices' means no economic profit, there is an infinite number of streams of future prices which are consistent with zero economic profit. To pin down the meaning of 'cost-based prices' we must impose some additional

---

22 Burness and Patrick (1992) also extend the Schmalensee result to consider the case where the regulator might give a cost of capital which is 'too high' or 'too low'. Unsurprisingly, a regulatory cost of capital that is 'too high' leads the firm to delay depreciation as long as possible. A regulatory cost of capital that is 'too low' causes the firm to seek to depreciate as rapidly as possible.
constraints on the path of desired prices. The textbook economic approach focuses on choosing the path of prices to minimise the deadweight loss. Once those prices have been chosen, Salinger argues that we can and should work backwards to work out the depreciation/cost allocation which justifies those prices. Echoing the argument of Baumol cited above, he emphasises in the introduction:

> At the outset … it is worth being clear on the order in which costs and cost-based prices are calculated. The prices are calculated first and depreciation, one of the components of cost, is calculated second. As a result, the term ‘price-based costs’ describes the calculations more accurately than does ‘cost-based prices’.

Salinger’s argument makes sense within the neoclassical perspective, with its focus on deadweight loss. However, under the transactions-cost perspective on regulation, Salinger’s argument can be called into question. Despite being heavily criticised by economists, notions of cost-based prices and cost allocation are deeply rooted in regulatory practice. As noted earlier, the transactions cost approach to public utility regulation suggests that these practices may have a legitimate rationale — as part of a long-term contract which seeks to protect and promote sunk investment. In other words, the notion of cost-based prices, despite being ambiguous, may in fact make more sense than Salinger’s proposed alternative of price-based costs (that is, working out the Ramsey prices and working backwards to determine the corresponding cost allocation).

Salinger also makes a number of secondary points which are interesting in themselves. He points out that where there is a constant probability of an asset being stranded or failing in some way, the required revenue stream is higher than if the life of the asset were known with certainty. Salinger also comments on what was (at the time) a brand new idea — the real options theory. Although concerns had been raised that regulation might cause firms to delay investment Salinger argues that, in principle, cost-based regulation eliminates any incentive to delay. He also notes the argument that investments ‘can create options as well as destroy them’.

None of the papers surveyed in this section address the vexed question of how to set the ‘opening RAB’ at the time of transition into a new regulatory regime. There have only been a couple of papers in the economics literature on this topic. 23

---

V. Averch–Johnson and Incentives under Rate-of-Return Regulation

As noted above, regulatory contracts will (almost always) involve some form of mechanism for periodically adjusting prices to reflect market developments such as changes in cost or demand. This might be through a regulatory reset or ‘rate hearing’ process. It has long been recognised that the mechanisms established for adjusting prices have important secondary effects on incentives – particularly incentives for productive efficiency. A cost-plus government procurement contract has quite different incentives for cost-cutting than does a simple fixed-price contract.

These incentive issues have been a central focus of the economic literature of the last forty years and are the subject of several parts of this survey. We first review here some of the key papers from the 1960s and 1970s which are historically important in the development of ideas which we now take for granted. These developments ultimately provided important ideas in the design of regulatory price adjustment processes.


Averch and Johnson’s famous paper from the early 1960s asks the simple question: If we assume that regulation acts to instantaneously adjust prices so as to maintain a constant target return on a firm’s capital stock, what incentives does the firm have to choose an efficient combination of inputs? Averch and Johnson show, using fairly straightforward maths, that if the regulator sets the regulatory rate of return above the firm’s true cost of capital, the regulated firm has an incentive to choose too much capital relative to labour – that is, there will be an inefficient capital–labour ratio. This observation sparked off a large empirical and theoretical literature exploring what came to be known as the ‘Averch–Johnson’ or ‘A–J’ effect.

The Averch–Johnson model assumed an extreme form of what is known as ‘rate of return’ or ‘cost of service’ regulation in which prices are continuously and rapidly adjusted so as to yield the desired return on capital stock. In reality, rate of return regulation as it was historically practiced in the US always involved an element of ‘regulatory lag’ – that is, a period of time before prices were adjusted to reflect changes in costs. This regulatory lag gives rise to some desirable incentives, as discussed further below. Even in 1962 it was recognised that allowing higher ex post rates of return was necessary to induce desirable incentives. Averch and Johnson note:

We have been told by representatives in both the industry and in regulatory agencies that justification exists for allowing a return in excess of cost to give firms an incentive to develop and adopt cost-saving techniques. If the firm is left only indifferent as among a wide range of activities it has no positive incentive to minimize costs for any given activity.

Averch and Johnson’s paper was, for a while at least, very widely cited, often incorrectly. The A–J effect later came to be synonymous with ‘gold-plating’ and with general inefficiency of the regulated firm. But these ideas cannot be strictly attributed to Averch and Johnson. Averch and Johnson only highlighted a potential tendency towards an inefficient mix of capital and labour – but they made no mention of the tendency to simply spend too much on all inputs – so-called ‘X-inefficiency’ as discussed further below.
Averch and Johnson’s paper kicked off a large literature exploring the incentives of regulated firms operating under a ‘rate of return’ constraint. Some papers explored how the conclusions of the model changed if it was assumed that regulated firms did not maximise profits, but instead maximised other measures – such as sales. Other papers asked the question: If the regulated firm responds in the way that A–J suggest, how should we set the regulatory allowed return so as to induce efficient outcomes?

A 1970 paper by Baumol and Klevorick surveys this literature. In the process they tidy-up and clarify some loose ends in Averch and Johnson’s work. Most of this part of their paper is merely of academic interest. However, Baumol and Klevorick make some very interesting comments on the role of regulatory lag – essentially laying the foundations for incentive regulation. Baumol and Klevorick suggest that the incentive for cost-reducing effort on the part of regulated firms is akin to investment in cost-reducing ‘knowledge’. The firm’s incentive to invest in this knowledge is critically linked to the length of the period until the next regulatory review – the notion of regulatory lag:

Suppose the equilibrium stock of knowledge increases with the length of the regulatory lag, because the firm undertakes more research the longer are the periods during which it enjoys the temporary benefit of an excess earning flow. Then the total discounted present value of the benefits generated by the regulated firm will rise when the length of the regulatory lag increases. For example, with instantaneous regulation – a lag of length zero – and technical progress which increased the productivity of capital and labour proportionately, one would expect the firm to undertake no research outlays since it could never reap any benefits from such an expenditure, while with no regulatory review – a lag of infinite length – one would expect the firm to accumulate a maximal stock of knowledge since it reaps the benefits of its research work forever.

Importantly, Baumol and Klevorick suggest the existence of a social-welfare trade-off between the incentives to engage in cost-reducing effort, and the share of the benefits consumers receive from those efforts. The shorter the regulatory period, the quicker any cost-reducing benefits are passed on to consumers; the longer the regulatory period, the greater the incentives to create those cost-reductions in the first place.

Regulators can then choose a length for the period between reviews which will attain the social optimum with respect to the amount and distribution of producer’s plus consumers’ surplus. If regulatory agencies have chosen the time period between reviews in some more or less ad hoc fashion, this model would suggest that they have neglected a potentially valuable instrument of control.

In their conclusion, Baumol and Klevorick suggest that the A–J over-capitalisation may be less significant in practice than what was then known as ‘X-inefficiency’ – the possibility for the regulated firm to be using all inputs inefficiently, rather than substituting capital for labour on the efficient frontier. With the benefit of hindsight the last paragraph of the paper seems prescient:

---

24 Klevorick (1971).
A rate-of-return ceiling by its very nature tends to contribute another X-inefficient influence. By ruling out all profits in excess of some predetermined rate as potential fruits of monopoly, a rate-of-return constraint also precludes extraordinary rewards for extraordinary entrepreneurial accomplishment. Indeed, if such a constraint were applied with complete inflexibility it might well eliminate altogether any financial reward for efficiency and innovation. If the firm cannot hope to earn more than the ceiling, even temporarily, and if the demands for its products are inelastic so that it can reasonably count on that profit rate in any event, what can management hope to gain by working for increased efficiency in its operations? Fortunately, as we have indicated, regulatory lag, like the Schumpeterian innovation process, by permitting temporary rewards for efficiency can perhaps provide just the necessary incentives. … While regulation may well be … the source of some non-negligible inefficiencies in the economy, it is not clear that the phenomenon encompassed by A–J analysis is the most disquieting of these.


Just a few years later a young newcomer on the regulatory-economics scene, Paul Joskow, launched an even more direct attack on the Averch–Johnson model. In the ten years since the publication of the original A–J paper, the A–J model had become the ‘accepted paradigm for the analysis of the effects of rate of return’ regulation of public utilities in the United States’ and ‘an accepted descriptive model which implicitly assumes that actual regulatory processes are characterized in this way’. Joskow’s core critique is that:

A–J models bear little correspondence with actual regulatory instruments. As a result, these models may be of little use in evaluating the effects of actual regulatory instruments on the behaviour of regulated firms.

The A–J model explicitly assumes that regulatory agencies continuously adjust the regulated firms’ prices so as to bring the firm’s cost of capital down to the regulatory-allowed cost of capital at each moment in time. Joskow had already spent many years studying the actual behaviour of regulatory agencies. He makes the following telling observation:

Contrary to the popular view, it does not appear that regulatory agencies have been concerned with regulating rates of return per se. The primary concern of regulatory commissions has been to keep nominal prices from increasing. Firms which can increase their earned rates of return without raising prices or by lowering prices (depending on changing cost and demand characteristics) have been permitted to earn virtually any rate of return that they can. … There is no ‘allowed’ rate of return that regulatory commissions are continually monitoring and at some specified point enforcing.

Joskow proposes an alternative model of the behaviour of regulator agencies. In this model, both firms and regulators were content to avoid formal legal involvement in lengthy regulatory proceedings. As a result: ‘As long as prices were not going up, regulatory commissions were happy to “live and let live”, engaging in little or no formal rate of return regulation’. During long stretches of the 1960s there were, in fact, very few regulatory rate hearings – rates stayed essentially unchanged in nominal terms for quite long periods of time.

Joskow’s model, of course, had important implications for the incentives of the regulated firm. In fact, as long as costs overall were broadly falling, the firm had strong incentives to produce efficiently, since the higher profits could be kept until at least the next formal
review, which could be arbitrarily far in the future. Joskow also points out that the A–J model explicitly assumes that the regulator allows all capital into the regulatory asset base. ‘But regulatory agencies need not be so stupid. … Actual inefficiencies that arise from the formal review process depend crucially on how careful the commission is in reviewing the “allowability” of the firm’s expenditures and other elements of the formal hearing process itself’.

In Joskow’s model a change in the environment, such as an increase in inflation, temporarily provokes an increasing in the number of formal regulatory hearings. But Joskow argues that this increase in rate hearings is temporary – in the medium to long term new regulatory adjustment mechanisms will be developed which allow for automatic adjustments, ‘unbinding’ firms from any kind of formal rate of return constraint. Interestingly, with this comment Joskow seems to anticipate the revolution of price-cap regulation that emerged in the 1980s which is discussed further in the next section of this paper.
VI. The Rise of ‘Incentive Regulation’

As just noted, in direct contrast to Averch and Johnson, Joskow (1974) argued that during the 1960s – a period of low and stable inflation – regulatory rate reviews were rare and regulated firms had substantial incentives to improve their productivity. However, this beneficial arrangement of low inflation, long periods between rate reviews, and strong incentives for productive efficiency, broke down in the 1970s. In the face of high inflation, regulated firms sought regulatory hearings more and more frequently – moving the regulatory framework closer to a ‘cost-plus’ approach. ‘Fuel adjustment clauses’ were introduced to allow regulated utilities to pass through higher fuel costs but these did not address the problem of higher costs of other inputs (such as labour) and did not provide any incentives for lowering fuel costs (by say, improving fuel purchasing arrangements).


In the early 1980s Baumol made a proposal that sparked off something of a revolution in regulatory practice. In order to extend the period between rate reviews Baumol proposed allowing a regulated firm to automatically increase regulated prices at a rate reflecting the average rise in input costs less a ‘productivity offset’ reflecting the expected increase in the firm’s productivity. Baumol notes:

[If the regulatory agency were to institute an automatic rate adjustment mechanism for, say, a three-year period, increasing rates by enough to make up for an appropriate portion of the input inflation effect, it will then be left up to the firm to make up the rest through its own effort. If the firm’s future productivity record turns out not to live up to the standards that can be expected from its past performance, then the company will suffer the financial consequences. But if by vigorous innovative effort the firm outdoes what was to be expected of it, then both the company and the public will profit.

Baumol does not, in this paper, specifically advocate the use of an index of the regulated firm’s own input prices, or measures of the firm’s own productivity. The reason is that:

Measurement of productivity is not a matter which is altogether straightforward. There are many alternative formulae for the measurement of productivity and the selection and use of such a formula may well embroil the regulatory agency in precisely the sort of protracted hearings and administrative complications whose elimination is one of our prime objectives.

Moreover, an index of the firm’s own input prices would have to be constructed de novo and would probably be subject to manipulation by the firm itself. Instead, Baumol recommends reliance on a broad price index – such as the rate of inflation, and the use of a productivity growth figure ‘based on national data or data for other industries … thereby taking into account productivity-affecting developments outside of the regulated firm’s control’.

Baumol emphasises that this arrangement would not be permanent: ‘No commission should or can be expected, after adoption of such a rule, to leave the task of rate adjustment entirely to the formula forever thereafter’. Instead, the firm would be subject ‘to a formal review process from time to time’.

Seldom in the history of regulation, I would suggest, has an idea had such an immediate and significant effect. Within a year of the publication of this paper, Stephen Littlechild (1983) proposed the use of an ‘RPI-X’ price-cap approach for the regulation of British...
Telecom. A few years later the US FCC (1987) recommended the use of a ‘price cap’ approach in the regulation of AT&T, and subsequently for the local exchange companies. Although in 1985 virtually every state regulator in the US used rate-of-return regulation for regulating the telecommunications industry, by the end of the same decade half of all the US state regulators used price-cap regulation, and by the year 2000 virtually every state in the US used some form of price-cap regulation. It is hard to imagine a more dramatic impact on regulatory practice.

What came to be called ‘incentive regulation’ revolutionised the standard regulatory approach to telecommunications and many other regulated sectors. However, as happens in many revolutions, it was accompanied by a degree of fervour, hope, and ideology that probably exceeded what was known to be true at the time. A good part of the academic literature of the next few years was aimed at tempering and moderating the initial claims and hopes on the part of price-cap devotees.


It seems that the academic community was largely taken by surprise by these developments in regulatory policy. Academic economists subsequently rushed to analyse the pros-and-cons of this new CPI-X regulatory approach. The first of these papers appeared in academic journals at the end of the 1980s.

In one of the first of these papers, Schmalensee compared a stylised or extreme form of ‘rate of return’ regulation, with a stylised or extreme form of ‘price cap’ regulation. Rate-of-return regulation was characterised as a regime in which the regulator rapidly and fully adjusts prices to the firm’s own average costs. Pure price-cap regulation, on the other hand, was characterised as a regime in which the regulator breaks the link between the firm’s prices/revenues and the firm’s costs entirely.

The basic trade-offs between these two approaches were already well-known by the end of the 1980s: ‘Regimes with cost sharing provide weaker incentives for cost reduction, but they permit price to respond to cost changes and reduce the ability of the regulated firm to profit from regulatory ignorance or favourable cost shocks’. On the other hand:

Price caps provide superior incentives for cost reduction, but the more uncertain the environment, the higher the cap must be set in order to keep the regulated firm profitable, and the greater the average ex post price-cost gap. Cost-plus regulation is thus preferred, at high levels of uncertainty.

Schmalensee concludes:

Generally, this study suggests that price caps have been oversold relative to simple alternatives, particularly if regulators are (or should be) more concerned with consumers’ surplus than with the profits of regulated firms. Regimes in which prices depend in part on actual costs may provide weaker incentives for productive efficiency, but nonetheless generally perform better in the presence of cost uncertainty and asymmetric information about the capabilities of the regulated firms. Regimes involving cost sharing are better than price caps at limiting the profitability of regulated firms and they allow prices to track costs more closely.

These results have been corroborated in subsequent studies. For example, Lyon (1996) compares ‘profit sharing’ or so-called ‘sliding scale’ regulatory regimes with both price caps and rate-of-return regulation. He finds: ‘some degree of profit-and loss-sharing
outside a deadband improves social welfare relative to pure price-cap regulation. Simulation results show that a significant departure from pure price-caps – that is, sharing a substantial portion of profits with ratepayers – may be welfare enhancing’. In fact, in practice, many of the new regulatory arrangements put in place in the US in the 1990s incorporated some form of ‘sliding scale’.


By 1992 it was well understood that so-called ‘price cap’ regulation, by fixing the period between regulatory hearings, strengthened the incentives on regulated firms to minimise their expenditure. However, the significance of the regulator’s actions at the time of the regulatory reset was not fully appreciated. Was price-cap regulation nothing more than rate-of-return with an institutionalised regulatory lag?

In a short paper which appeared in 1992, Ellen Pint compared two possible approaches to resetting the firm’s revenues/prices at the next regulatory hearing. The first approach – based on the longstanding practice associated with rate-of-return regulation – bases the firm’s future revenues/prices on the firm’s out-turn costs in a single year – known as the ‘test year’. The second approach bases the firm’s future revenues/prices on the basis of the average of the firm’s out-turn costs in each year of the previous regulatory period.

Pint shows that when the firm’s future revenues/prices are based exclusively on the firm’s cost out-turns in a single year, the firm has strong incentives to distort the costs (and the capital–labour ratio) in that year. She shows that the regular regulatory resets give rise to a pattern of cost-reducing effort which cycles with the regulatory period, with strong incentives for cost-reducing effort near the beginning of the regulatory period, and weaker incentives (and, indeed, incentives to raise costs) at the end (in the ‘test year’).

In contrast, basing future revenues/prices on the average of the firm’s costs over the previous regulatory period reduces the size of these cycles – giving rise to incentives which are smoother over time.


By 1994, a synthesis was emerging between the new theoretical developments in regulatory economics that arose in the 1980s (discussed above) and the developments in regulatory policy, reflected in the rapid growth of price-cap regulation and its variants. At that time it was recognised that the incentives created by a regulatory regime depend primarily on three factors: (a) the length of the regulatory period; (b) the mechanism for adjusting prices during the regulatory period (by ensuring that regulated prices move broadly in line with exogenous movements in cost, the price-adjustment mechanism allows for longer regulatory periods without risk that the regime will prove unsustainable); and (c) the mechanism for resetting prices for the next regulatory period – and, in particular, the information the regulator takes into account when setting the new prices.

As we saw above, one of the primary claimed benefits for price-cap regulation was that it would increase the power of the incentive on the regulated firm to reduce costs. But, as already noted above, higher-powered incentives are not always better than low-powered incentives. Higher-powered incentives place greater risk on the regulated firm and
increase the likelihood that the regulatory regime will be found to be unsustainable (that is, will be renegotiated, with that renegotiation initiated either by the government or by the regulated firm). Higher-powered incentives on some desirable objectives may also distort the efforts of the firm to concentrate on those objectives (such as reducing costs) at the expense of other desirable objectives (such as improving service quality). Finally, higher-powered incentives also give rise to greater incentives on the regulated firm to behave strategically at regulatory reviews to induce the regulator to loosen the performance targets (e.g., set higher cost targets).

In a survey paper which appeared in 1994, Sappington and Weisman emphasised taking a holistic view to the design of a regulatory regime – rather than focusing on just one amongst many desirable incentives:

The key to success in designing effective incentive regulation is relatively simple: Anticipate all of the incentives that will ultimately come to bear, and structure regulatory policy in advance to limit any adverse incentives.

Sappington and Weisman also emphasised that higher-powered incentives would test the ability of the regulator to commit to the regulatory regime, especially if the regulated firm is able to reduce its costs and earn substantial apparent excess returns. They also mention the problem of the ‘ratchet effect’: if past performance is used to set future performance targets, incentives to improve performance will be weakened:

If future standards are raised each time superior performance is observed, the incentives for superior performance are reduced. In essence, ratcheting up performance standards in response to superior performance under price-cap regulation is akin to tying earnings to production costs under cost-based regulation. In both cases, no good deed goes unpunished.

In addition, Sappington and Weisman mention two further ideas. The first was the possibility of giving the regulated firm a ‘menu of options’ from which it can select. This allows firms with higher scope for performance improvement to choose stronger incentives, but also lower average prices, while firms with lower scope for performance improvement can choose weaker incentives and higher prices.

The second idea was to base the adjustments in regulated prices on factors outside the control of the firm. We have already seen the notion of basing the allowed changes in prices on changes in a broad index of input prices. But what if the firm is subject to some industry-specific change in demand or technology? Instead, why not base the allowed prices of a regulated firm on the costs of other similar firms. This is the notion of ‘yardstick competition’:

The performance of other firms that operate in similar environments can sometimes serve as a benchmark against which to assess the regulated firm’s performance. … When the regulated firm in question is shown to perform better than other firms in comparable settings, evidence of greater diligence or ingenuity on the part of the regulated firm is provided. Such evidence can help to justify enhanced compensation for the firm. Of course, it is critical that the comparison group of firms be carefully selected. Observed differences in performance must be due to differences in diligence or ingenuity, not to exogenous environmental differences, if they are to motivate the regulated firm and enhance perceptions of fairness.

Yardstick regulation is discussed further in a subsequent section of this paper.

In the last few years there have emerged a number of excellent survey articles reviewing the theoretical and practical issues associated with incentive regulation. One of the best of these is a survey by Paul Joskow that has not (to my knowledge) been published.

Rather than attempt to summarise this comprehensive paper, I will just extract a few notable paragraphs. For example, Joskow emphasises the importance of good monitoring and auditing frameworks:

If regulators are going to apply incentive regulation mechanisms that are cost contingent they must have some consistent mechanism for measuring the regulated firm’s actual costs. … Capital cost accounting issues have largely been ignored in the theoretical literature on incentive regulation. Although it has been of limited concern to contemporary economists, any well functioning regulatory system needs to adopt good cost accounting rules, reporting requirements for costs, output, prices, and other dimensions of firm performance, and enforce auditing and monitoring protocols to ensure that the regulated firm applies the auditing rules and adheres to its reporting obligations. Much of the development of US regulation during the first half of the 20th century focused on the development of these foundation components required for any good regulatory system that involves cost contingent regulatory mechanisms. … If incentives are to be extended to the quality dimension as well, as they should be, then these quality dimensions must be defined and associated performance indicia measured by the firm, reported to the regulator, and must be subject to auditing protocols.

Joskow carefully explains the regulatory framework established by Ofgem in the UK. Many of the issues are familiar to regulatory authorities in Australia. He notes that the five-year regulatory period ‘potentially leads to different incentives for cost reduction depending on how close the firm is to the next price review’. For capital expenditures this is addressed through ‘a formula for rolling adjustments in the value of capital assets used for regulatory purposes so that outperformance or underperformance incentives and penalties are reflected in prices for a five-year period’. Joskow also highlights the potential problems that arise from applying different incentive regimes to capex vis-à-vis opex:

The power of the incentive schemes for operating costs and capital costs appears to be different for at least two reasons. First, the sliding scale mechanism applies to capital cost variations but not operating cost variations. In addition, there is not a well-defined line between what is an operating cost that is expensed in a single year and what costs can be capitalised. The firms may have an incentive to capitalize operating costs to beat the opex incentives during the current review period in the hope that they will be included in the RAB during the next review period.

Another issue which is familiar in Australia is the potential for deferral of expenditure across regulatory periods:

A third set of issues relates to incentives to reduce both operating and capital costs today to increase profits during the current price control period, but with the result that service quality deteriorates either during the current review period or in subsequent periods. Deferred maintenance (e.g., tree trimming) and deferred capital expenditures may lead to the deterioration of service quality in either the short run or the long run or both. Regulated firms may hope that they can use adverse service quality trends to argue for higher allowances for operating and capital costs in future price reviews. The UK regulatory process tries to deal with the relationships between operating and capital cost
expenditure and service quality in two ways. First, there are service quality performance norms and incentives … Second, Ofgem reserves the right to ‘claw back’ capital cost savings if they are clearly not the result of efficiencies but rather reflect efforts to cut services in the short-run or the long run. This is not an ideal approach since operating expenditures, capital expenditures and service quality are related in complex ways over time and space. Indeed, it sounds like ‘prudence reviews’ that are a component of traditional cost of service regulation in the US.

Joskow’s comments on the UK approach are telling and could equally apply to Australia:

The basic price-cap regulatory mechanism used to regulate electricity, gas and water distribution and transmission companies in the UK is often contrasted with characterizations of cost-of-service or ‘cost plus’ regulation that developed in the US during the 20th century. However, I believe that there is less difference than may first meet the eye. The UK’s implementation of a price-cap based regulatory framework is best characterized as a combination of cost-of-service regulation, the application of a high powered incentive scheme for operating costs for a fixed period of time, followed by a cost-contingent price ratchet to establish a new starting value for prices. The inter-review period is similar to ‘regulatory lag’ in the US context except it is structured around a specific RPI-\(x\) formula that employs forward looking productivity assessments, allows for automatic adjustments for inflation and has a fixed duration.

A considerable amount of regulatory judgment is still required by Ofgem. The regulator must agree to an appropriate level of the starting value for ‘allowable’ O&M as well as a reasonable target for improvements in O&M productivity during the inter-review period. The regulator must also review and approve investment plans ex ante and make judgments about their reasonableness ex post … It does this without statistical benchmarking studies which are unavailable. An allowed rate of return must be determined as well as compatible valuations of the rate base (capital stock) and depreciation rates. Cost accounting and cost reporting protocols are required to implement sound incentive regulation mechanisms.

Thus, there are many similarities here with the way cost-of-service regulation works in practice in the US. Indeed, perhaps the greatest difference is philosophical. Ofgem takes a view which recognizes that by providing performance-based incentives for regulated utilities to reduce costs, it can yield consumer benefits in the long run by making it profitable for the firm to make efficiency improvements. If the firm over-performs against the target, consumers eventually benefit at the next price review. It has generally (though not always) been willing to allow the regulated firms to earn significantly higher returns than their cost of capital when these returns are achieved from cost savings beyond the benchmark, knowing that the next ‘ratchet’ will convey these benefits to consumers. Under traditional US regulation, the provision of incentives through regulatory lag is more a consequence of the impracticality of frequent price reviews and changing economic conditions than by design.

Finally, Joskow’s comments on price-caps versus rate-of-return regulation are cogent:

There is a lot of loose and misleading talk about the application of price caps in practice. From a theoretical perspective the infatuation with price caps as incentive devices is surprising since price caps are almost never the optimal solution to the trade-off between efficiency and rent extraction when the regulator must respect the regulated firm’s budget-balance constraint (Schmalensee 1989) and raise service quality issues. However, price caps in practice are not like ‘forever’ price caps in theory. There are ratchets every few years which reduce the power of the incentive scheme and make it easier to deal with excessive or inadequate rents left to the firm. They are not so simple to implement because defining the relevant capital and operating costs and associated benchmarks is
challenging. Price caps are also typically (eventually) accompanied by other incentive mechanisms to respond to concerns about service quality. Evaluating the performance of price cap mechanisms without taking account of the entire portfolio of incentive mechanisms in place can lead to misleading results. Effective implementation of a good price cap mechanism with periodic ratchets requires many of the same types of accounting, auditing, capital service, and cost of capital measurement protocols as does cost of service regulation. Capital cost accounting and investment issues have received embarrassingly little attention in both the theoretical literature and applied work on price caps and related incentive mechanisms, especially the work related to benchmarking applied to the construction of price cap mechanisms. Proceeding with price caps without this regulatory information infrastructure and an understanding of benchmarking and the treatment of capital costs … can lead to serious performance problems.

Incentive regulation in its various forms has had a big impact on regulatory practice. However, this literature has largely neglected to embed these questions within a broader regulatory contract designed to protect and promote sunk investment. Crocker and Masten (1996) argue that the transactions cost perspective sheds important light on the design of incentive arrangements:

Recent research on price adjustment processes, however, warns of the hazards of employing overly formulaic price adjustment processes in precisely the types of settings the characterize public utility transactions. … Overall, the research on price adjustment in private contracts suggests that price-cap regulation could work adequately for short periods or under relatively stable conditions but that too rigid a regulatory structure poses the risk of political and legal conflicts if shifts in technological or demand conditions leave either consumers or utilities greatly disadvantaged. Issues in the design and evaluation of price-cap regulation for which transaction-cost research can offer insights include: what adjustment processes should be employed, how often prices should be adjusted, when formulaic prices should be abandoned or the formula itself be modified, and how such modifications should be effected. … Efforts to instill regulation with contract-like, high-powered incentives begs the question: why regulation rather than franchise contracting? If the design is to replicate market incentives, why not the real thing?25

25 Pages 30-31.
VII. Pricing IV: The Weighted Average Price Cap

In regulatory practice the notion of a ‘price cap’ has two senses. The first sense, which was highlighted in the previous chapter, relates to the notion that the price is set independent of the costs of the regulated firm. The second sense relates to a cap on the weighted average of a basket of prices. Regulatory practice has tended to combine these, although their theoretical justification is quite separate. This section looks at the theoretical foundation for the use of a weighted average price-cap on a basket of prices.

Following the paper by Baumol and Bradford (1970), Ramsey–Boiteux pricing was all the rage amongst regulatory economists. But some regulators complained that Ramsey pricing was impractical because it required detailed information about the shape of demand and marginal cost curves. This raised the question: Could there be some regulatory mechanism which achieves Ramsey pricing without requiring the regulator to collect additional information? If that mechanism could also give rise to incentives for cost efficiency, then, economists believed, we would have found the long-sought ‘holy grail’ of regulation: a simple practical mechanism achieving both efficient pricing and efficient production.


This was the promise of a paper by Vogelsang and Finsinger which appeared in the late 1970s. At that time, although regulated firms could in principle adjust or re-balance their regulated tariffs, as long as the firm earned no more than its regulated rate-of-return overall, in practice the regulator often opposed such rebalancing. Vogelsang and Finsinger’s key innovation was the observation that the regulated firm might have more information than the regulator about the shape of the demand curve it faces. In this context, allowing the regulated firm some discretion over its prices – provided that discretion was held in check by a suitable regulatory mechanism – might allow overall efficiency to be improved.

The particular mechanism that Vogelsang and Finsinger proposed was that the regulated firm be allowed to adjust its regulated prices subject to the requirement that a weighted average of the firm’s prices be less than a threshold. Specifically, they proposed the now-familiar requirement that the regulated firm be allowed to adjust its prices subject to the requirement that an index of those prices (known as a Laspeyres index) is decreasing over time. Specifically, the new proposed prices multiplied by the previous year’s quantities divided by the previous year’s prices multiplied by the previous year’s quantities should be less than one.

The merit of this pricing restriction is that, in theory, it ensures that consumers as a group are not left worse off by the pricing adjustments proposed by the regulated firm – some prices may go up, but other prices must go down in a way that leaves consumers as a group no worse off overall.

Vogelsang and Finsinger go on to prove that under conditions of increasing returns to scale, if the regulator allows the firm to adjust its prices each period according to this rule, and if the firm myopically maximises its profit each period subject to this pricing constraint, the resulting prices will converge, as if by magic, to Ramsey prices. As Brennan (1989) summarises:
Permitting prices to be flexible, subject to maintaining a constant weighted average where the weights are based only upon the quantities sold for each service, can approximate the behaviour of a firm maximizing profit subject to an aggregate consumer surplus constraint. Flexibility to change prices is in this sense ‘optimal’ …\textsuperscript{26}

Moreover, as Vogelsang and Finsinger suggest, the ‘regulatory lag’ built into this process will yield incentives for productive efficiency in the manner we have discussed earlier.

In effect, Vogelsang and Finsinger’s major innovation had two parts: first, they proposed the idea of allowing the regulated firm the discretion to adjust its prices subject to a cap on a basket of tariffs. If the weights in that basket are chosen ‘optimally’ the regulated firm will choose the efficient Ramsey prices. Second, they showed how, even if the weights in the basket were not chosen optimally at the outset (and, instead, were simply proportional to the observed quantities consumed), over time the firm would be induced to choose the Ramsey optimal prices. These two central claims are striking.

Vogelsang and Finsinger’s paper triggered something of a revolution in thinking. Perhaps the most significant development in regulatory practice of the subsequent decade was the proposal by Professor Stephen Littlechild, just a few years later, to use a weighted-average price cap to regulate the newly-privatised British Telecom. Professor Littlechild’s proposal combined the suggestion of Vogelsang and Finsinger to allow the regulated firm the flexibility to adjust its prices, with the suggestion of Baumol (discussed earlier) to allow the level of the overall prices to increase over time at the rate of inflation (known as CPI in Australia) less a pre-determined efficiency factor (known as X). This became known as ‘CPI-X’ regulation. The combination of flexibility under a weighted average price cap, with the cap allowed to evolve at CPI-X, has subsequently been adopted as the standard approach in the regulation of many industries around the world.\textsuperscript{27}

Vogelsang and Finsinger’s results are somewhat more limited than it may first appear. To begin with, they assume the cost and demand conditions do not change over time. If cost and/or demand are changing over time the proposed mechanism may not converge to Ramsey pricing or may not converge at all. Moreover, they assume that the regulated firm will behave ‘myopically’ – that is, will maximise its profits in each period, ignoring the effect on future periods. It seems more realistic to assume that the firm will quickly recognise that how it sets its prices today will affect the regulatory limits imposed on it in the future – and will take this into account in its optimisation.

Bradley and Price (1988, discussed below) point out that the rate at which the prices converge to Ramsey prices depends on the starting point:

The speed of adjustment towards Ramsey pricing depends not only on X but also on the weights in the tariff basket formula. The nearer the initial weights are to Ramsey quantities the quicker the adjustment will be. This is pertinent in the case of newly privatised industries. Management of these industries when nationalised was notorious for the ambiguity of both objectives and the means by which they were to be achieved. It is unlikely that they were profit maximisers or that much more than lip service had been

\textsuperscript{26} Brennan (1989), page 134.

\textsuperscript{27} In fact, the combination of these two policies has become so common that many commentators use the term ‘price cap regulation’ to refer to both, when in principle they could each be adopted independently of the other.
paid to the notion of marginal cost pricing. ... In such circumstances attempts to estimate Ramsey quantities as weights for the tariff basket formula could be very worthwhile.

Vogelsang and Finsinger’s mechanism fits well within the neoclassical tradition in economics, with its focus on deadweight loss. The transactions-cost perspective, with its focus on protecting the sunk investments of customers, suggests that although it may be valuable to allow the service provider to adjust its tariffs – that adjustment should be subject to the constraint that no individual customer is left worse off.

This concern to protect individual customers can be seen clearly in regulatory practice. Where firms have been given discretion to change their prices subject to a weighted average price cap, it is very common for regulators to impose additional limits on the rate of change of individual prices, known as ‘side constraints’. For example, it is common to limit the increase in individual prices from one year to the next to CPI+2%. If Vogelsang and Finsinger’s analysis is correct, side constraints are not only unnecessary, they actively impede the achievement of efficient pricing. Yet side constraints persist in regulatory practice. We might suggest that these side constraints reflect a concern to ensure that no individual customer is left worse off from a change in tariffs.


Within just a few years of Professor Littlechild’s original proposal, a form of price-cap regulation had been applied to three newly-privatised companies – British Telecom (BT), British Gas (BG), and the British Airports Authority (BAA). But the form of the price cap applied to BG and the BAA differed in an important way: while the price cap on BT effectively placed a limit on a weighted average of prices (as discussed above), the cap on BG and BAA placed a limit on the average revenue earned by BG or BAA. In the case of BG the limit was on the average revenue per gigajoule of energy delivered; in the case of BAA the limit was on the average revenue per passenger.

The ‘average revenue cap’ approach to regulation is still in use in Australia today. The National Electricity Rules allow distribution businesses to propose a ‘revenue yield’ approach to regulation as one of the permissible forms of control. In the past, several distribution businesses have used this form of control.

However, since this approach was first adopted in the UK, academic economists have consistently criticised this approach, starting with this paper by Bradley and Price. Bradley and Price point out that, under average revenue regulation, the prices the firm will choose – that is, the prices which maximise the regulated firm’s profits – will have no necessary relationship to Ramsey prices. That is, there is no tendency for the firm to choose efficient prices. Cowan (1997) shows that, in fact, the regulated firm can have an incentive to set prices so high under average revenue regulation that the customers would have been better off if no regulation had been applied!

Armstrong and Sappington (2007), in their survey of the theory of regulation, also emphasise the inefficiency of average revenue regulation. They point out that:

---

28 Such limits are explicitly set out in the National Electricity Rules, for example.
(a) consumer welfare will be lower under average revenue regulation than if the firm were forced to charge the same price for all of its services(!); and

(b) a reduction in the regulated average revenue can leave consumers worse off.

Under average revenue regulation the regulated firm effectively receives the mandated or regulated average revenue for each additional unit of output that it sells – no matter to whom, when, or where it sells that output. The firm clearly has an incentive to try to increase the quantity it sells for those services where the marginal cost is less than this regulated average revenue. It can do that by lowering the price it charges for those services with low marginal cost – potentially lowering the price well below marginal cost for those services, with little or no impact on its overall revenue. At the same time, it has an incentive to raise the price for services with a higher-than-average marginal cost.

Overall, the message is clear: average revenue regulation leads to inefficient outcomes. More than two decades ago, Bradley and Price conclude:

An average revenue constraint is clearly inefficient. … In determining the appropriate form of price regulation for privatised UK industries, the average revenue constraint seems to have been preferred wherever a common unit of supply could be defined. Future regulation of privatised industries in the UK (particularly electricity and water) and proposals for price-capping in the US should not depend merely on the identification of some apparently common supply unit. Although such a unit may appear obvious … an average revenue constraint based on this would be inferior to a tariff basket restriction.
VIII. The Rise of ‘The New Economics of Regulation’

In section VI we explored the rapid growth in interest in incentive regulation that emerged in the 1970s and early 1980s. In parallel there were key developments in economic theory emphasising the primary role of information and incentives in all economic transactions – and, in particular, in decisions as to the design of contracts and the choice of a contracting partner. It was realised that certain key economic relationships are best characterised as involving an ‘information asymmetry’ – where one party to the transaction is better informed than the other.

Two types of asymmetric-information models emerged in the literature. In the first type of model, one of the parties has private information about a characteristic that is unique to that party and that existed prior to contracting. These are known as ‘hidden information’ or ‘adverse selection’ models. In the second type of model, the two parties have the same information when they enter into the contract, but one of the parties subsequently takes an action which is not perfectly observed by the other party to the transaction. These are known as ‘hidden action’ or ‘moral hazard’ problems.

Both of these models have important applications throughout economic life. But, most importantly for our purposes, they have direct application to the theory of economic regulation – particularly the design of the optimal long-term regulatory contract.


In the early 1980s, Baron and Myerson were the first to propose that the regulatory problem could be viewed as an asymmetric information problem. In their framework the regulated firm has an unknown cost parameter (e.g., an unknown fixed cost, or an unknown marginal cost – so this is an ‘adverse selection’ model) and the task of the regulator is to choose a regulatory framework which maximises a weighted average of consumers’ surplus and producers’ surplus, with more weight given to consumers’ surplus. Importantly (and unrealistically) they assume that the regulator is allowed to make transfers directly to the regulated firm out of government funds.

Under these assumptions, if there was no information asymmetry, the efficient regulatory mechanism would simply involve pricing at marginal cost. In the presence of the information asymmetry, however, there arises a trade-off between ‘rent extraction’ and ‘efficiency’. Baron and Myerson show that the optimal approach for the regulator involves setting a price equal to marginal cost for low-cost firms, but allowing them potentially substantial rents in the form of transfers. The highest-cost firms, on the other hand, earn no rent overall, but their price is distorted above marginal cost (interestingly, potentially even above the monopoly price).

The central idea of a trade-off between rent-extraction and efficiency is important and is common to all asymmetric-information models. The relevance of the Baron and Myerson paper to real-world regulatory policy is limited, but this paper remains an important milestone.


The Baron and Myerson paper assumed that the firm’s cost was fixed and not under the control of the management of the firm. This assumption seems unrealistic – the longer
the firm is regulated the greater the proportion of the firm’s costs which are directly controllable by the management of the firm. In the medium term, it seems more reasonable to assume that costs are primarily the result of managerial effort.

Laffont and Tirole introduce a model of regulation incorporating both moral hazard and adverse selection. They assume that managerial effort (moral hazard) can reduce the firm’s costs from some unknown starting level which we might call their ‘inherent cost’ (adverse selection). A key difference with their model is that the regulator is able to observe the firm’s costs ex post – after conducting some form of audit. However, they retain the assumption that the regulator is able to make transfers to the regulated firm, which is not realistic in most industries. Their model is perhaps better viewed as a model of procurement – where the government is seeking to purchase a public good from a firm at least cost.

Laffont and Tirole show that the optimal regulatory (or procurement) contract is linear in the firm’s out-turn costs. In other words, the optimal contract has the form of a fixed payment plus a sliding scale which depends on the firm’s out-turn costs. Importantly, Laffont and Tirole introduce the idea of offering the regulated firm a ‘menu of contracts’. Under this approach the regulator would allow the firm to choose between, say, a high-powered incentive (a ‘fixed price’ or ‘price cap’ arrangement) and a lower-powered incentive (a ‘cost of service’ arrangement) which allowed the firm’s prices or revenues to vary with its out-turn costs. In their model, firms with a low inherent cost choose a fixed-price contract with corresponding strong incentives for cost-reducing effort. On the other hand, firms with a higher inherent cost choose a contract in which the regulator reimburses a share of the cost differences from some target. The lower the inherent efficiency of the firm, the greater the extent of cost-sharing (the closer the arrangement is to ‘cost of service’ regulation).


A number of other papers emerged in the late 1980s and early 1990s. Some are worth mentioning here, including a survey by Baron (1989), published in the *Handbook of Industrial Organization*, two papers by Lewis and Sappington (1988a, 1988b) which explored the implications for uncertainty in demand, and the technically difficult textbook by Laffont and Tirole: *A Theory of Incentives in Procurement and Regulation*.

In 1994, Jean-Jacques Laffont took the opportunity of his Presidential Address to the Econometric Society to set out a synthesis of the new theory of regulation. He started by reviewing the criticisms of rate-of-return regulation from the 1960s and 70s, emphasising that these criticisms are little more than ‘a list of defects of this procedure with no clue whatsoever on whether those defects are the outcomes of optimal trade-offs’. Without some framework for analysing the key trade-offs we can’t be certain that the key features of rate-of-return regulation – although those features are admittedly imperfect – are not in fact the least imperfect arrangements given the constraints of imperfect information and inability to commit to long-term contracts. One of the key strengths, he argues, of the new approach to regulation is explicit recognition of the information and incentive problems that lie at the core of the regulator’s task.

As in his previous work, Laffont sets out the basic regulatory problem as a ‘principal–agent’ problem, with the regulator as the ‘principal’ and the regulated firm as the ‘agent’.
One of the important contributions of this paper is to set out the conditions under which the regulator can separate the task of setting efficient prices from the task of designing incentives. In the Baron and Myerson model the optimal regulatory contract involved distorting prices above marginal cost for higher-cost firms. However, Laffont shows that under certain conditions on the cost function, the regulator can deal with these two issues separately – setting theoretically efficient prices, on the one hand, while designing an efficient incentive framework, on the other. This is known as the incentive-pricing dichotomy. Laffont also discusses the implications for regulation when incentive contracts are auctioned (as Demsetz proposed in 1968) and when they are repeated over time (as they are, of course, in practice).

Finally, Laffont asks a further important question: We can model the regulator/regulated firm interaction as a principal–agent relationship, but what about the relationship between the state or government and the regulator? Shouldn’t we also model that arrangement as a principal–agent relationship, with the government as the principal and the regulator as the agent? In other words, shouldn’t we model a ‘hierarchy’ of governance relationships?

Laffont points out that when there is a risk that the regulator will be ‘captured’ by the regulated firm, it can be optimal to lower the power of the incentive on the regulated firm, so there is less rent to be shared with the regulator, even if that results in lower cost efficiency outcomes. More generally, Laffont argues that we will not be able to fully explain the regulatory practices we observe without a model which integrates political processes and constraints.

This new theory of regulation represented something of a revolution in the thinking of theorists that paralleled, at least in time, the revolution in policy associated with the introduction of ‘price cap’ and other forms of ‘incentive regulation’. The new theory seemed to offer significant new insights, especially with regard to the fundamental role of information asymmetries. But the models tend to be complex, with quite restrictive assumptions, and the policy implications have been limited. Some respected academics have been quite critical of this theory.

For example, Crew and Kleindorfer (2002) argue that the ‘Achilles heel’ of these theories is – paradoxically – that they rely on information that is not available to the regulator. In particular, the theories often assume that certain information, such as the probability distribution of key parameters, is ‘common knowledge’. Crew and Kleindorfer argue that this assumption is possibly acceptable when the agent can walk away from the contract offered by the principal – but this is not the case when the agent is a regulated firm with substantial sunk costs. In addition, Crew and Kleindorfer make the following argument:

> The promise of these mechanism-design-style theories was ostensibly considerable. They promised none other than the holy grail of X-efficiency, something previous regulation had manifestly failed to deliver. X-efficiency, however, was only achieved if two conditions … were met. The first condition was that achievement of the promised X-efficiency required that the regulator concede some information rents to the firm. … Why this would not be a fatal flaw in the whole scheme was never considered. … No regulator can even admit that it allows the firm to retain information rents let alone commit to such a practice. For the regulator this is a congenital problem of far greater magnitude than has been recognized in economic theory. How do these rents differ so much from the old style monopoly rents that would make them acceptable to the regulator when it was monopoly rents that were the principal motivation of regulation in the first place? … In effect, the theory proceeds by ignoring an immutable institutional
constraint, namely that neither commitment nor its associated information rents are reasonable assumptions. Other than being a rich source of classroom exercises, this theory seems to have found no takers in practice.

Perhaps a further criticism we can make is that these theories fail to embed the incentive problem within a context of protection and promotion of sunk investments – and the resulting need for a long-term contract in the first place.


As noted in the previous section, a number of excellent papers have emerged in the last few years which provide a valuable synthesis and summary of the literature on incentive regulation. One of these papers (Joskow, 2007) was discussed above. Another valuable survey was prepared by Armstrong and Sappington for the third volume of the Handbook of Industrial Organization. This paper is a tour de force which defies simple summary here. The presentation of the key insights of the new theory of regulation is straightforward and well organised. The authors conclude by identifying the following five general principles:

First, when a regulated firm has privileged information about the environment in which it operates, the firm typically is able to command rent from its superior information. Second, to help limit this rent, a regulator can design options from which the firm is permitted to choose. When designed appropriately, the options induce the firm to employ its superior industry knowledge to realize Pareto gains. Third, the options intentionally induce outcomes that differ from the outcomes the regulator would implement if he shared the firm’s knowledge of the industry. These performance distortions serve to limit the firm’s rent. Fourth, a benevolent regulator is better able to limit the firm’s rent and secure gains for consumers via the careful design of regulatory options when he is endowed with a broader set of regulatory instruments and more extensive commitment powers. Fifth, when the regulator’s commitment powers are limited, it may be optimal to limit the regulator’s access to information, in order to limit inappropriate use of that information.

Armstrong and Sappington recognise that this new theoretical approach to regulation has its drawbacks:

While this normative approach can provide useful insights for the design and evaluation of real-world regulatory policy, the approach has its limitations. In particular: (i) all relevant information asymmetries can be difficult to characterize precisely; (ii) the form of optimal regulatory policies is not generally known when information asymmetries are pronounced and multidimensional; (iii) a complete specification of all relevant constraints on the regulator and firm can be difficult to formulate; (iv) some of the instruments that are important in optimal reward structures (such as transfers) are not always available in practice; and (v) even the goals of regulators can be difficult to specify in some situations. Therefore, although formal models of optimal regulatory policy can provide useful insights about the properties of regulatory policies that may perform well in practice, these models seldom capture the full richness of the settings in which actual regulatory policies are implemented.

Armstrong and Sappington go on to survey a variety of ‘simple, practical regulatory policies’ and draw the following interesting (if overly-generic) conclusions:

The analyses of optimal and practical regulatory policies together provide at least four important observations. First, carefully designed regulatory policies often can induce the
regulated firm to employ its superior information in the best interests of consumers. Although the objectives of the regulated firm typically differ from those of society at large, the two sets of objectives seldom are entirely incongruent. Consequently, Pareto gains often can be secured. Second, the Pareto gains are secured by delegating some discretion to the regulated firm. The (limited) discretion afforded to the firm is the means by which it can employ its superior knowledge to secure Pareto gains. The extent of the discretion that is optimally afforded the firm will depend upon both the congruity of the preferences of the regulator and the firm and the nature and extent of the prevailing information asymmetry. Third, it generally is not costless to induce the firm to employ its superior information in the best interests of consumers. The firm typically will command rent from its superior knowledge of the regulatory environment. Although the regulator may place little or no value on the firm’s rent, any attempt to preclude all rent can eliminate large potential gains for consumers. Consequently, the regulator may further the interests of consumers by credibly promising not to usurp all of the firm’s rent. Fourth, the regulator’s ability to achieve his objectives is influenced significantly by the instruments at his disposal. The regulator with fewer instruments than objectives typically will be unable to achieve all of his objectives, regardless of how well informed he is about the regulatory environment.
IX. Yardstick Regulation and Benchmarking

Achieving incentives for short-term productive efficiency is largely a matter of breaking the link between the regulated firm’s prices/revenues and its own expenditures. However, breaking this link creates its own potential problems – for example, where revenues are de-linked from costs, a large change in the regulated firm’s costs may induce the service provider or its customers to use litigation or political pressure to seek to re-open the regulatory determination (or, at least, to cut service quality and investment) potentially resulting in a political change to the regulatory arrangements.

Ideally, to overcome these problems, we would like to de-link the regulated firm’s prices/revenues from its own controllable costs, while allowing the firm’s prices/revenues to vary according to shifts in costs that are outside its control. As we have seen, the original proposal for CPI-X was an attempt to allow prices/revenues to better track one possible measure of broad industry-wide costs, allowing for fixed and longer regulatory periods. But can we do better?


In the 1980s Andrei Schleifer proposed setting a regulated firm’s revenue/prices equal to the average of the costs of other comparable firms. If two firm’s costs are jointly affected by factors which are partly outside the control of each firm (i.e., if they are correlated), information on the relative performance of each firm relative to the average performance contains more information about the extent of managerial effort at keeping costs down at each firm individually, than does the out-turn level of cost for each firm separately. Rewarding a firm on the basis of changes in its own costs results in rewards/penalties which depend on factors that are partly outside the control of the firm (such as industry-wide changes in input costs). However by rewarding the firm when its own costs go down relative to a comparator firm, the regulator can eliminate the effects of industry-wide cost shocks, thereby more directly rewarding the desirable management behaviour the regulator wishes to achieve. As Schleifer observes:

By relating the utility’s price to the costs of firms identical to it, the regulator can force firms serving different markets effectively to compete. If a firm reduces costs when its twin firms do not, it profits; if it fails to reduce costs when other firms do, it incurs a loss.

Schleifer presents a simple model based on N firms which are known to be identical ex ante. Each firm is offered a simple two-part tariff with a variable part equal to the average of the observed marginal cost of the other firms; and a fixed part equal to the average of the observed fixed cost of the other firms. This simple regulatory scheme induces each firm to choose the socially-efficient level of cost-reducing effort, and results in pricing at the socially-efficient marginal cost.

But what if the regulated firms are not identical? In this case, Schleifer proposes carrying out a regression to estimate a ‘cost function’ which reveals how costs vary with all known cost parameters. If the regulator is able to account for all of the factors that result in diversity in costs, this mechanism will also yield the first-best efficient level of managerial effort. But, if some heterogeneity between the firms is not accounted for (so that the R-squared of the regression is less than one), Schleifer points out that the first-best outcome will not be attainable and the scheme may result in insufficient revenue for the regulated firms.
This scheme, by breaking the link between a firm’s revenues and its own cost out-turns, results in high-powered incentives to reduce costs. But, as discussed in section VI, high-powered incentives are not always desirable. Schleifer recognises some of these problems, such as the need for a high level of regulatory commitment and the risk of cutting quality to cut costs:

Yardstick competition works because it does not let an inefficient cost choice by a firm influence the price and transfer payment that that firm receives. It is essential for the regulator to commit himself not to pay attention to the firms’ complaints and to be prepared to let the firms go bankrupt if they choose inefficient cost levels. Unless the regulator can credibly threaten to make inefficient firms lose money (or, alternatively, can prove in court that firms chose to be inefficient and that their practices were imprudent), cost reduction cannot be enforced.

A form of yardstick regulation is used in the US Medicare system, but not without problems. Schleifer notes that hospitals that are compensated for each procedure on the basis of the average expenditure of other comparable hospitals retain incentives to deny services, turning away patients, or reducing patients’ length of stay, or discharging them and then readmitting them as a new patient, or simply foregoing new and better procedures. Similar incentives to reduce quantity and quality, he notes, would be expected to arise in a regulatory context and would need to be adequately controlled.

Finally, Schleifer recognises the potential problem of collusion. If each firm’s revenue depends on another firm’s reported costs there is scope for an agreement where all firms agree to report higher costs. Schleifer suggests that the regulator could investigate and prosecute instances of collusion. In any case, he suggests, where the number of comparator firms is large, collusive strategies may not be sustainable.

Schleifer’s central idea was not taken up immediately. As we saw earlier, during the late 1980s and 1990s regulators were busy exploring RPI-X forms of regulation. However, in the late 1990s some regulators around the world started investigating the possibility for using comparative performance information in regulatory practice. In particular, some regulators started using techniques which have come to be known as ‘benchmarking’. Benchmarking compares the relative productive efficiency of a regulated firm with a set of comparator companies and then bases the regulated firm’s revenue allowance at least in part on the cost out-turn of those other companies. The most common practice does not set the revenue allowance directly on the average of the cost out-turns of other firms; rather, the most common practice involves setting a firm-specific ‘X factor’ which, in effect, reduces the allowed revenue for those firms which are above the industry average (or, in some cases, above an assumed hypothetical efficient frontier).

There is considerable enthusiasm for regulatory benchmarking in some circles. This enthusiasm parallels the initial enthusiasm for price-cap regulation – as it seemed to offer the promise of high-powered incentives for productive efficiency with minimal regulatory involvement and oversight. Over time, this enthusiasm for price-cap regulation has been tempered, as in practice price-cap regulation more closely resembles traditional rate-of-return regulation, with a built-in fixed regulatory lag. However, enthusiasm for various forms of benchmarking – such as the use of total factor productivity (TFP) – still remains.
A number of papers on benchmarking have appeared in recent years. Many of these urge caution in the use of benchmarking in regulatory processes. The problems that have been highlighted include:

- Comparing productivity of different firms in a single period of time (such as a year) requires some methodology for allocating a share of the long-lived sunk investment to that year. There is no one correct theoretical way to do this. Different choices will result in different apparent relative efficiency. Failure to include capital investment in the benchmarking risks introducing incentives to substitute capital for labour or vice versa.

- Even if it were possible to accurately allocate capital services to a particular period, there remains the problem that different firms provide a very large number of different services using different technologies, different physical environments, and different proportions of inputs. It is not possible, with reasonably obtainable data on inputs and outputs, to take account of all the possible legitimate different cost drivers.

- There is no one correct methodology for measuring costs or estimating cost functions. Different methodologies yield differences in apparent efficiency.

- Higher-powered incentives brought about by benchmarking increases the incentive on the regulated firm to behave strategically to manipulate information (or to collude or merge) so as to change its perceived efficiency.

- Benchmarking implicitly assumes that all of the difference in the apparent efficiency of the firm relative to the benchmark (defined as either the average or the frontier) is due to a lack of effort on the part of the managers of the firm which can be eliminated within a certain period of time and is not due to measurement error or mis-specification of the underlying model. This assumption is very strong and is unlikely to hold in practice.


In one paper from recent years, Jamasb, Nillesen and Pollitt look at how benchmarking affects the incentives of regulated firms to engage in strategic behaviour (which they define to be profit-maximising behaviour not directly targeted at improving productivity). They observe that:

Regulators should recognise that their benchmarking exercise inevitably shapes the efforts and directs considerable resources of the firms towards the make-up and variables of these models. However, while benchmarking can measure ‘true’ performance improvements, gaming can sometimes produce illusive or ‘virtual’ performance improvements. Therefore, benchmarking models need to strike a balance between reflecting the main performance drivers of the business in question and


30 These problems are in addition to the problems with high-powered incentives discussed above.
reducing incentives for engaging in unproductive method or model-induced strategic behaviour.

The authors identify several different forms of gaming. One form of gaming relates to efforts by the regulated firm to influence the selection of the benchmarking model that is adopted, including the choice of method, model, variables (and their weighting), and the methodology for translating efficiency scores into X factors. Another form of gaming relates to cost shifting (i.e., between capex and opex) to present the firm’s performance in a more positive light. The firm may also defer or bring forward expenditure to affect apparent relative performance at a given point in time. Finally, firms may engage in strategic mergers – particularly in the case of frontier techniques such as data envelopment analysis (DEA) – so as to shift the apparent performance of other firms in the sector.

The paper sets up a possible benchmarking framework based on DEA and shows how the regulated firms might act in order to strategically manipulate the outcome. The authors conclude:

Regulatory benchmarking does not eliminate the issue of asymmetric information on firm’s costs and efficiency improvement effort as known under rate of return regulation. Rather, it adds new dimensions to this issue and the ways in which firms can behave strategically. Countering strategic behaviour can partly be overcome by increasing data accuracy and improving data collection procedures. The information requirement for reliable regulatory benchmarking therefore appears to be higher than initially expected. The continued efforts made by regulators using benchmarking to improve data quality are testament to this fact.


Graham Shuttleworth of NERA has been one of the more cogent critics of benchmarking. His paper from 2005 summarises the experience with regulatory benchmarking and highlights its shortcomings. He concludes:

at best, benchmarking can help to focus regulatory enquiries, but … it shows no prospect of becoming a substitute for detailed evaluation of each regulated utility’s own costs.

Shuttleworth notes that there are a variety of methodologies for comparing efficiency across firms, including both parametric and non-parametric techniques, and methodologies based on average efficiency versus those which seek to determine frontier efficiency:

Regulators have a wealth of benchmarking techniques at their disposal. However, the sheer variety of methods represents a problem in regulation, since each will produce results that differ, sometimes significantly. Indeed, the problem with benchmarking is not a lack of techniques, but their unsuitability for a contentious and financially important task like regulation.

Shuttleworth’s first criticism is that there is, at best, only enough data available to determine broad cost relationships, and certainly not enough to determine an accurate overall cost function incorporating every factor that could possibly shift a firm’s costs:

A moment’s consideration should be enough to convince anyone that the costs of an electricity distribution network depend upon a large number of factors, including
As noted above, there is no correct way to allocate the costs of a sunk investment to a particular period for the purposes of efficiency comparison. As a consequence some regulators focus their benchmarking on opex alone. But this creates its own distortions, as Shuttleworth explains:

In 1999, Ofgem applied benchmarking to opex, and found that the Southern distribution network was on the frontier, whilst the Seeboard distribution network lay some way above it. Ofgem gave Seeboard a revenue allowance for opex that demanded a rapid reduction. However, it became apparent during the review that Southern had favoured the use of capex (e.g., asset replacement) rather than opex (e.g., asset maintenance), whilst Seeboard had favoured opex rather than capex. The difference between their respective opex performance therefore reflected different choices over the mix of inputs, rather than differences in efficiency. In its final proposals, therefore, Ofgem had to undo the effects of its opex benchmarking, by awarding Seeboard additional revenue … and penalising Southern for excessive investment, thereby rendering the benchmarking largely irrelevant. In the end, it is likely that both companies received revenue allowances in line with their expected costs.

Perhaps the most important of Shuttleworth’s criticisms relates to the interpretation of the results of a benchmarking exercise:

Most literature on benchmarking interprets the gap between the frontier and any observation as ‘inefficiency’, i.e., as unnecessary costs caused by inefficient management of the company, and not due to the conditions in which the company operates. … The residual gap between the frontier and any observation could, in principle, be due to any factor not contained in the model. Given a small sample size, benchmarking models cannot contain more than a few variables … However, the costs of each network depend on a large number of factors, some highly specific to its location, or possibly even unique. For example, the current design of a network will depend on the history of its development and whether demand growth occurred in large lumps … or gradually …

No benchmarking model can ever hope to capture all these factors in the model specification. Some of the omitted factors or unique factors explain part of the ‘residual’, i.e., the unexplained gap between observed costs and the estimated frontier. Until anyone can claim with certainty that a benchmarking model has captured every possible cost driver, it is incorrect and misleading to ascribe the residual to ‘inefficiency’ or to describe the benchmark as a measure of ‘efficient costs’. Instead, one must acknowledge that the residual measures no more than the element of observed costs that the model has failed to
explain. On that basis, it provides no grounds for disallowing certain costs or anticipating rapid rates of cost reduction.

Shuttleworth offers the following concluding paragraphs:

Benchmarking may still have a role to play in regulation, as long as regulators recognise that the residual merely measures the extent to which the model has failed to explain costs, and not the extent to which companies are inefficient. Stable and predictable regulatory techniques have to rely on other forms of evidence. Benchmarking techniques can nevertheless help in the following cases:

- To decide which companies out of a large sample deserve closer (and more objective) examination, so that scarce investigative resources are allocated efficiently; …

- As a preliminary step in the detailed investigation of each company’s costs, aimed at identifying decisions that should be investigated further for ‘imprudence’.

As an interim step in an investigative procedure, benchmarking may help regulators to appraise large volumes of data on costs and outputs. However, benchmarking techniques are not robust and cannot replace detailed investigation of costs. Any attempt to rely entirely on benchmarking to set revenue allowances is bound to involve subjective and arbitrary choices. For the sake of transparency and stability in regulation, therefore, it will be necessary to regard benchmarking as an investigative technique, not an alternative method of setting revenues.
X. One-Way Access Pricing

Historically many natural monopoly industries were highly vertically-integrated – from the upstream suppliers to the final consumers. For example in electricity, the state electricity monopolies owned and operated generation, transmission, and distribution facilities and even, in many cases, coal mines. In telecommunications, the US company AT&T was historically integrated into equipment manufacturing, and the provision of long-distance and switching services, through to the leasing of customer-premises equipment. In both industries these firms were regulated as monolithic vertically-integrated entities.

One of the key insights of the major regulatory reform movements of the past two decades was the observation that, in many cases, price regulation could be limited to just the natural monopoly components of these industries. In the case of electricity, the natural monopoly elements are the wires businesses – transmission and distribution. In telecommunications, the remaining natural monopoly is probably in the ‘last mile’ wired (fibre or copper) connection to individual homes in suburban and rural areas. Focusing the regulation on these wholesale services allows scope for other firms to enter and compete in the potentially-contestable upstream and downstream markets. This approach can simplify the regulatory task.

As a general rule, it is more difficult to regulate natural monopoly businesses which are also involved in competitive (non-regulated) activities. The reason is that vertically-integrated monopoly service providers have an incentive to restrict or prevent the development of competition downstream so as to allow them to re-capture some of the monopoly rent that is otherwise lost to regulation.

For this reason, in many liberalised industries the natural monopoly businesses were separated from the competitive activities and prevented from re-integrating into these related sectors. For example, in electricity, transmission and distribution businesses were separated from generation and retailing. Airports in Australia are not allowed to own or be owned by airlines. The operator of the interstate rail track network in Australia (ARTC) is not permitted to run trains over that track. In the US, in the early 1980s the local telephone service business was separated from the long-distance business.

But such vertical separation was not enforced in every case. In most countries, for example, the incumbent telecommunications provider was both subject to access regulation at the wholesale level and allowed to compete against rivals in related downstream businesses. In postal services, no attempt has been made to separate local postal delivery services (i.e., the local mailman) from the rest of the postal business. The third-party access regime in Part IIIA of the Australian Competition and Consumer Act focuses on regulation of the ‘essential facility’ component of a business. It does not also impose structural separation – thereby often creating access pricing problems.

The problem of regulating the price of an essential input is, first and foremost, the same problem of regulating a natural monopoly that has been discussed. All of the material discussed in the sections above still applies. However, when the provider of the essential input also competes downstream, a few issues (which, to an extent were already present) become particularly important. In particular, unlike ordinary sales to independent

---

31 A similar vertical separation was applied to British Rail in the UK.
consumers, the sale of an input to a rival might reduce the sales of the monopolist to its own downstream subsidiary. This creates new pricing issues, which are the focus of the papers discussed below.


The classical access problem arises when a regulated firm is forced to sell an essential input to rivals with whom it competes in providing a homogeneous service in an unregulated downstream market. In this situation, if the monopolist can eliminate competition from the downstream market it can sell downstream at a high (unregulated) price, thereby essentially evading any regulatory controls. If the monopolist makes a sale of the essential input to a rival at a price which allows that rival to undercut the incumbent in the downstream market, the monopolist is essentially ‘shooting itself in the foot’ – i.e., reducing the profit it could earn by selling that product itself. As a consequence, the monopolist has a strong incentive to actively oppose any attempts to force it to sell to its rivals at a price which erodes the monopoly rent it might earn in this supply chain. On the other hand, entrants have traditionally sought access at prices based on the ‘cost’ of providing the input (i.e., either marginal cost or incremental cost, perhaps with some mark-up to cover joint and common costs).

These debates came to a head in the early 1990s as the first OECD countries after the US started to liberalise their telecommunications markets. A dispute that arose in New Zealand between a new entrant (Clear Communications Ltd) and the dominant incumbent (Telecom NZ Ltd) featured prominently in both the policy and academic debates. In these debates, like so many key regulatory debates of the previous twenty years, William Baumol made an important, but (in this case) highly controversial, contribution. This contribution – the ‘efficient component pricing rule’ – was described in a paper published in the *Yale Journal on Regulation* and also in a short book called *Toward Competition in Local Telephony*, both appearing in 1994.32

According to Baumol and Sidak, the efficient price for an essential input is equal to the cost of producing that input plus the opportunity cost of that sale to the monopolist. In other words:

$$ECP = \text{Cost of producing input} + \text{Opportunity cost of sale of input to rival}$$

If the sale of one unit of the input to the rival results in the loss of one unit of sale of the retail product to a downstream consumer, the opportunity cost of the sale to the monopolist is equal to the foregone profit – that is, the retail price less the cost of producing the retail product less the cost of producing the input. Therefore, another way to write the efficient component pricing rule is:

$$ECP = \text{Retail price} - \text{Cost of producing retail product from the essential input}$$

32 The ECPR seems to have been originally proposed by Willig (1979).

33 Armstrong, Doyle and Vickers suggest that this version of the ECPR be given a different name – they suggest the ‘margin rule’.
As Baumol and Sidak point out, variants of this pricing rule had been around for some time under different names, such as the ‘imputation requirement’, the ‘principle of competitive equality’, or the ‘parity principle’.

Baumol and Sidak argue that the ECPR is necessary to achieve economic efficiency. They allege that this price will allow rivals to compete downstream if and only if they are the cheaper provider of the downstream (retail) service. Any price below the ECP will, it is alleged, allow inefficient entrants to enter and compete in the market, to the detriment of overall welfare. Baumol and Sidak also appeal to the ‘competitive market’ standard:

The efficiency of this optimal component-pricing rule is confirmed indirectly by the fact that it yields a price level set in precisely the same way it would be in a perfectly competitive or perfectly contestable market. To see this, imagine a set of landlords competing to rent retail space to tenants. Suppose further, … the space can be used for the landlord’s own profitable retailing establishment. No landlord who can use the property in this way will rent it to anyone for less than the direct incremental cost of the tenant’s occupation of the property plus the opportunity cost to the landlord of the rental arrangement. If the landlord can earn $90,000 by using the property, the tenant will be required to make good the $90,000 that is foregone by renting the property. … Consequently, even in the most competitive of markets, no landlord will rent for less than the fee determined under the efficient component-pricing rule …

A price lower than one set in accordance with the rule … always constitutes an inter-firm cross-subsidy and so invites the assumption of the supplier’s role by a firm that is not the most efficient provider. This result should not come as a surprise. It is well known that economic forces set component prices in competitive markets in this way, and competitive market prices are generally those necessary to achieve economic efficiency. Thus, our efficient result also follows immediately through this indirect route, using the competitive-market standard as the guide to efficient pricing.

Unfortunately – and this is one of the most severe criticisms we can apply to a paper from such an eminent scholar as Baumol – this paper is misleading. Baumol and Sidak’s arguments only apply when the downstream retail price is itself regulated. If the downstream price is regulated, the arguments in this paper make sense – any price lower than the ECP will result in inefficient entry, raising production costs and reducing overall welfare. Baumol, Ordover and Willig (1997), in a later paper, emphasised that: ‘We have … always emphasised that efficiency requires both ECPR and some arrangement that prevents overpricing of both final products and bottleneck input and, consequently, that removes all monopoly profit from the opportunity cost component’.

But, if the retail price was regulated, why would we bother with regulating the wholesale price at all? In fact, we could leave to the incumbent the ‘make-or-buy’ decision: that is, whether to produce the downstream service itself or whether to purchase the downstream service from some other firm. There would be no need for upstream regulation at all.

Moreover, if the downstream price is not regulated and if there is effective competition downstream, downstream competition will ensure that the downstream service is provided by the most efficient provider no matter what the level of the access price.\textsuperscript{34} If the

\textsuperscript{34} Kahn and Taylor (1994) observe that the access price of $7 per ton in conjunction with a final price of $10 a ton for transport ‘is indeed fully compatible with efficient competition between the two, as they point out. That demonstration must, however, not be permitted to obscure the fact that price combinations of $6 and $9, $5 and $8, $4 and $7, and $3 and $6 would likewise ensure that result’.
incumbent is the most efficient downstream provider it will lower the retail price to the point where other downstream rivals are excluded, and no further, and this is the efficient outcome. If some other provider is more efficient than the incumbent, it will compete the retail price down to the point where the incumbent exits the downstream market. Again, that is the efficient outcome. The analogy with the competitive market in the quote above is unhelpful. It is true that no landlord in a competitive market will rent for less than the fee determined under the ECPR. It is also true that no monopoly will voluntarily sell its output for less than the unregulated monopoly price.

In a response to Baumol and Sidak in the same issue of the *Yale Journal on Regulation*, Kahn and Taylor (1994) point out:

> The absolute level of the interconnection charge and of the opportunity costs or contribution it contains are, however, of genuine economic importance. It is this other side of the coin that we wish Baumol and Sidak had emphasized more than they do. True, as they observe, the fact that a firm subject to intense competition will seek to recover the net profits that it loses as a result of making any of its facilities available to competitors means that such a charge cannot be regarded in itself as monopolistic. But a monopolist, too, will seek to recover these ‘opportunity costs’ and by so doing recoup in its charges for the essential input such monopoly profits as it was previously earning from its direct retail sales. And so while efficient component pricing will ensure that the retailing function subject to competition is indeed performed by the most efficient of the rivals, it will not fulfil the other important function of competition – the erosion of monopoly profits.

With the benefit of hindsight it has become clear from theory that a pricing relationship such as that given by the efficient component pricing rule is, indeed, the efficient relationship between efficient wholesale and retail prices under certain specific assumptions. This does not imply, however, that it is good policy to impose such a relationship between wholesale and retail rates, ex ante, without otherwise ensuring that one or other of those prices are set efficiently.

The ECPR does have some advantages. It allows the incumbent to preserve any cross-subsidies necessary to, say, finance a non-commercial service obligation. In addition, it ensures that any and all discrimination in the final, retail prices is fully reflected in the wholesale or access prices – ensuring that entrants are not excluded from some downstream markets. In addition, in principle, it leaves the incumbent indifferent whether or not to provide access – thereby offsetting some of the incentives of the incumbent to restrict or prevent the development of competition.

The ECPR has not been widely taken up by regulators in mainstream regulatory practice. Echoes of the ECPR can be seen in different areas. For example, ‘retail-minus’ pricing, often used in telecommunications regulation, has some of the flavour of ECPR. In many countries, ‘imputation testing’ (that is, testing for a ‘price squeeze’ between wholesale and retail prices) uses principles very similar to the ECPR. I am not aware of other instances where ECPR has been used explicitly.

---

35 Kahn and Taylor (1994), in responding to this section of Baumol and Sidak’s paper, note: ‘This argument is misleading or subject to misconstruction: those ‘opportunity costs’ could just as well be monopoly profits’.

As often seems to be the case in regulatory economics, following the proposal by Baumol and Willig of the ECPR, and its endorsement by the UK Privy Council in the *Clear v Telecom* case, theorists scrambled to catch up with this new policy problem and propose solutions.

In this paper Armstrong, Doyle and Vickers explored the question of the optimal access charge in the case where, if the monopolist sells one unit of the input to its downstream rival, it loses less than one unit of sales from its own downstream subsidiary. This might arise because either the rival’s retail product is not a perfect substitute for the retail product of the incumbent, or because the downstream rivals can partially substitute away from the bottleneck product (i.e., the downstream production technology does not have ‘fixed coefficients’). In this case the opportunity cost of sales to the downstream rivals is lower than in the classic ECPR problem, and the optimal access charge is correspondingly lower. In effect, in this model, the optimal access price is a weighted average of the cost of providing access, on the one hand, and the ECPR access price, on the other. The ECPR itself is seen as a special case of this more general theory.

In a survey paper, Armstrong (2002) summarises this result as follows:

In the guise of the margin rule, [the ECPR] has the virtue of simplicity and being informationally undemanding (at least when compared to the Ramsey approach). All that needs to be known is the incumbent’s retail price and its avoidable costs in the retail segment. However, this margin rule is simply not appropriate except in a few special cases … If the margin rule is used as a basis for policy then inefficiencies will result in situations where (i) the rival’s product does not substitute one-for-one for the incumbent’s or (ii) where rivals have some ability to bypass the incumbent’s access service. On the other hand, our preferred version of the ECPR … is, outside these same special cases, informationally demanding, and various awkward elasticities appear in our expressions for opportunity cost. This suggests that the apparent contrast between the ‘simple’ ECPR approach and the ‘complex’ and informationally demanding Ramsey approach may just be a misleading artefact of simple examples with extreme elasticities.


Laffont and Tirole remind their readers that optimal access pricing is just a special case of the more general problem of optimal pricing by a regulated monopolist. That problem has the general solution given by Ramsey and Boiteux – specifically, where there are increasing returns to scale, the price for each service (including the access service) should be marked-up above marginal cost by an amount which reflects the elasticity of the demand for the service. However, since sales of the access services are, in a sense, a substitute for sales by the incumbent of retail services, this substitution must be taken into account when setting the appropriate Ramsey mark-ups. In certain special cases (such as when access services and retail services are perfect substitutes) the relationship between the access prices and the retail prices satisfies the efficient component pricing rule. In other cases the optimal access price may be larger or smaller than the price given by the ECPR.

Of course, Ramsey pricing has long been criticised as being informationally demanding. To partially overcome this problem Laffont and Tirole propose subjecting the incumbent
to a cap on the weighted-average of its prices, in exactly the same manner as discussed above. The primary innovation proposed by Laffont and Tirole is to include access services in the price cap – hence the term ‘global’ price cap.

As is well known, a price cap induces a firm to select the proper Ramsey structure as long as all goods (including, here, access goods) are included in the definition of the cap and the weights are exogenously fixed at the level of output that will be realized. This result holds for any demand structure and in particular allows for the possibility of strong substitutability between access goods and final goods. That is, a global price cap in principle allows a proper usage based pricing structure apparently without a need for the regulator to know the demand functions.

Laffont and Tirole also make the important point that the incentive on the incumbent to use various non-price tools to foreclose competition depends on the nature of the regulation. Providers of essential facilities which also compete with downstream firms are often accused of pursuing a range of non-price strategies to restrict or prevent downstream competition, such as delays in interconnection, refusal to unbundle, or imposing costly technical requirements. Laffont and Tirole point out that this incentive depends on the relative ‘tightness’ of the final product regulation relative to the access product regulation. When the access regulation is ‘tight’, while final product regulation is loose (or non-existent), the regulated firm has a strong incentive to attempt to force out rivals from the downstream final product market; ‘Current regulations tend to unevenly put more pressure on access charges than on prices of services in competitive segments, and encourage [foreclosure]’. Conversely, when the access regulation is loose or non-existent, but the final product regulation is particularly tight, the firm may voluntarily choose to provide access, and may withdraw from the downstream market. Under a global price cap, in contrast, the regulated firm faces ‘balanced’ regulation for its final and access services, and has no particular incentive to favour one market over the other.

Importantly, Laffont and Tirole also suggest that it may be efficient to encourage entry in the competitive segment, even if the entrant offers an identical product to the incumbent (so that there are no gains to consumers from product differentiation) and even if such entry incurs a fixed entry cost. The reason is that downstream entry may facilitate regulation of the incumbent – by yielding a benchmark for the regulator, in what is known as ‘yardstick competition’. Laffont and Tirole conclude that:

entry may be desirable even if entrants face a fixed entry cost and produce a close substitute on the competitive segment. The gains from yardstick competition are larger, the less regulators are informed about the [incumbent]’s technology on the competitive segment, and the more similar the [incumbent]’s and the competitor’s technologies. If there is a fixed cost which renders entry unprofitable when [using the ECPR] … then the only way to induce entry is by lowering [the access price] further, and it may be worth doing so even if the variety advantage brought by the entrant is small, because the entrant now brings a second advantage, namely, the yardstick.

In some respects, Laffont and Tirole’s contribution was the ‘last word’ on one-way access pricing. In effect, their paper puts access pricing squarely back in the domain of optimal price regulation more generally, with the same tools (such as the price cap) and the same issues. In the subsequent years theorists turned their attention to the more complex and varied questions of two-way network interconnection, which is discussed shortly.
XI. The Regulation of Quality

The previous sections of this paper have focused on matters related to incentives and optimal pricing. But the regulation of prices is only one part of the task of a regulator – another key part is the regulation of what might loosely be called ‘quality’. How should the long-term contract between the service provider and its customers be designed so as to ensure an efficient level of service quality is provided and maintained over time?

The term ‘quality’ refers to different dimensions of the nature of the regulated products or services. The term can be used to describe either a characteristic of a product or service or to describe the range of the goods and services provided by the regulated firm. Quality is usually distinguished from other variations in the characteristics of goods and services in that, other things equal, it is assumed that all consumers prefer higher levels of the ‘quality’ dimension.

Introducing the possibility that the regulated firm might change the nature or range of regulated products on offer complicates the task of the regulator. In earlier sections we saw how, in pricing matters, a regulated firm might seek to vary the price from the efficient level, or might seek to discriminate between different customers on price, or might seek to offer a menu of different tariff options designed to attract different groups of customers. Exactly the same issues arise in relation to quality. As we will see, a regulated firm might seek to vary the quality from the efficient level, might seek to offer different levels of quality to different customers, or might offer a menu of different price–quality options. A key issue, as we shall see, is the ability of the regulated firm to offer different quality levels to different customers, and the different incentives brought about by different regulatory regimes.


According to conventional economic theory, an unregulated monopolist choosing a simple linear price, common to all customers, will choose a price which is above marginal cost, bringing about a reduction in welfare known as the deadweight loss. But what is the corresponding result for quality? Will an unregulated monopolist, choosing a single level of quality common to all customers, choose a level of quality which is too high or too low?

Interestingly, this question wasn’t answered until 1975. In that year Michael Spence showed that an unregulated monopolist, choosing a single level of quality for all its customers, might choose a level of quality that is either higher or lower than the socially-efficient level. Spence pointed out that it depends on how a change in quality changes the demand curve for the product. In Spence’s model, a small change in quality shifts the demand curve slightly. The monopolist is then able to raise the price slightly to the marginal consumer (and to all other consumers). But the social benefit from the quality increase depends not just on how much the marginal consumer values the change in quality but also on how much all of the infra-marginal consumers value the change in quality.

If consumers with a lower marginal valuation for the product value the increase in quality by more (roughly speaking, if the change in quality rotates the demand curve around the vertical or price intercept), the marginal consumer will value the increase in quality by more than the average consumer and the monopolist will have too much incentive to offer
quality – it will choose a level of quality higher than the socially-efficient level. On the other hand, if consumers with a higher marginal valuation for the product value the increase in quality by more (roughly speaking, if the change in quality rotates the demand curve around the horizontal or quantity intercept), then the marginal consumer will value the increase in quality by less than the average consumer and the monopolist will choose a level of quality lower than the socially-efficient level.

Importantly, however, this result says relatively little about how a firm will behave when it is regulated. The level of quality chosen by a regulated firm (when that firm is forced to choose a single level of quality for all its customers) will depend on factors such as marginal cost of additional quality, the effect of that additional quality on sales, the margin between price and marginal cost on those additional sales, the responsiveness of the regulated price to changes in the quality level, and any other rewards or penalties associated with changes in quality.

The simplest case to consider is the case where the regulator sets a fixed price that the firm must charge and imposes no other rewards or penalties for choosing a particular level of quality. In this case the regulated firm will always choose a level of quality too low. The actual level of quality chosen depends on price – marginal cost mark-up. The higher this mark-up, the greater the incentive on the firm to spend money on quality in order to stimulate demand (which increases its profits since price is above marginal cost). If the regulator squeezes the marginal price down to marginal cost, even this incentive disappears and the regulated firm will choose the lowest possible level of quality.  

Laffont and Tirole summarise this key role of the price – marginal cost margin:

> Provided the quality of a service is observed by consumers when purchasing the service … the demand-expansion effect of an increase in quality provides the utility (just like any unregulated firm) with an incentive to supply quality. However, there is no guarantee that the provision of quality will be appropriate. … When quality is underproduced one can manipulate price-cap weights so as to introduce an upward distortion in the price with respect to Ramsey pricing to improve the quality performance [or vice versa]. Such action … raises the incentive for provision of quality. … Price cap regulation is about constraining margins. With low margins, the regulated firm has mild incentives to provide quality. It bears the full cost of the provision of quality and reaps a small fraction of the benefits … It is for this reason that price cap regulation is often accompanied by the introduction of measurements of new indicators of quality.

But the regulator could probably do better than simply manipulate the price – marginal cost margin. The regulator could offer to directly reward the firm with higher prices if it chose a higher level of quality, and vice versa. Spence points out, correctly, that ‘the appropriate strategy is to confront the firm with a schedule of prices and qualities that correspond to an iso-consumer surplus line’. In other words, the regulator should ensure that any change in quality is linked to a matching change in price which leaves the

---

36 It has long been recognised in the regulatory community that the imposition of regulatory controls on prices leads to a need to also control service quality. Alfred Kahn, in his classic 1970 text of regulation, emphasises: ‘Price really has no meaning except in terms of an assumed quality of service; price is a ratio, with money in the numerator and some physical unit of given or assumed quantity or quality in the denominator. Price regulation alone is economically meaningless’. (Kahn 1970, Vol. I, page 21)

consumers indifferent overall. The firm, in seeking to maximise its profits, will then choose a level of price and quality which is socially efficient.

However Spence is not sanguine about the possibilities of identifying the right price–quality trade-off at the margin. The problem is that changes in quality affect the welfare of all consumers – not just those at the margin – so identifying the right price–quality trade-off requires knowledge of how quality affects the entire demand curve. Spence notes that rate-of-return regulation may not be as bad as a pure price-cap regulation as it effectively allows the regulated firm to increase price in response to increased quality, especially when increasing quality increases the stock of capital.

Spence’s model assumes that the monopolist offers a single level of quality to all consumers. Another important strand of the literature on quality deals with the case where consumers have different tastes for quality and the monopolist is able to offer a menu of different quality levels at different prices, and allows consumers to select the price–quality level they prefer. In this case it can be shown that the unregulated monopolist will reduce the quality offered to the customers with the lowest willingness-to-pay for quality below the socially-efficient level.\(^\text{38}\)

As above, this result does not carry over to the case of the regulated firm. Nevertheless, a few papers have explored what minimalist forms of regulation might increase welfare in these circumstances. Besanko, Donnenfeld and White, in a couple of papers\(^\text{39}\), show that in this context either minimum quality standards or maximum price regulation can help reduce the quality distortion for the lowest willingness-to-pay customers – the reason is that the price regulation reduces the gains the monopoly can achieve by forcing some customers to switch to the high-quality service. These authors argue that a modest degree of price regulation will, in this context, always increase welfare, whereas the effects of minimum quality standards are more ambiguous (due to the fact that minimum quality standards may exclude some low willingness to pay customers). However, these papers are primarily targeted at analysing minimal regulations in a context where the dominant firm is largely unregulated. Although interesting, these results do not directly apply to industries where prices and services are primarily determined through regulation.


Quality is, in practice, significantly multi-dimensional. For example, for a regulated telecommunications company, service quality might include factors such as the level of noise on the communications connection, the frequency of congestion, reliability (frequency of outages), time to repair (duration of outages), helpfulness of staff, number of billing errors, and so on. Lynch, Buzas and Berg point out that, at least in the 1990s, regulators were recognising these different dimensions of service quality, but were grading regulators on a simple binary ‘pass/fail’ classification on each service quality dimension. In turn, the regulated firm would be rewarded or penalised (through cost disallowances, lower weighted average cost of capital (WACC), or lower regulated prices) for failing to meet these standards.

\(^{38}\) Mussa and Rosen (1978)

\(^{39}\) Besanko, Donnenfeld and White (1987, 1988)
But such simple incentive schemes have drawbacks: it is unlikely that the regulatory standard on each dimension is set efficiently over many years; there is no incentive for out-performance, even if doing so is desired by customers; there is no scope for trade-offs between service quality dimensions (again, where doing so would be preferred by customers); and there is insufficient variation in the rewards/penalties across different service dimensions – the regulated firm will choose not to meet the costliest quality standards, even where those quality standards are highly valued by consumers:

The standards themselves are clear and precise, but two major classes of problems arise in their use to monitor and reward quality. First, by evaluating performance relative to a pass/fail cutoff, distinctions among various levels of substandard and superstandard performance are ignored. As a consequence, companies are given targets to achieve, but little incentive to exceed these targets. Standards are rarely set by formal economic analysis equating the marginal benefits of improvements along each dimension equated to the marginal costs so that meeting the standards exactly would enhance consumer welfare. … For example, optimal cutoffs would often vary among companies, but political considerations dictate a single set of standards for all – e.g., what is currently feasible for the least efficient of a set of regulated utilities.

Faced with service quality measurements on a large number of different performance dimensions, how are regulators to aggregate the data to obtain an overall assessment of quality for regulatory purposes? Lynch, Buzas and Berg point out that a simple criterion such as counting the number of dimensions on which the firm meets the regulatory standard leads to misleading and inefficient outcomes – since a firm might meet the performance standard on dimensions which are less valued by consumers and might fail to meet the performance standard on the one dimension which is highly-prized by consumers.

Lynch, Buzas and Berg propose creating a function which maps the different performance measurements to a single, aggregate measure of quality. This function is constructed by surveying the responses of telecommunications experts, using a procedure described in their paper. Using various survey techniques they showed that the quality function could be best represented as a linear function of the various performance dimensions. They found that there was substantial consensus among the experts on the appropriate weights in this quality function. They claim:

The proposed methodology is clearly preferable to the present regulatory regime. It has applicability beyond telecommunications to pressing policy areas such as government payment systems for health care and pollution regulation. By developing and publishing weights, regulatory agencies can communicate clearly how they define quality, leaving each company free to attain its desired target level of overall quality by the mixture of sub-standard and super-standard performance that is most efficient given its cost structure.

Despite this optimism, there are reasons to be concerned with the approach proposed by Lynch, Buzas and Berg. If consumers differ in their weighting of the different service quality dimensions, then if the firm raises some service quality dimensions and lowers others, some consumers may be made strictly worse off. In this circumstance it may be preferable to have minimum standards on at least some dimensions, or, at least, maximum rates of change of service quality over time – just like the ‘side constraints’ which limit the rate of change of individual tariffs in weighted-average price-cap regulation.

Recently Sappington has surveyed the literature on regulation of service quality. Perhaps the strongest message to emerge from this paper is the notion that effectively incentivising service quality involves many of the same issues as does the design of incentives on other desirable regulatory outcomes – such as incentives for minimising expenditure.

For example, Sappington notes one of the classic problems of time-consistency of incentive schemes. If a firm currently operates under a low-powered incentive to reduce expenditure and anticipates a shift to a high-powered incentive to reduce expenditure, coupled with a high-powered service quality incentive, in the future, it has a strong incentive to undertake any quality-enhancing expenditure in the current period when expenditure is not penalised, reaping the gains in the form of a reward through lower costs and higher quality in the future:

Consider a setting where a monopoly supplier presently operates under a ‘low-powered’ reward structure (such as cost-plus regulation or rate-of-return regulation) in which prices are adjusted frequently to match realized costs. In addition, suppose the firm anticipates an impending switch to a more high-powered reward structure such as price cap regulation, in which prices are not linked explicitly to realized operating costs. In such a setting, the firm may perceive financial gain from undertaking non-recurring costs under the low-powered regime that ensure increases in service quality that will persist under the high-powered regime, even as operating costs are reduced under the latter regime. … When the costs of securing the ongoing supply of high levels of service quality – and the associated increased profit – can be recovered via higher prices under the low-powered regime, the regulated firm may implement even more than the welfare-maximizing level of service quality.

Sappington also highlights the point that design of effective service quality incentives requires, as a prerequisite, the ability to verifiably measure all of the relevant service quality dimensions. If there are dimensions of service quality which are difficult to measure (such as the courtesy with which a firm treats its customers), placing the firm under a strong incentive to, say, reduce its expenditure, may lead the firm to reduce quality on the unverifiable dimensions. This is one of the classic arguments against the use of high-powered incentives to reduce expenditure. When a firm has high-powered incentives to reduce costs:

the regulated firm may have substantial incentive to allow (unverifiable) quality to decline if doing so reduces operating costs and thereby increases profit margins. … The optimal resolution of this conflict depends upon the relative importance of reducing costs and increasing quality. When the delivery of high-quality services is of primary importance, low-powered reward structures … can be optimal.\(^{40}\)

\(^{40}\) Laffont and Tirole (2000, page 54) point out: ‘This point is well known. Indeed, the U.S. Department of Defense has often invoked it to motivate the use of cost-plus contracts in contexts in which quality is a sensitive issue and its specifications are hard to pin down … Similarly, the argument has been made several times that the introduction of incentive regulation for power companies conflicts with the safe operation of nuclear power plants. As a last illustration, quality started deteriorating shortly after British Telecom’s 1984 privatization and design of more powerful incentives in the form of a price cap, and quality standards and verification mechanisms had to be set up as a consequence’.
In addition, Sappington again emphasises the role of the price-marginal cost margin in providing some incentives for service quality, which operate even if the regulator cannot directly observe key dimensions of service quality. The precise role of the price-marginal cost margin depends on whether quality and quantity are substitutes or complements (i.e., if quality goes up, do consumers purchase more or less?).

When [quality and quantity] are complements, the higher is the quality of the service, the more of the service consumers will purchase. In this case, the regulator can provide enhanced incentives for the firm to increase the quality of its services by allowing higher prices than she otherwise would. The higher prices increase the profit margin the firm secures on each unit of output it sells and thereby increase the firm’s incentive to sell more units of output. The firm can be sure to sell more units of output by increasing the quality of the services it supplies. In contrast, price reductions can induce higher levels of service quality when quality and quantity are substitutes for consumers. When the two are substitutes, consumers require fewer units of a service to satisfy their needs the higher is the quality of the service. If the marginal price of a service is set below its marginal cost of production, the firm will have a financial incentive to reduce consumer demand for the service. When quality and quantity are substitutes, the firm can reduce consumer demand by increasing quality.

At the end of the day, quality of service is just another of the desirable objectives the regulator would like the regulated firm to pursue. The regulator must design mechanisms which provide the right balance of incentives on the regulated firm to achieve these objectives. These incentive mechanisms will depend on factors specific to each case, such as the value to consumers of different dimensions of service quality, the observability of quality of service measures, and the cost to the firm of increasing quality.
XII. Vertical Integration and Non-Price Discrimination

The previous section focused on the regulation of quality of service delivered to downstream customers. However, the incentives on the regulated firm to deliver service quality are quite different when the downstream customers are themselves competitors to the regulated firm in another market. As we will see, when the regulated entity also operates in an upstream or downstream related market (i.e., where we have vertical integration), there can arise incentives for non-price (i.e., quality) discrimination against upstream or downstream rivals. This is also known as ‘sabotage’.


The potential problem of allowing regulated public utility businesses to operate in unregulated markets has been known for some time. Alfred Kahn, in his 1970 textbook, highlights various ‘line of business’ restrictions on regulated firms dating back to the beginning of the twentieth century. However, the break-up of the US telecommunications industry in the early 1980s brought these issues back to the forefront of regulatory thinking.

In 1987 Timothy Brennan helpfully summarised the arguments in favour of keeping regulated firms out of unregulated markets. His summary focuses on three distinct arguments:

- First, there is a *cost-shifting* argument. When a regulated firm also provides unregulated goods or services, the firm has an incentive to shift as much of the costs of providing the unregulated services into the regulated cost base, thereby raising the apparent cost of the regulated services, ultimately inducing the regulator to raise the prices it can charge in its regulated business (while having no effect on its prices in the unregulated business). In addition, depending on which costs are shifted to the regulated business, the unregulated business may be able to undercut other businesses in the contestable market, expanding its market share, even when it is a higher-cost producer in that market. Furthermore, potential competitors in the contestable market may be deterred from entering if they fear competing against an incumbent who is able to ‘cross-subsidise’ its competitive activities from its uncompetitive activities. In addition the regulated business may be induced to choose an inefficient production technology which increases the extent of economies of scope, precisely in order to increase the extent to which it can shift costs in this way.

The ability to shift costs also extends to the ability to shift risks. A regulated business may be able to absorb some of the risks of the unregulated business into its regulated business. As before, this effectively ‘cross-subsidises’ the contestable business. For example, the regulated business may be able to raise funds at a lower cost of capital than a stand-alone competitive business. In this case, the regulated business is, in effect, using its ability to pass on cost-shocks to
consumers, as a way of lowering its cost of borrowing in the unregulated business.  

- The effects noted above are compounded if the regulated firm also provides unregulated goods or services which are an input into the regulated service. In this case there is also scope for reducing the transparency of input costs. If the regulated firm purchases an input supplier, the regulated firm may be able to raise the apparent price at which it sells inputs to itself. If successful, this would induce the regulator to raise the regulated price. Brennan writes:

  Suppose the regulated firm enters the business of providing inputs needed to produce the regulated product. It then may be able to sell to itself these inputs at a price higher than the costs of producing them. If it can do this, the price in the regulated market will increase, as a result of the increasing costs attributed to the production of the regulated production. The profits of producing the regulated product remain apparently controlled, but the profits of charging the higher price in the regulated market will be recovered as the difference between the costs of producing the inputs and the prices paid by the regulated firm to its upstream affiliate. …

  The solution, Brennan observes, is structural separation:

  [E]fficiency may be enhanced by taking the ability to cross-subsidize away from the regulated firm. This would be guaranteed by preventing the regulated firm from entering other businesses, particularly those in which the inputs used were similar.

- Third, the problems noted above are even more significant when the regulated firm also competes in a downstream market which makes use of the regulated service as an input. In this case the regulated firm may be able to discriminate in the price or quality of the service it provides to downstream rivals, reducing or eliminating downstream competition, and allowing it to increase the downstream price. This strategy might be known as discrimination against downstream rivals. Such discrimination, if effective, both raises prices downstream, may eliminate more efficient rivals, and may reduce downstream product variety, all to the detriment of overall welfare. The regulator may actively seek to prevent upstream discrimination of this kind, but the regulated firm has an incentive to constantly innovate to find new ways to delay or deter effective downstream entry.

  Again, the solution is structural separation:

  The obvious way to prevent regulatory evasion using this tactic is to prevent the entry of the regulated firm into downstream markets or markets for complementary goods. Without the vertical integration or entry, the regulated firm lacks the unregulated market in which it can charge prices and capture returns exceeding competitive levels. It therefore has no incentive to discriminate among unaffiliated providers of the related competing services.

---

41 One of the primary concerns with ‘competitive neutrality’ between government-owned and private businesses is that government-owned businesses may benefit from a lower cost of borrowing, due to their lower perceived risk of default.
Brennan developed these arguments slightly more formally in a later paper published in the *Journal of Regulatory Economics*. In that paper he noted several caveats to the arguments above: first, these results depend on the assumption that the regulator allows prices to adjust to reflect forecast costs; if the regulator could set prices independently of costs, this incentive would be eliminated. Second, entry of the regulated firm into the unregulated business could be welfare enhancing if the regulated firm were, in fact, the lowest-cost provider of unregulated services. In addition, if there are substantial shared costs in producing the regulated and unregulated services, allowing the regulated firm to earn profits on unregulated activities may increase its incentives to produce the regulated services efficiently (in effect, in this case the regulator can ‘free ride’ on the natural incentive of the regulated firm to keep its costs down to remain competitive in the unregulated market). Finally, as Brennan notes, these theories assume that the regulator has some difficulty observing the true efficient cost of the regulated firm. But, nevertheless, the theory assumes that the regulation still has some effect in controlling prices. If the regulator could not observe the true cost of the firm at all, the regulated firm would not need to engage in these strategies to evade regulation!

During the 1990s there was an active regulatory debate about the appropriate role for structural separation in regulatory reform. In some instances structural separation was actively pursued. Vertical structural separation was one of the three key pillars of the 1995 Hilmer reforms. In Australia, structural separation was imposed in a few newly liberalised industries – particularly in the electricity sector and the national interstate rail network.

But structural separation has remained somewhat controversial, particularly in the telecommunications sector. As competition was introduced into telecommunications sectors around the world, no other countries followed the lead of the US by structurally separating the telecommunications industry – most retained a vertically integrated and regulated incumbent. In the US the local exchange companies (LECs) actively lobbied to be allowed to enter the long-distance market.

In the early 2000s the OECD published a report advocating structural separation (with some caveats). A year later, Crandall and Sidak (2002) responded, arguing that there is no need for structural separation in telecommunications. They argued that regulation already effectively controls the incentive for price and non-price discrimination, that the failure for competition to develop in telecommunications is due to the entrants’ own business models rather than any anti-competitive behaviour by incumbents, and that arguments for structural separation are the self-interested lobbying of downstream rival firms who want to ensure their own business success by handicapping incumbents:

Mandatory structural separation is unnecessary because the putative benefits that it would produce are in fact nonexistent. Mandatory structural separation cannot be necessary to increase competition in local exchange services because regulators already (1) prohibit discrimination by means of more direct but less-intrusive behavioral policies and (2) regulate not only the Incumbent Local Exchange Carrier (ILEC’s) prices for end services sold to consumers, but also its prices for Unbundled Network Elements (UNEs) and wholesale services sold to Competitive Local Exchange Carriers (CLECs). Given

42 Brennan (1990)

43 OECD (2001). I was the author of that report.
this multiple overlay of regulation, the ILECs surely cannot exercise market power in the sale of end services to consumers or in the sale of inputs to competitors. No malady exists for mandatory structural separation to cure. Mandatory structural separation, however, would clearly impose substantial costs on the ILECs. Because those costs are unnecessary to advance any public-interest objective, they are also social costs – a waste of economic resources.


Once again regulatory theory for a while lagged regulatory practice. In the late 1990s and early 2000s several papers emerged exploring the incentives on a regulated firm to vertically-integrate and discriminate against downstream rivals. These include papers by Economides (1998) and Weisman (1998).

Beard, Kaserman and Mayo contributed a model in which a dominant firm at the upstream stage supplies an essential input to a Bertrand-competitive differentiated product downstream industry. They asked two questions: (i) Whether the dominant firm has an incentive to vertically-integrate downstream and (ii) if so, whether the dominant firm has an incentive to degrade the quality of the input it supplies to rivals at the downstream stage – which they describe using the (slightly emotive) term ‘sabotage’. They summarise their findings as follows:

An incentive for vertical integration is present whether the upstream stage is regulated or not. An incentive for sabotage … arises only in the presence of a binding regulatory constraint on the input price. Thus, while regulation may be imposed to push the input price closer to costs, such regulation can create a perverse incentive for quality degradation that may be extremely difficult to monitor or prevent. Finally, the conditions under which an incentive for sabotage arises are shown to depend upon: (1) the margin between the regulated price and cost at the upstream stage; (2) the price-cost margin at the downstream stage; and (3) the intensity of competition present in the upstream market [since they assume that the dominant firm faces some upstream competition from a competitive fringe].

Beard, Kaserman and Mayo conclude that the regulator has three policy options: (a) prevent vertical integration (that is, to impose structural separation) – this eliminates the incentive to engage in sabotage but prevents the achievement of potential vertical economies; (b) permit vertical integration and attempt to control sabotage through direct regulation of the quality – they note that ‘enforcement efforts are unlikely to be fully successful in preventing sabotage as quality degradation is difficult to monitor and may be mistakenly attributed to exogenous and/or random occurrences’; (c) wait until competition develops upstream (that is, wait until the regulated firm no longer has a monopoly over the input).

Beard, Kaserman and Mayo observe that the incentive to engage in sabotage depends in part on the relative ‘tightness’ of the regulation upstream and downstream. If regulation is tight upstream and loose (or non-existent) downstream, the firm has an incentive to restrict or prevent competition downstream in an attempt to recover some of the monopoly rent lost to competition upstream. But if regulation is relatively tight on the integrated firm downstream and loose upstream, the firm has an incentive to encourage sales to downstream rivals.
This observation raises the possibility that the incentive to engage in sabotage might be eliminated by applying the same regulatory ‘pressure’ upstream and downstream. One way to do that is through a price cap that includes both upstream (‘access’) services and downstream (‘retail’) services in the same weighted average price cap, as in the ‘global price cap’ proposed by Laffont and Tirole. They note: ‘Such caps would limit the ability of the downstream affiliate of the dominant upstream provider to exploit the higher costs imposed by sabotage through price increases when the cap applied to retail services as well as inputs’. However, they conclude that the usefulness of global price caps remains an open question.


David Mandy usefully summarises the key factors which affect the incentive to engage in sabotage:

1. The cost of carrying out the sabotage. Naturally, the higher the cost of engaging in sabotage, the less the incentive to engage in it. Some of the costs of sabotage (such as engaging in lengthy negotiations, litigation or regulatory proceedings) may be material, especially if there is a risk of a regulatory or antitrust backlash if sabotage is convincingly revealed.

2. The downstream cost structure of the incumbent. If the marginal cost of downstream production is increasing, it becomes increasingly less desirable to attract customers back from the rivals. This is particularly relevant when the incumbent faces a downstream capacity constraint on its production.

3. The relative efficiency of downstream production. The higher the downstream cost of the incumbent relative to its rivals, the lower the incentive to divert sales to itself, relative to sales of the input to rivals.

4. The degree of integration / the degree of autonomy of the downstream subsidiary. The greater the degree of autonomy of the downstream subsidiary, the less the integrated firm may be able to insulate its downstream affiliate from the effect of the sabotage.

5. The extent of downstream product differentiation. If the downstream products of the integrated incumbent are a good substitute (in the eyes of consumers) for the products of rivals, the incumbent is more likely to be able to steal sales back from the rivals by engaging in sabotage.

6. The extent of competition for the essential input. If rivals can switch to alternative suppliers, the incumbent gains less from sabotage.

7. The size of the upstream profit margin. A large upstream margin discourages sabotage, other things equal, as it makes input sales relatively more profitable. ‘An upstream monopoly would be killing the (downstream) goose that laid the golden egg if it conducted sabotage against relatively efficient downstream rivals when its upstream margin is high and its downstream subsidiary’s margin is low.

At the end of the day, Mandy concludes, whether or not firms will engage in sabotage depends on their particular circumstances. Using telecommunications industry data he
predicts that the local exchange companies in the US will indeed have incentives to engage in the ‘foreclosure level’ of sabotage. This ‘prediction stems from a combination of a sufficiently low upstream margin, a sufficiently small downstream efficiency differential, and insufficiently intense downstream competition’. Sabotage, it seems, is something to be concerned about – as regulatory experience bears out.

A related question is whether or not regulators can control sabotage by imposing a ‘non-discrimination’ requirement – i.e., a requirement that the integrated firm must provide service of exactly the same quality to rivals as it provides to itself. Interestingly Sappington and Weisman (2005) show that, even in the absence of discrimination, sabotage might still pay for the integrated firm – that is, it might be profit-maximising for the integrated firm to increase the cost or reduce the quality of its upstream product, even when its own downstream arm faces the same price and quality as the non-integrated rivals.
XIII. Risk, Uncertainty and Cost of Capital Issues

One of the core elements of the regulatory contract between the service provider and its customers is an assurance that the service provider will earn at least a normal return on its investments. Put another way, the service provider must be able to offer its investors a ‘just and reasonable’ return on its investment, commensurate with the risks that those investors face. This was spelt out by the US Supreme Court in the landmark Hope decision. In that case the court observed that:

The return to the equity owner should be commensurate with returns on investments in other enterprises having corresponding risks. That return, moreover, should be sufficient to assure confidence in the financial integrity of the enterprise, so as to maintain its credit and to attract capital.\(^{44}\)

But how do we estimate the return required by investors commensurate with the risks that they face? Estimating the return required by debtholders is relatively straightforward, since the debt of a firm of a given credit rating and duration is a good substitute for the debt of another firm of the same credit rating and duration, and it is fairly straightforward to estimate an appropriate return to debtholders using information on current prices for bonds in the credit markets. But how do we estimate the return required by equityholders to be induced to hold the equity of a regulated firm?\(^{45}\)

In the mid 1960s a new tool appeared for determining the return required by investors to hold an uncertain cash-flow stream. This new tool, developed by Sharpe (1964), Lintner (1965) and Mossin (1966), was called Capital Asset Pricing Model (or ‘CAPM’). The CAPM provided a simple and mathematically elegant methodology for estimating the return required by investors to be induced to hold a particular cash-flow stream.


It didn’t take long for the value of this new tool in regulatory proceedings to be recognised. In 1972, Stewart Myers (the co-author of the famous textbook on corporate finance ‘Brealey and Myers’) argued that this new finance theory should be applied in regulatory rate determinations. Specifically, the innovation proposed by Myers was the estimation of the parameter known as ‘beta’ to determine the appropriate risk class for the regulated firm and therefore the appropriate cost of capital.

It appears that this proposal was wholeheartedly embraced by regulators. Within a decade, the use of CAPM in regulatory proceedings was commonplace.

Over the subsequent two decades, the theory of cost of capital in regulatory proceedings developed along several distinct strands: the first strand sought to improve on the CAPM, by relaxing some of the key assumptions of the CAPM, or by replacing it altogether. The


\(^{45}\) Here it is assumed that investors are called on to finance the entire firm in ‘one shot’. In other cases an established firm will seek financing for an extension or for a new project. The cost of capital for a distinct project could be significantly different than for the firm as a whole. In some cases it may make sense to reflect that different cost of capital in a separate cash-flow stream (and a separate RAB) specific to the project.
second strand sought to take into account the implications of corporate and personal taxes. The third strand addressed questions of the circularity of the use of historic estimates of beta as a measure of future systematic risk. Finally, a more recent strand seeks to take into account the implications of sunk costs, management flexibility and so-called real options for the cost of capital. The first three of these strands of research are briefly surveyed here.

**Alternatives to the CAPM**

In the 1970s several authors proposed extensions to the basic CAPM. For example, Black (1972) relaxed the assumption that investors can borrow and lend at the risk-free rate. The resulting version of the CAPM (known as the ‘Black CAPM’) predicts higher returns on equity for low-beta stocks, and lower returns for high beta stocks. Merton (1973) explored the implications of relaxing the assumption of a single-period in the standard CAPM. He proposed a two-factor model in which the required expected return on an uncertain cash-flow stream is a linear function of both the standard beta multiplied by the market risk premium and an additional term, known as theta, which is the covariance of returns on the stock with reductions in the market’s future investment opportunity set.

More recently, in the 1990s, Fama and French (1993) proposed a three-factor model, in which the required expected return on an uncertain cash-flow stream is a function, not just of the standard Sharpe CAPM beta times the market risk premium, but also a ‘size factor’ relating to the size of the firm (related to the increased default risk of small firms) and a ‘value factor’ relating to the ratio of the book value of the firm to its market value. The Fama-French model has had some empirical success but has been criticised as not based on an underlying theory justifying the choice of factors.

In practice, most regulators and practitioners in the Australian market use a version of the standard Sharpe CAPM with adjustments for taxes. Although there have been a few advocates of using alternative models in the context of regulation (see, for example, Bower, Bower and Logue 1984), the CAPM remains the standard for now. For this reason none of these variations or alternative approaches make it into the top fifty most important papers.

**Taxes and the CAPM**

The standard Sharpe CAPM ignores corporate and personal taxes. Brennan (1970) was the first to present a version of the CAPM which took into account the effects of personal taxation. Following the introduction of a dividend imputation scheme in Australia in 1987 there was a flurry of interest in developing a cost of capital formula which took into account the effects of dividend imputation.

Dividend imputation allows companies to rebate to domestic shareholders the company tax they have already paid (to prevent double taxation of corporate earnings). This can be handled mathematically either as an additional or supplemental dividend from the tax bill of the company, or as a lower tax rate on dividend payments. A series of papers in the early 1990s by Monkhouse (1993, 1996) and Lally (1992) explored the implication of dividend imputation for the standard cost of capital formulae.
The most widely cited paper on the effect of dividend imputation on the cost of capital formulae in Australia is a paper by Officer that appeared in 1994. Officer assumes the simplest possible framework – involving constant in-perpetuity cash-flows. This assumption virtually never holds in practice, but it has the advantage of keeping the mathematics simple. Officer considers various possible definitions of the cash-flow of the firm and derives the corresponding cost of capital for each definition of the cash-flow.

Officer also makes important simplifying assumptions about the nature of taxes and the handling of imputation credits. Specifically, he assumes that dividend and interest income is taxed at the same rate in the hands of investors as income from capital gains. He also assumes that the volume of imputation credits paid out is a simple constant fraction of the tax paid (this assumption is natural in a framework of constant in-perpetuity cash-flows). The fraction he denotes as gamma (γ). Gamma depends on both the share of the imputation credits that are paid out and whether or not those imputation credits have value to investors.

Officer shows that the appropriate weighted average cost of capital for a post-tax cash-flow stream is just equal to the standard ‘plain vanilla’ WACC.46 In effect, in this model, corporate tax and imputation credits have just two effects relative to a model without taxes:

• First, the building block model must ‘scale up’ the post-tax cash-flow stream just enough so that the pre-tax cash-flow stream – from which the allowed revenue is determined – is high enough so that, after deducting the appropriate amount of taxes, the investors earn the return given by the plain vanilla WACC. This is achieved in the building block model through the addition of a ‘tax expenditure’ term as one of the building blocks.

The amount by which the post-tax cash-flow stream is scaled up (and therefore the amount of the ‘tax’ building block) depends, among other things, on the extent to which corporate taxes can be paid out to investors in the form of imputation credits. If all investors are eligible to receive imputation credits, any tax on the earnings of the firm is rebated to its investors. In effect, the firm pays no taxes, and no adjustment for taxes is required. On the other hand, the greater the proportion of the investors which do not benefit from imputation credits, the higher the pre-tax cash-flow stream of the firm must be, so that, after deducting taxes, the investors can earn the required post-tax return.

• Second, the return on equity in the WACC (recall that the WACC is a weighted average of the required return on equity and the required return on debt) must include the effects of imputation credits on the overall return – in particular, when computing the return on equity using the CAPM we must use a market risk premium which takes into account the additional return from imputation credits.

46 In other words, there is no need to make any adjustment for taxes in the WACC formula. This is in contrast to the standard corporate practice, which values projects by discounting an ungeared cash-flow with a WACC formula which includes a tax term in the return on debt component.
The Officer approach has become well established in Australia and, in fact, has been codified in the National Electricity Rules.


A New Zealand academic, Martin Lally, has perhaps been the most strident critic of the Officer approach. In a series of papers he has criticised the Officer model on two primary grounds:

- First, Lally argues that there is an inconsistency in the treatment of foreign investors. The ACCC/AER has consistently used a domestic market approach to CAPM. That is, both the market risk premium and the ‘beta’ of a stock are measured relative to a portfolio of Australian stocks. This is justified on the basis that national stock markets are largely ‘segmented’, with Australians predominantly holding Australian stocks and foreigners primarily holding foreign stocks. But Australian investors receive the full benefit of the imputation credit rebate. If Australian stocks are predominantly held by Australians, the value placed on imputation credits (i.e., the value of ‘gamma’) should be very close to one. Lally argues that the practice of using a domestic CAPM and a gamma of 0.5 is inconsistent. Either we should use a domestic CAPM and assume that imputation credits are valuable to investors (gamma equal to one), or we should use an international CAPM, and use a much lower value of gamma. The current approach, he argues, is a form of ‘cherry-picking’ by the regulated firms. Lally recommends the former approach, with a gamma equal to one.

  Critics of this view argue that Lally has pushed this argument too far. These critics note that the use of a domestic CAPM is primarily a convenience reflecting the domestic bias of investors, while the use of a gamma less than one reflects the undisputed fact of the existence of foreign investors in the Australian market.

- Second, Lally argues that the simplifying assumption in the Officer model, that income and capital gains are taxed at the same rate, ‘could lead to an error in the estimated cost of equity of up to 2 per cent’. This is quite a large potential error. ‘Arguably this is a sufficiently large sum to justify use of a cost of equity formula that recognizes differential personal taxation of capital gains and ordinary income’.

In New Zealand, where there is no capital gains tax, the assumption that income and capital gains are taxed at the same rate is not credible. The consensus approach in New Zealand, unlike Australia, is to use the so-called ‘simplified Brennan-Lally model’ which assumes no tax on capital gains.


The standard approach to estimation of the appropriate cost of capital of a firm involves measurement of the *historic* beta of the firm. This is then used to forecast what the cost of capital of the firm will be in the future. But this creates a problem – what if past outcomes are not a reliable indicator of a firm’s future cost of capital? Even more significantly, what if the regulatory decision itself changes a firm’s future cost of capital?
This criticism dates back to the original proposal by Myers to use the concept of a firm’s ‘beta’ in regulatory proceedings. In the very next issue of the same journal, Breen and Lerner (1972) point out that estimates of beta are not stable over time. They suggest this may be due in part to the fact that beta depends, in part, on managerial actions, which change from period to period. Furthermore, Breen and Lerner (1972) suggest that:

It is reasonable to believe … that regulatory decisions themselves directly affect the value of , for they influence the corporation’s growth rate, stability, size, and payout. If this conjecture is borne out, the regulatory body must be prepared to assess not only the impact of its decisions on the allowable return, but also its effect on the riskiness of the company.

Two papers appeared in the early 1980s directly addressing this concern – Brennan and Schwarz (1982) and Marshall, Yawitz and Greenberg (1981). The core problem is that, for a regulated firm, the firm’s future beta may itself depend on the regulatory decision. But regulatory decisions routinely depend on estimates of the historic beta. Marshall, Yawitz and Greenberg observe:

Although the arguments by Myers and others in support of the use of the CAPM are persuasive, the conventional method of implementation is improper. Typically, the financial risk of the regulated firm is measured by estimating the covariability of its return with that of the market, using historical stock price data. Then, the fair rate of return is determined from the risk-return relationship observed for a reference group of nonregulated firms. Finally, prices are set such that at expected levels of demand owners can anticipate earning the required rate of return on their investment.

But a firm’s systematic risk (as measured by its beta) may itself depend on the regulation:

The process described above is doomed to failure by the implicit and inappropriate assumption that a firm’s financial risk is independent of the activities of regulators and managers. It will be evident from our analysis that risk is endogenous. The prices chosen by regulators serve to determine the impact of demand uncertainties on profits. Further, managers affect risk by their choice of a production technology. Put simply, the choice of technology determines the covariability of costs and revenues and, thereby, the sensitivity of profits to variations in demand.

Marshall et al. model the uncertainty the firm faces explicitly. ‘By basing the analysis directly on the production relationships and on the riskiness of the utilities’ customers, one is able to utilize the CAPM to cut through the circularity that characterizes the conventional approach to the problem’.

The problem of time-varying beta and the possible endogeneity of beta has been the focus of several other more recent papers. Binder and Norton (1999) present evidence that tighter regulation reduces the beta of regulated firms (consistent with the view that regulation partly shields regulated firms from external shocks). Buckland and Fraser (2001) show that betas of regulated utilities vary considerably over time, but in a way that is consistent across different firms, in response to regulatory events and external political influences.
XIV. Two-Way Interconnection

Previous sections of this paper looked at principles for setting prices in general, and specifically in the context of one-way access pricing. In this section we explore the specific issues that arise in ‘two-way access’ problems – that is, where two competing firms both need to purchase inputs from each other. This problem arises in telecommunications, postal services, and rail services. Similar issues arise in banking (the interconnection of ATM networks) and airlines (interconnection of major airline networks).


In the first half of the 1990s a few countries were experimenting with liberalisation of the telecommunications sector. New Zealand was one of the pioneers in this area. As noted earlier, Baumol and Willig had proposed one possible methodology for pricing access to essential inputs, known as the Efficient Component Pricing Rule, but this model didn’t seem to fit the new world in which competing telecommunications networks would each require access to each other’s networks to terminate calls.47

Again, it took a few years for regulatory theory to catch up with the issues ‘on the ground’. In the second half of the 1990s a number of papers appeared clarifying that the interconnection of two telecommunications networks is, indeed, different from classic ‘access to essential facilities’ problems, with a range of subtle and interesting possible outcomes.

One of the first of these papers was a paper by the young British theorist Mark Armstrong. This paper was the first to introduce the distinction between ‘one-way’ and ‘two-way’ access problems.

Armstrong explored the issues that arise when two mature, symmetric telecommunications networks seek to interconnect with each other. Armstrong’s model assumed that both telecommunications networks only used a simple linear retail price and didn’t discriminate between on-net and off-net calls. He assumed that the retail tariffs (the prices charged by each network to their customers for making a call) were unregulated. Similarly, he assumed that the wholesale access charge (the ‘interconnection charge’) was unregulated but was required to be reciprocal (i.e., the same in both directions). The telecommunications customers were assumed to be homogeneous and to have an equal probability of calling anyone else (so that the networks each have balanced inter-network calling patterns, no matter what their relative size).

Armstrong shows that, in stark contrast to the case of one-way access, the two interconnecting networks should be able to agree on the level of the common access charge. But, he points out that the higher the reciprocal access charge, the higher the retail tariffs of the two firms. In this context the access charge could be used as an instrument of collusion – to soften competition in the retail market:

47 The New Zealand Government published a discussion paper at that time raising these issues (written by the present author) entitled ‘The Regulation of Access to Vertically-Integrated Natural Monopolies’ which was cited by Armstrong and Laffont, Rey and Tirole in the papers surveyed here.
The reason for this is that high access charges increase the cost of reducing retail prices unilaterally, since such an action causes the deviating network to have a net outflow of calls to the rival network which incur costly call termination payments. Therefore, even with symmetry – so that access payments cancel out in equilibrium – the level of the access charge may affect downstream competition. In sum, a major contrast between one-way and two-way models of access is that in the former the chief danger is that high access charges could be used to *anti-competitive* effect – perhaps to drive out rivals from the downstream market – whereas in the latter case high access charges may be used as an instrument of *collusion*.

In Armstrong’s model, interconnection is in the mutual interests of the two networks. But the regulator cannot limit its role to that of a mere arbitrator of disputes. In fact, the optimal reciprocal access charge is below the marginal cost of providing access. There is a role for the regulator to actively intervene to set a low reciprocal access charge.


It turns out, however, that Armstrong’s results are sensitive to his assumptions about the structure of the prices that the competing networks can charge. Laffont, Rey and Tirole point out that when the networks are allowed to charge two-part charges to their customers, if they are forced to charge a symmetric (reciprocal) access charge, the profit level of the firms is independent of the access charge – the networks are therefore *indifferent* as to the level of the access charge. The networks have no reason not to agree on pricing access at marginal cost.

Laffont, Rey and Tirole go on to consider the effect of allowing the networks to charge different prices for on-net and off-net calls – that is, different prices for calls to customers on the same network as for calls to customers on the rival network. This again, changes the predictions of the analysis. Whereas in the model with linear retail prices, higher access charges led to higher retail charges (and therefore could facilitate collusion), with the ability to differentiate between on-net and off-net calls, this result no longer holds. In fact, raising the access charge in this framework induces an increase in the off-net call price, and also induces each network to compete more intensely for market share (by lowering the on-net call price and the fixed charge), so as to reduce the overall cost of serving its customers. This is summarised in a later survey by Armstrong and Sappington (2007) as follows:

Firms can determine whether positive or negative network size effects arise through the access tariffs they set. When the firms establish a reciprocal access price in excess of the marginal cost of delivering a call, the equilibrium price for an off-net call will exceed the equilibrium price for an on-net call and a positive network size effect will arise. This effect will intensify competition for subscribers. In contrast, if the firms set the access price below marginal cost, the equilibrium price for an on-net call will exceed the equilibrium for an off-net call. The resulting negative network size effect will induce relatively weak competition for subscribers, and thereby increase the profit of established competitors. Therefore, in this setting, the firms will wish to price access below the socially desirable access price. The regulator’s task in this setting, then, will be to ensure the firms do not negotiate access payments that are unduly low.
Price discrimination between on-net and off-net calls, which is neither cost- nor demand-based, normally introduces an economic distortion and reduces welfare. But Laffont, Rey and Tirole show that in circumstances where the two networks are relatively poor substitutes for each other (so that the competition between them is limited), price discrimination of this kind can alleviate the double-marginalisation effect. As a result, some price discrimination of this kind can be beneficial.

However, price discrimination between on-net and off-net calls can also hinder small-scale entry:

Here, a full-coverage incumbent can squeeze a small-coverage entrant by insisting on a high access price. The high access charge translates into high off-net prices, creating a de facto lack of interconnection. We thus conclude that freely negotiated access prices raise serious anticompetitive concerns under price discrimination unless entrants can and do find it profitable to quickly build a large-coverage network.


In addition to the papers mentioned above, Carter and Wright (1999, 2003) have also made important contributions to the theory of two-way access pricing. However, one of the most valuable papers in this area is another of the survey articles published in the Handbooks in Economics series.

In this survey article, Armstrong further relaxes some of the assumptions of the papers above, such as the assumption of a common access charge. The assumption that the access charge is the same in both directions doesn’t make sense for the interconnection of networks with quite different cost structures, such as the interconnection of fixed and mobile networks. In addition, Armstrong shows that some of the key conclusions of the earlier papers – namely that (a) access charges can be used to sustain collusive outcomes, or (b) that access charges have no effect on equilibrium profits – are not robust to small changes in the assumptions.

Armstrong also emphasises that the more tasks the access charge is required to perform, such as:

(a) the need to give entrants good make-or-buy incentives; (b) the need to induce a desirable amount of entry, given the incumbent’s retail tariff distortions, and (c) the need to control the incumbent’s retail prices when those were not directly controlled by regulation – the more complicated and informationally-demanding the various access pricing formulas become.

Finally, Armstrong also notes that even in the most ‘competitive’ of markets – such as the mobile sector – there is likely to remain a role for regulation of two-way access charges. In the later survey paper with David Sappington he summarises:

In sum, the developing literature on two-way access pricing suggests a number of continuing roles for regulation, even when competition for subscribers is sufficiently vigorous to limit the need for explicit retail tariff regulation.
XV. The Design of Regulatory Institutions

Within the transactions cost approach to public utility regulation, the regulator plays the role of the neutral arbiter – establishing, in the event of dispute, the terms and conditions that the parties would-have-agreed-on if they had negotiated a complete contingent contract at the outset.

However, even within this paradigm, there remain many questions to be resolved. For example, to what extent should the regulatory contract be formalised (perhaps in writing), or to what extent should the regulator be allowed discretion? What rules should be ‘set in stone’? What is the role of the wider judicial system or political system in influencing the optimal form of the regulator or the regulatory contract? How might the regulatory contract and regulatory institutions vary over time or across countries? What can we say about the design of the regulatory authority itself – should it be part of government or independent? Should it consist of a single person (as many UK regulators were in the past) or should it consist of more than one person (as is common in Australia). What is the ideal scope for regulatory responsibility: Should regulators have responsibility for a single firm, an entire industry, or more than one industry?

A number of papers emerged in the 1990s addressing these questions, particularly out of the World Bank. The World Bank is particularly interested in these issues as they apply to developing countries. In developing countries the primary regulatory problem is attracting sufficient funds to achieve adequate investment in infrastructure. Perhaps for this reason, this literature has tended to focus almost exclusively on the problem of fostering investment by the regulated firm (and has ignored the need for investment by its customers).

According to this perspective, the primary deterrent to infrastructure investment is the risk of expropriation – that is the risk that the firm, having made the investment, will be forced to provide services at prices that do not fully cover the cost of the investment. The primary task for the government therefore is designing the regulatory framework so as to establish a credible regulatory commitment – a commitment that the regulated firm will be allowed to earn an adequate return on its investment.

This task is perhaps harder than it may first appear. In fact, perversely, the more power a government retains for itself, the harder it will find it to make such a credible commitment – since it cannot bind itself to not change its mind in the future. A key question in this literature, therefore, is how governments commit themselves through the design of regulatory institutions to not expropriating the legitimate sunk investment of regulated firms – through constitutional requirements, through the establishment of an independent judiciary, or through the establishment of an independent regulatory authority, with certain rules and objectives.

In 1991, Lewis and Sappington explored the incentives for investment by a firm whose investment outlives the lifespan of the regulator. They assumed the regulator cares only about the welfare of current consumers. They find (unsurprisingly) that under-investment is the likely outcome. In a 1994 paper, Gilbert and Newbery also assumed that regulators care only about consumer interests – but in their model, the regulation operates under a constitutional commitment to an adequate rate of return. They find that an efficient investment programme can be supported for some parameters – especially when the regulator is permitted to disallow certain assets in the regulatory asset base as not ‘used and useful’.
One of the more important papers from this time period is a paper by Brian Levy of the World Bank and Pablo Spiller of the University of California, Berkeley.

Levy and Spiller do a good job of explaining the core problem they seek to address:

The combination of significant investments in durable, specific assets with the high level of politicization of utilities has the following result: utilities are highly vulnerable to administrative expropriation of their vast quasi-rents. Administrative expropriation may take several forms. Although the easiest form of administrative expropriation is the setting of prices below long-run average costs, it may also take the form of specific requirements concerning investments, equipment purchases, or labour contract conditions that extract the company's quasi-rents. Where the threat of administrative expropriation is great, private investors will limit their exposure.48

The primary regulatory problem is how to provide a guarantee to the regulated firm that such expropriation will not occur in the future. Levy and Spiller point out that the economics literature has instead focused on questions such as optimal pricing, cross-subsidies, entry, interconnection and so on. They argue that these questions are inevitably secondary to questions of how to establish the central regulatory commitment.

Levy and Spiller emphasise the importance of taking into account a country's endowment of political, judicial and regulatory institutions, such as: 'explicit separation of powers between legislative, executive, and judicial organs of government; a written constitution that limits the legislative power of the executive and is enforced by the courts; two legislative houses elected under different voting rules; an electoral system calibrated to produce either proliferation of minority parties or a set of parties whose ability to impose discipline on their legislators is weak; and a federal structure of power, with strong decentralization even to the local level'.

These factors in combination determine place constraints on how the regulatory framework of a country can be designed. For example, countries with a strong and independent judiciary, with a reputation for impartiality, will find it easier to rely on the judiciary to enforce contractual undertakings or to limit administrative discretion. In the absence of a strong judiciary it might be necessary to rely on government ownership of the regulated firm or some form of international guarantees. Even where the judiciary is strong and independent, if the country has a strong central executive and if 'executive powers alternate between political parties with substantially different interests, specific legislation need not constitute a viable safeguard against administrative discretion as changes in the law could follow directly from a change in government.'49 These countries may need to rely on formal contractual arrangements or constitutional protections for regulated investment. Another factor is the administrative capabilities – 'the ability of the nation's professionals (e.g., academics, lawyers, bureaucrats) to handle complex regulatory concepts and processes in a relatively efficacious manner, without triggering excessive disputes and litigation – are of particular relevance. These capabilities will

---

48 Page 204.
49 Page 207.
determine the potential for the successful implementation of complex regulatory designs.\(^{50}\)

Levy and Spiller write:

Our evidence suggests that regulatory commitment can indeed be developed in what appear to be problematic environments, that without such commitment long-term investment will not take place, that achieving such commitment may require inflexible regulatory regimes that go against prevailing academic views, that in some cases public ownership of utilities is the default mode of organization, and furthermore, that such ownership may be the only feasible alternative.\(^{51}\)

At the end of the day, the core message of Levy and Spiller’s paper is that institutional arrangements matter. In particular, the ability of a country’s institutions, collectively, to guarantee to a firm that his/her investments will not be expropriated ex post. No matter how much effort is put into designing, say, optimal pricing principles, efficient outcomes will not be achieved overall if the regulated firm does not have confidence that it will recover an adequate return on its investment.

Although this message is primarily relevant to designers of regulatory regimes, for practising regulatory authorities it highlights the critical importance of trust and reputation. If the regulatory authority can foster a reputation for acting in a manner which protects investment (either investment by the regulated firm or its customers), and which is ‘fair’ and seen to be ‘fair’ to the parties involved, effective outcomes overall are more likely to be achieved. Conversely, if the regulatory authority cannot be trusted to adequately protect investment, it may find that its administrative discretion is curtailed, or important regulatory tasks delegated elsewhere.

In 1999 there appeared a survey paper on the design of regulatory institutions by Estache and Martimort. This paper is not particularly well written, and it has not appeared in a major economics journal (it was published in the World Bank Policy Research Working Paper series), but covers a very large range of issues, such as: (a) the incentive problems that arise when the regulator acquires new information on, say, the firm’s performance over time; (b) the separation of powers across different regulatory institutions; (c) the design of incentives on regulators; (d) preventing of capture of regulators by the regulated industry; (e) the optimal degree of decentralisation; and (f) whether regulators should be appointed or elected.

Estache has also been involved in a number of papers exploring empirical issues in the design of regulatory institutions. One recent paper, for example, found that the existence of a regulatory authority was positively correlated with greater productive efficiency and higher levels of social welfare.\(^{52}\)

\(^{50}\) Page 208.

\(^{51}\) Page 202.

One more paper, which appeared at the end of the 1990s, deserves inclusion in the fifty most important papers in the economics of regulation. The paper, by David Martimort (like the Gilbert and Newbery paper mentioned above) starts from the observation that regulation is a repeated game between the regulator and the regulated firm – which gives opportunities for significantly larger rewards and penalties and significantly richer strategies than in a simple one-shot game. In particular, the repeated nature of the game increases the scope for ‘regulatory capture’ – that is, the tendency for a regulatory agency to act in the interests of the regulated firm over time.

Interestingly, political scientists and regulators have long argued that regulatory authorities go through a ‘life-cycle’. As Martimort reports:

Casual observations suggest that agencies start to behave in the public interest and then become increasingly inefficient, bureaucratized and more eager to please private interests. This point has been forcefully made by several political scientists … who have put forward the view that government efficiency cannot be judged from a static perspective but should instead be understood as a dynamic phenomenon.

Martimort goes on to quote from Alfred Kahn, in his classic text on the economics of regulation:

There have been wide differences between [regulatory] commissions and in their legislative mandates, and changes over time in the political environment in which they operate. Both the differences and changes … are illustrated, for example, in the oversimplified but illuminating generalization that regulatory commissions tend to go through a life cycle, setting out as vigorous, imaginative, and enthusiastic protagonists of the public interest, reflecting the public concern and ferment that had to be mobilized to legislate them into existence in the first place, defining their responsibilities broadly and creatively, then gradually becoming devitalized, limited in their perspective, routinized and bureaucratized in their policies and procedures and increasingly solicitous and protective of the interests of the companies they are supposed to regulate, resistant to changes, wedded to status quo.

Martimort’s fundamental thesis is straightforward. He recognises the key role of information in the regulatory process. In his model, the regulator learns information about the regulated firm in the course of the regulatory process and can choose whether to act on this information or not. By choosing not to act on the information, the regulator leaves rent to the regulated firm, for which the regulator seeks to be rewarded – perhaps in the form of future job opportunities with the regulated firm. Recognising this problem, the government seeks to limit the administrative discretion of the regulator. Over time as more opportunities for colluding with the regulated firm arise, administrative rules must increase in number and restrictiveness to ‘curb the pursuit of private interests’. In fact, the optimal policy response by the government in Martimort’s model is a reduction over time in the discretion left to the regulator.

53 Here there are referring to all regulatory bodies – not just natural monopoly regulators.

54 Pages 938-939.

Interestingly, the long-run outcome is the same as if side-payments were permitted directly from the regulated firm to the regulator. The effect of disallowing bribery is merely to slow the rate at which the overall outcome approaches a situation of capture. On the other hand, changes in the design of the regulatory institutions – such as changes in the length of the regulator’s tenure, changes in the design of the regulatory tasks, can slow the tendency to capture.

Although Martimort’s paper is, as before, more directly relevant for designers of regulatory institutions, the implications for practising regulatory authorities are worrying. The overall regulatory framework for public utilities in Australia is fairly new. Already there is some tendency towards greater prescription in the regulatory process (for example, the encoding of the ACCC electricity guidelines into the National Electricity Rules). Will there be a long-run tendency towards a reduction in regulatory discretion and a routinisation of day-to-day regulatory operations?
Commentary

This paper is primarily a literature survey. However, I have in many respects been critical of much of the direction of this literature. In any public policy problem it is essential to first understand the problem we are trying to solve. In my view, a case can be made that neoclassical economists have misunderstood the fundamental economic rationale for public utility regulation. Regulators do not, in practice, behave as though their primary concern is the minimisation of deadweight loss. Instead, according to the transactions-cost perspective, public utility regulation is best thought of as a mechanism for protecting the sunk investments of both the monopoly service provider and its customers.

In my view, the economics literature could have been more valuable to regulatory practitioners if economists had taken seriously the approach first put forward by Goldberg and Williamson in 1976. It is not that these ideas were simply ignored. After all, the new field of transactions cost economics emerged from these papers, for which Williamson won the Nobel Prize in 2009. However, where these ideas have been applied in public utility regulation, they have focused only on the need to protect the sunk investment by the regulated firm. The need for sunk investment by customers of the service provider has been largely ignored.

In my view, in the field of public utility regulation much remains to be done. For example, many of the conventional practices of regulators are still poorly understood by economists. For example, although the transactions-cost approach to utility regulation provides some suggestion as to why regulators focus on ensuring stable prices, there is not yet a theoretical foundation for this practice. What degree of stability can be promised in the face of uncertainty? What promises can or should the service provider make to its customers regarding price stability? What role do cost allocation mechanisms play in ensuring price stability? Similar issues arise with common pricing practices such as the control of price discrimination. Precisely what forms of discrimination should be allowed and when?

Furthermore, there is a need for a better understanding of the proper role for regulators themselves. What does it mean for a regulator to play the role of a dispute resolution mechanism in a long term contract? What is the proper scope and grounds for appeal from regulatory decisions? What functions should the regulator carry out? When should a regulator keep and when should it break its promises? What is the best division between ex ante regulatory rules and ex post regulatory discretion? In addition, there is a need for a better understanding of the potential for restrictions on the regulatory process to limit the time and resources consumed in resolving regulatory disputes.

There is also a need for a better understanding of the role of customers in the regulatory process. How should customers’ interests be aggregated and represented in regulatory processes? Most of the states in the US have established institutions to represent customer interests in utility rate-making processes. These institutions are hardly ever mentioned in the economics literature. What exactly is the theoretical role for such institutions?

There is also significant scope for further interaction between regulatory policy-makers and other designers of long-term contracts. For example, what can regulators learn from government agencies involved in the design of long-term contracts, such as procurement agencies, or agencies involved in designing long-term public-private partnerships? There have been important economic studies of long-term private contracts which have yielded
a few insights for utility regulation, but there remains much to be understood about the kinds of provisions and clauses that service providers and their customers will include in private long-term contracts. It seems to me that regulatory economists have only begun to tap the resources available around them.

**Conclusion**

This paper is, at one level, merely a literature survey. On the other hand, it is not an attempt to survey each and every paper in the field of regulatory economics. Rather, it is an attempt to identify and summarise the fifty most important papers in the field of public utility regulation – putting each paper into its historical context and explaining the resulting developments in theory and practice. The fifty papers were selected from the perspective of a regulatory practitioner. It is the answer to the question: If I only have time to read fifty papers in the economics of regulation, what fifty papers should I choose?

Of course, the selection of the fifty most important papers is inherently somewhat subjective. I have focused on papers which are important either because they stimulated entirely new lines of enquiry, or because they had a major impact on regulatory practice, or because they summarise the state of the art. I have focused on papers which would be accessible to a regulatory practitioner. In addition the selection of papers reflects my conviction of the importance of the transactions cost approach to public utility regulation, which has yet to be fully explored.

As this survey has shown, over the past few decades considerable advances have been made in developing better regulatory policies. In particular, the past few decades have seen considerable progress in understanding one-way and two-way access pricing, the foundations of the building block model, and the key role played by information. Important advances have also been made in understanding regulatory incentives and in the application of finance theory to the regulatory task.

Interestingly, many of the key advances in regulatory thinking were in response to, rather than leading, regulatory reforms. In many instances key theoretical papers emerged after, rather than before, key reforms. Where economic theory has been pursued independently (such as the theoretical work on incentive regulation) its application to regulatory practice has been limited. The impact of the neoclassical approach to public utility pricing remains somewhat limited. 54

In a sense, the study of public utility regulation is at a crossroads. Many respected economists (such as Baumol) have recognised that their advice has not always met the needs of policy-makers. In my view, as I have emphasised throughout this review, this is due to the neglect by most economists of the importance of sunk investments by customers and the consequent need for mechanisms to protect and promote that investment. Taking Baumol out of context, if economists do not take these sunk investments seriously, ‘those who seek the help of microeconomists are likely to continue to feel that they have asked for bread and we have given them only stone’. 55


57 Baumol (1986).
Appendix: Top Fifty Papers in Regulatory Economics


Appendix: Fifty Most Important Papers in Regulatory Economics


References


Biggar, Darryl (2009), ‘Is Protecting Sunk Investments by Consumers a Key Rationale for Natural Monopoly Regulation’, *Review of Network Economics*, 8(2), 128-152


Cowan, S. (1997), ‘Tight Average Revenue Regulation Can be Worse than No Regulation’, *Journal of Industrial Economics*, 45(1), 75-88


